Gesture and Speech in Interaction - 4th edition (GESPIN 4)

Gaëlle Ferré, Tutton Mark

UNIVERSITY OF NANTES

FRANCE

Gesture and Speech in Interaction

4th edition

Gaëlle FERRÉ & Mark TUTTON

2 - 4 SEPTEMBER 2015

Contents

Foreword

The fourth edition of *Gesture and Speech in Interaction* (GESPIN) was held in Nantes, France. After Poznan in Poland, Bielefeld in Germany and Tilburg in the Netherlands, it has been our pleasure to host this international conference. With more than 40 papers, these proceedings show just what a flourishing field of enquiry gesture studies continues to be. Although the majority of the participants were European, we were delighted that non European countries were represented as well. This shows the will of researchers – both junior and senior – to come together to present the findings of their research in what is a very exciting and thriving domain.

The keynote speeches of the conference addressed three different aspects of multimodal interaction: gesture and grammar, gesture acquisition, and gesture and social interaction. In a talk entitled *Qualities of event construal in speech and gesture: Aspect and tense*, **Alan Cienki** presented an ongoing research project on narratives in French, German and Russian, a project that focuses especially on the verbal and gestural expression of grammatical tense and aspect in narratives in the three languages. **Jean-Marc Colletta**'s talk, entitled *Gesture and Language Development: towards a unified theoretical framework*, described the joint acquisition and development of speech and early conventional and representational gestures. In *Grammar, deixis, and multimodality between code-manifestation and code-integration or why Kendon's Continuum should be transformed into a gestural circle*, **Ellen Fricke** proposed a revisited grammar of noun phrases that integrates gestures as part of the semiotic and typological codes of individual languages. From a pragmatic and cognitive perspective, **Judith Holler** explored the use of gaze and hand gestures as means of organizing turns at talk as well as establishing common ground in a presentation entitled *On the pragmatics of multi-modal face-to-face communication: Gesture, speech and gaze in the coordination of mental states and social interaction.*

Among the talks and posters presented at the conference, the vast majority of topics related, quite naturally, to gesture and speech in interaction – understood both in terms of mapping of units in different semiotic modes and of the use of gesture and speech in social interaction. Although it would be too long to quote every single author and paper in this short foreword, we will give the reader an outline of the variety of approaches presented at GESPIN this year. Several presentations explored the effects of impairments (such as diseases or the natural ageing process) on gesture and speech. The communicative relevance of gesture and speech and audience-design in natural interactions, as well as in more controlled settings like television debates and reports, was another topic addressed during the conference. Some participants also presented research on first and second language learning, while others discussed the relationship between gesture and intonation. While most participants presented research on gesture and speech from an observer's perspective, be it in semiotics or pragmatics, some nevertheless focused on another important aspect: the cognitive processes involved in language production and perception. Last but not least, participants also presented talks and posters on the computational analysis of gestures, whether involving external devices (e.g. mocap, kinect) or concerning the use of specially-designed computer software for the post-treatment of gestural data. Importantly, new links were made between semiotics and mocap data.

Finally, we would like to take this opportunity to acknowledge the work of a certain number of people at the University of Nantes who were crucial to the successful hosting of this conference. Firstly there is Myriam Lecoz, the secretary of our research laboratory (LLING), who provided us with continuous advice and support and dealt with many key administrative tasks. Secondly, there are our wonderful student volunteers: Anne-Laure Besnard, Quentin Brisson, Manon Lelandais, and Benjamin Lourenço, without whose help the logistical organization of the three conference days would have been impossible. Many thanks also go to the members of the scientific committee, who did a very fine job in ensuring the reviewing process went as smoothly as possible. We are also grateful to the *Linguistics Laboratory of*

Nantes (LLING), the *English Department* of our university, the *University of Nantes* itself, as well as to the *Infrastructure de Recherche pour les Corpus Oraux et Multimodaux (IRCOM)*, all of whom granted us funding. In doing so, they enabled this conference to take place. We sincerely hope that is has been as enjoyable for every participant as it has been for us.

September 2015
Gaëlle Ferré & Mark Tutton

Committees

Local organizing committee

- Gaëlle Ferré (LLING, Nantes University, Gaelle.Ferre@univ-nantes.fr)
- Mark Tutton (LLING, Nantes University, Mark.Tutton@univ-nantes.fr)
- Manon Lelandais (LLING, Nantes University, Manon.Lelandais@etu.univ-nantes.fr)
- Benjamin Lourenço (LLING, Nantes University, Benjamin.Lourenco@etu.univ-nantes.fr)

Scientific board

- Mats Andrén (U. Lund, Sweden)
- Dominique Boutet (Evry, France)
- Jana Bressem (TU Chemnitz, Germany)
- Heather Brookes (U. Cape Town, South Africa)
- Alan Cienki (VU Amsterdam, The Netherlands & Moscow State Linguistic U., Russia)
- Doron Cohen (U. Manchester, UK)
- Jean-Marc Colletta (U. Grenoble, France)
- Gaëlle Ferré (U. Nantes, France)
- Elen Fricke (TU Chemnitz, Germany)
- Alexia Galati (U. Cyprus)
- Marianne Gullberg (U. Lund, Sweden)
- Daniel Gurney (U. Hertfordshire, UK)
- Simon Harrison (U. Nottingham Ningbo, China)
- Judith Holler (MPI, The Netherlands)
- Ewa Jarmolowicz (Adam Mickiewicz University, Poland)
- Konrad Juszczyk (Adam Mickiewicz University, Poland)
- Maciej Karpinski (Adam Mickiewicz University, Poland)
- Sotaro Kita (U. Warwick, UK)
- Stefan Kopp (U. Bielefeld, Germany)
- Emiel Krahmer (U. Tilburg, The Netherlands)
- Anna Kuhlen (U. Humbolt Berlin, Germany)
- Silva H. Ladewig (Europa-Universität Frankfurt, Germany)

- Maarten Lemmens (U. Lille 3, France)
- Zofia Malisz (U. Bielefeld, Germany)
- Irene Mittelberg (HUMTEC Aachen, Germany)
- Cornelia Müller (EUV, Viadrina Gesture Centre, Germany)
- Asli Ozyürek (Radboud University Nijmegen and MPI, The Netherlands)
- Katharina J. Rohlfing (U. Bielefeld, Germany)
- Gale Stam (National Louis University, USA)
- Marc Swerts (U. Tilburg, The Netherlands)
- Michal Szczyszek (Adam Mickiewicz University, Poland)
- Marion Tellier (U. Aix Marseille, France)
- Mark Tutton (U. Nantes, France)
- Petra Wagner (U. Bielefeld, Germany)

GESPIN conference board

- Ewa Jarmolowicz (Adam Mickiewicz University, Poland)
- Konrad Juszczyk (Adam Mickiewicz University, Poland)
- Maciej Karpinski (Adam Mickiewicz University, Poland)
- Zofia Malisz (U. Bielefeld, Germany)
- Katharina J. Rohlfing (U. Bielefeld, Germany)
- Michal Szczyszek (Adam Mickiewicz University, Poland)
- Petra Wagner (U. Bielefeld, Germany)

Plenary Speakers

Qualities of event construal in speech and gesture: Aspect and tense

Alan Cienki
a.cienki@vu.nl

This talk will present preliminary results of an international project on verbal and co-verbal means of "event construal" (understood as per Langacker 1987 and Croft 2012). The project is based at Moscow State Linguistic University but involving teams of colleagues from France and Germany as well as Russia. The choice of French, German, and Russian as languages for analysis was motivated by the differing morphological and/or lexical means that are used, or not, in the three languages (and in some other members of the Romance, Germanic, and Slavic language families) for talking about different types of events. Whereas French and German rely on a variety of tense forms, particularly to talk about events in the past, German prefixes on verbs additionally highlight numerous distinctions of manner of action (Aktionsart or "lexical aspect"), while Russian's simple tense system (past, present, future) is complemented by a distinction of two grammatical aspect categories as well as categories of Aktionsarten marked by verbal affixes.

The results reported on here will focus on the relations of grammatical aspect and tense to the qualities of speakers' coverbal gestures. In previous research, Duncan (2002) showed that the duration of gesture strokes tends to be longer and more agitated with event descriptions in the "imperfective" than with those using "perfective" verb forms in English and in Mandarin Chinese, findings that were further confirmed for English by McNeill (2003) and Parrill et al. (2013). In the present study, narratives about different types of events were elicited using a protocol from Becker et al. (2011) with native speakers of French, German, and Russian. Coverbal gestures were analyzed using a system of "boundary schemas" developed for this project, based on (Müller 2000). These have to do with whether the stroke of a given gesture phrase involves a pulse of effort ("bounded") or not ("unbounded"). Bounded gestures, hypothesized to correlate more with perfective aspect and perfect tenses, show greater effort markedly exerted at the onset of the stroke, the offset, both, or repeated throughout the stroke. Unbounded gestures, by contrast, involve effort spread evenly over the stroke. The analysis of effort involves attention to kinesiological parameters (Boutet 2010) based on physiological features, e.g., the relation of the form of the movement to the structure of the hand, wrist, arm, etc. Initial findings suggest that certain grammatical distinctions concern qualities of event structure that are also expressed in patterns of speakers' coverbal motoric behavior. The results will be compared to findings from PhD research by Wang (VU Amsterdam & Xiamen U.) on the use of gesture in Mandarin Chinese with clauses with aspectual particles (e.g., *le* marking actualization of an action, *zai* marking progressives, *zhe* marking duration).

While the categories customarily used to characterize event construal in grammar, such as different aspectual distinctions, show certain connections to speakers' use of gesture, they do not carry over to gesture in a straightforward way. We see from these studies how research on gesture from a linguistic perspective (Müller et al. 2013) can provide more nuanced insights into a process Slobin (1987) characterized as "thinking for speaking", investigated here in terms of how the construal of events appears in embodied expression while speaking.

Abbreviated references:

- Becker, R., Cienki, A., Bennett, A., Cudina, C., et al. 2011. Aktionsarten, speech and gesture.
- Boutet, D. 2010. Structuration physiologique de la gestuelle: Modèle et tests.
- Croft, W. 2012. Verbs: Aspect and causal structure.
- Duncan, S. 2002. Gesture, verb aspect, and the nature of iconic imagery in natural discourse.
- Langacker, R. 1987. Foundations of cognitive grammar. Volume 1. Theoretical prerequisites.
- McNeill, D. 2003. Aspects of aspect.
- Müller, C. 2000. Zeit als Raum.

- Müller, C., Ladewig, S., & Bressem, J. 2013. Gestures and speech from a linguistic perspective.
- Parrill, F., Bergen, B., & Lichtenstein, P. 2013. Grammatical aspect, gesture, and conceptualization.
- Slobin, D. 1987. Thinking for speaking.

Gesture and Language Development: towards a unified theoretical framework

Jean-Marc Colletta

jean-marc.colletta@u-grenoble3.fr

Children communicate their needs through bodily behavior and begin to gesture way before talking. Together with expressions of emotions, gestures, such as pointing or waving goodbye, constitute the principal means of interacting with others before the emergence of the first words. Children continue to gesture during their second year as they start talking and gesturing in bimodal language production. Older children carry on using speech associated gestures through to adulthood as their language repertoire fulfills new social-interactional needs and incorporates new discourse genres. Thus, as a number of studies have demonstrated over the past twenty years, verbal language does not replace gestures as children grow up. Rather, language is to be considered as a compound of audio-linguistic signs and visual-kinesic signs whose use and forms evolve together in the course of age.

To present an overview of early and later gesture and language acquisition is too big a scope for this presentation, considering today's vast literature on the subject. In this presentation, I will rather present a set of a priori unrelated observations and results on early emblems and representational gestures, gestures of the abstract, changes in gesture production and in the relation between speech and gesture during childhood, gesture variation in situational and discourse context, as well as teacher's gestures during language and maths class. I will then discuss these results within a unified theoretical framework that builds on "mimesis theory" as introduced by Marcel Jousse in his "Anthropologie du geste" (Calbris, 2011), René Girard's mimetic theory and Jordan Zlatev and collaborators's work on mimesis (Zlatev, 2002; Zlatev et al., 2008). Language acquisition is then to be seen as an embodied process fully embedded into sensory and motoric experience of both the physical and the social world, and gesture as a shared representation mechanism that both grounds and extends linguistic means for communication among human beings.

References:

- Calbris, G. (2011). *Elements of Meaning in Gesture*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Zlatev, J. (2002). Mimesis: the "missing link" between signals and symbols in phylogeny and ontogeny? A. Pajunen (Ed.), *Mimesis, Sign and Language Evolution* (pp.93-122). Publications in General Linguistics, 3. Turku: University of Turku Press.
- Zlatev, J., Racine, T.P., Sinha, C. Itkonen, E. (2008). *The Shared Mind. Perspectives on intersubjectivity*. Amsterdam/Philadelphia, John Benjamins Publishing Company.

Grammar, deixis, and multimodality between code-manifestation and code-integration or why Kendon's Continuum should be transformed into a gestural circle

Ellen Fricke

ellen.fricke@phil.tu-chemnitz.de

Until recently, the idea that a multimodal approach to grammar is necessary was by no means evident. Most grammarians so far focused their grammatical analyses on written and spoken language without considering co-speech gestures. Yet the progress in gesture studies offers a new perspective on the grammatical capacity of gestures accompanying speech (Fricke 2008, 2012, 2013, 2014a, b, c; Harrison 2008, 2009; Ladewig 2011; Bressem 2012). Not only is human speech composed of articulations of the mouth, primarily perceived by ear, but also of visible articulations of other body parts affecting the eye (e.g., Kendon 2004; Müller 1998, McNeill 1992, 2005, for an overview see Müller, Cienki, Fricke et al. 2013 and 2014). In this regard, the movements of the hands play a special role: the sign languages of the deaf show that movements of the hand alone can function as articulators of fully established languages (Wundt [1900] 1904). If it is the case that movements of the hand inherently have the potential for establishing a grammar, what are the grammatical implications of all those hand movements that accompany the speech of hearing people?

Are single languages like French, English, or German partially multimodal? How far is the faculty of language bound to a particular mode of manifestation? If we conceive multimodality as a global dimension of linguistic and semiotic analysis which is generally applicable to language and other systems of signs then we have to broaden our perspective by also including grammars of single languages and the human faculty of language. With respect to linguistics and by focusing on the example of noun phrases, I will show that this extension of perspective on multimodality reveals two basic principles: Firstly, multimodal code-integration of gestures within grammars of single languages on the level of the language system; secondly, processes of multimodal code-manifestation of certain structural and typological aspects on the verbal and gestural level provided by the codes of single languages as well as the general human faculty of language.

With regard to gesture studies, evidence of multimodal grammatical structures and functions (e.g., multimodal modication in noun phrases or constituency and recursion in syntax (Fricke 2012, 2013)) could challenge the current view of Kendon's Continuum (McNeill 1992) as a straight line from left to right. If spoken langages are conceived of as being basically multimodal, then it is necessary to take into consideration speech and co-speech gestures as a unified whole when comparing them to sign languages. In the light of these findings, we propose transforming the straight line that joins them in Kendon's Continuum into a gestural circle, which may more adequately represent their close relation.

References:

- Bressem, Jana (2012). Repetitions in Gesture: Structures, Functions, and Cognitive Aspects. Dissertation, Europa-Universität Viadrina, Frankfurt (Oder).

- Fricke, Ellen (2008). Grundlagen einer multimodalen Grammatik: Syntaktische Strukturen und Funktionen. [Foundations of a Multimodal Grammar: Syntactic Structures and Functions]. Habilitation, Europa-Universität Viadrina, Frankfurt (Oder).

- Fricke, Ellen (2012). *Grammatik multimodal: Wie Wörter und Gesten zusammenwirken.* Berlin and Boston.

- Fricke, Ellen (2013). Towards a unified grammar of gesture and speech: A multimodal approach. In: Cornelia Müller, Alan Cienki, Ellen Fricke et al. (eds.), 733-754.

- Fricke, Ellen (2014a). Deixis, gesture, and embodiment from a linguistic point of view. In: Cornelia Müller, Alan Cienki, Ellen Fricke et al. (eds.), 1803-1823.

- Fricke, Ellen (2014b). Between reference and meaning: Object-related and interpretant-related gestures in face-to-face interaction. In: Cornelia Müller, Alan Cienki, Ellen Fricke et al. (eds.), 1788-1802.

- Fricke, Ellen (2014c). Syntactic complexity in co-speech gestures: Constituency and recursion. In: Cornelia Müller, Alan Cienki, Ellen Fricke et al. (eds.), 1650-1661.

- Harrison, Simon (2009). Grammar, Gesture, and Cognition: The Case of Negation in English. Dissertation, Université Bordeaux 3.

- Kendon, Adam (2004). *Gesture: Visible Action as Utterance.* Cambridge, CUP.

- Ladewig, Silva H. (2011). Syntactic and Semantic Integration of Gestures into Speech: Structural, Cognitive, and Conceptual Aspects. Dissertation, Europa-Universität Viadrina, Frankfurt (Oder).

- McNeill, David (1992). *Hand and Mind: What Gestures Reveal about Thought.* Chicago.
- McNeill, David (2005). *Gesture and Thought.* Chicago.
- Müller, Cornelia (1998). Redebegleitende Gesten: Kulturgeschichte – Theorie – Sprachvergleich. Berlin.
- Müller, Cornelia, Jana Bressem, Silva H. Ladewig (2013). Towards a grammar of gesture: A form-based view. In: Cornelia Müller, Alan Cienki, Ellen Fricke et al. (eds.), 707-733.
- Müller, Cornelia, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill and Sedinha Teßendorf (eds.) (2013). *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction* (HSK 38.1). Berlin and Boston.
- Müller, Cornelia, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill and Jana Bressem (eds.) (2014). *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction* (HSK 38.2). Berlin and Boston.
- Wundt, Wilhelm ([1900] 1904). *Völkerpsychologie. Eine Untersuchung der Entwicklungsgesetze von Sprache, Mythus und Sitte.* Vol. 1: Die Sprache. Leipzig.

On the pragmatics of multi-modal face-to-face communication: Gesture, speech and gaze in the coordination of mental states and social interaction

Judith Holler

Judith.Holler@mpi.nl

Coordination is at the heart of human conversation. In order to interact with one another through talk, we must coordinate at many levels, first and foremost at the level of our mental states, intentions and conversational contributions. In this talk, I will present findings on the pragmatics of multi-modal communication from both production and comprehension studies. In terms of production, I will throw light on (1) how co-speech gestures are used in the coordination of meaning to allow interactants to arrive at a shared understanding of the things we talk about, as well as on (2) how gesture and gaze are employed in the coordination of speaking turns in spontaneous conversation, with special reference to the psycholinguistic and cognitive challenges that turn-taking poses. In terms of comprehension, I will focus on communicative intentions and the interplay of ostensive and semantic multi-modal signals in triadic communication contexts. My talk will bring these different findings together to make the argument for richer research paradigms that capture more of the complexities and sociality of face-to-face conversational interaction. Advancing the field of multi-modal communication in this way will allow us to more fully understand the psycholinguistic processes that underlie human language use and language comprehension.

Talks and Posters

Compensation for a large gesture-speech asynchrony in instructional videos

Andrey Anikin, Jens Nirme, Sarah Alomari, Joakim Bonnevier, Magnus Haake

Lund University Cognitive Science, Sweden

rty_anik@yahoo.com, jens.nirme@lucs.lu.se, sjawdat@hotmail.com,
joakim.bonnevier.213@student.lu.se, magnus.haake@lucs.lu.se

Abstract

We investigated the pragmatic effects of gesture-speech lag by asking participants to reconstruct formations of geometric shapes based on instructional films in four conditions: sync, video or audio lag (±1,500 ms), audio only. All three video groups rated the task as less difficult compared to the audio-only group and performed better. The scores were slightly lower when sound preceded gestures (video lag), but not when gestures preceded sound (audio lag). Participants thus compensated for delays of 1.5 seconds in either direction, apparently without making a conscious effort. This greatly exceeds the previously reported time window for automatic multimodal integration.

Index Terms: gesture-speech synchronization, multimodal integration, temporal synchronization, comprehension

1. Introduction

Manual gestures facilitate speech production, evidenced by the fact that they persist when blind people speak among themselves [1] or when the listener is not visible [2]. Furthermore, gestures may improve listening comprehension, especially when speech is ambiguous [3] or when there is a lot of background noise [4]. But how exactly are gestures temporally related to speech? How important is this temporal relation to successful communication?

An influential view is that speech and gesture share a common origin and are best seen as two forms of the same communicative process [5],[6]. Their temporal relationship is determined by the semantic and pragmatic synchrony rules: if speech and gestures co-occur, they must either present the same semantic information or perform the same pragmatic function. It is well established that gestures are generally initiated simultaneously with – or slightly before – the onset of their lexical affiliates [7],[8],[9],[10]. But a new question immediately arises: Are they synchronized because this is necessary for successful comprehension or simply because speech and gesture stem from the same "idea unit"? [5],[6]

One way to answer this question is to see how a disruption of the natural synchronization affects comprehension. Since speech and gesture exploit different modalities, this is a case of multisensory integration, which is affected by the synchronicity of the two channels [11]. Of course, the time-window of tolerance for asynchrony varies depending on the nature of stimuli.

Several studies have found effects of gesture asynchrony on event-related potentials elicited around 400 ms after the onset of a word (N400) indicative of integration difficulty. Habets et al. [12] found a greater N400 to mismatched versus matched gesture-speech sequences only when speech lagged by 0 and 160, but not by 360 ms. The authors conclude that gesture and speech are integrated automatically when they fall within 160 ms of each other, so that a gesture which does not semantically match speech leads to effortful processing. Obermeier and Gunter [13] found an N400 effect for gestures related to either dominant or subordinate meanings of an ambiguous word from approximately -200 ms (speech lag) to +120 ms (gesture lag). Other studies have found a greater perceived emphasis on words when they are synchronized with gestures [10],[14].

A view emerges that gesture and speech may be integrated either automatically or with some conscious effort, depending on how precisely they are synchronized. The window for automatic integration is, however, well within the time frame reported for naturalistic conversations. For example, Morrel-Samuels & Krauss [15] discovered that gestures were never preceded by their lexical affiliates in their data bank, but the onset of gesture usually preceded its lexical affiliate. In fact, the mean reported delay was 1 second, and for less familiar words it could be as long as 3.8 seconds! This result emphasizes the simple fact that we should not underestimate the variability of gesture-speech synchronization in natural conversation. In fact, delays too large for automatic integration may be a normal feature of conversation, for which humans must possess a compensatory mechanism.

Practical implications of gesture-speech lag remain relatively unexplored, partly due to methodological problems with generating naturalistic sequences with mismatched speech and gestures. In particular, lip movements quickly give the manipulation away, unless the face is hidden or computer animation is used to separate facial from bodily movements. Practical implications of gesture-speech synchronization are, however, more relevant today, when digital agents are becoming increasingly common as chatterbots or virtual service desk personnel. Woodall & Burgoon [16] found that actors who purposefully delayed their gestures by up to 1 sec were perceived as less persuasive, and this delay impaired recall. However, in this paradigm speech is not identical in different conditions. Further study of the effects of gesture-speech (de)synchronization on overall comprehension as well as the perceived competence of digital agents is an essential part of the effort to make computer-human communication smooth and effortless. The results could drastically change the way digital agents speak and move.

There is some preliminary evidence that people tolerate much larger speech-gesture delays than the window in which multimodal integration occurs automatically. In a study by Kirchhof [17] 60% of participants accepted gesture-speech pairs as natural with delays from -600 to +600 ms. Furthermore, when asked to synchronize the audio and video tracks, participants chose delays from approximately -1.8 s (gestures first) to +1.2 s (speech first). The author concludes that gestures and speech are more closely synchronized in production than is necessary for successful comprehension. A limitation of Kirchhof's approach is that the perceived naturalness of a clip or the chosen audio-video offset time are both explicit measures that tap into subjective evaluation rather than implicit comprehension. To examine the latter, we would need to assess the pragmatic effects of the multimodal message on observable behavior.

Accordingly, we designed a practical task that required the participant to integrate the visual and the auditory channels, so that performance could be a measure of how successfully speech was integrated with gestures at different time lags. The perceived difficulty of the task and quality of instructions were assessed in a short questionnaire and provide explicit measures of the effects of gesture-speech lag. The main question is how overall comprehension is affected by a large audio or video lag and whether it is associated with subjectively experienced cognitive effort and/or dissatisfaction with the speaker.

2. Methodology

2.1. Participants

83 participants were students recruited and tested at Lund University. Data collection followed the recommendations of *Good Research Practice* by the Swedish Research Council [18] with respect to information to participants, consent, debriefing, confidentiality, and data use.

2.2. Experimental task

Participants were asked to recreate arrangements of five geometric shapes (Figure 1) after watching short videos, in which an instructor spoke and used gestures but did not show the physical objects. The shapes had to be selected from an array of 8 objects: two boxes, two sticks, a ball, a can, a tube, and a small cylinder. The task could be performed incrementally; missing a single step of the instructions did not preclude successful completion of later steps. The videos were presented in one of four conditions (sync, video-lag, audio-lag or audio-only). "Sync" in this case stands simply for original, unmodified clip; the files were not manipulated to ensure perfect synchronization of gesture strokes with their semantic referents. The lag conditions operated with delays of 1,500 ms. This value was chosen based on the results of a pilot study with 11 participants and delays up to +/- 2 s. In the audio-only condition the soundtracks were presented without any accompanying video.

2.3. Materials

Six short instructional videos from 42 to 82 seconds in duration were filmed with a hand-held camera, which was placed high over the shoulder of the instructor so as to provide an unobstructed view of his manual gestures but not his face. The instructor was a male student, who was told that the focus of this study was the effectiveness of communication but was naive to the exact purpose of the study. He was not specifically asked to gesture but encouraged to describe what needed to be done "as well as he could". For each trial, a picture of the target formation was shown on the screen of a smartphone, which the instructor kept in his lap (off camera) while explaining how to build the formation. The recordings were split into separate video- and soundtracks. Each soundtrack was converted to a sample rate of 44100 Hz, filtered to remove background noise, and normalized. In case of mistakes or unwanted noise, such as the sound of a hand slapping the desk, a new take was filmed. All 11 participants in a pilot study confirmed that they could hear the instructions clearly and were not bothered by the camera angle.

One of the authors identified gesture phases in the videos [5]. The gestures produced by the instructor were short in duration (duration: M = 580 ms, SD = 270 ms). The videos contained 124 gestures in total with a gesture stroke on average every 4.80 words. A concern with our video manipulations was the lack of control over what the temporally offset gestures ended up being synchronized with. Even with the large temporal offset used, chances are that gesture strokes still overlap with congruent speech (referring to the same position, orientation or shape of an object as the gesture). We thus categorized the overlapping speech in the manipulated videos as being either congruent or incongruent (overlapping with irrelevant or contradicting speech or silence). The proportion of congruence was very similar in the video lag (40.1%) and audio lag (38.8%) conditions. In the synchronized videos some natural "asynchrony" was present, but generally it was well within the magnitude of the delay introduced in the manipulated videos: 28.2% of the gesture strokes preceded the stressed syllable of their lexical affiliates (median offset 130 ms) and, conversely, 5.6% of the affiliated stressed syllables preceded the onset of the gesture strokes (median offset 80 ms).

2.4. Procedure

At the beginning of the experiment participants were asked to evaluate their ability to read maps (a skill judged to be functionally similar to the demands of the main test). As a practical pre-test, they also had to arrange small pieces of paper "furniture" in the drawing of a room based on verbal instructions. The instructions were read by the experimenter slowly, but without repetitions, and the resulting arrangement was informally assessed on a scale of 1 (poor) to 3 (good).

After this the participants were randomly assigned to an experimental condition (except the audio-only group, which was tested after the others) and started the main experiment, which consisted of six trials. In each trial the participant was asked to reconstruct a formation of five geometric shapes after watching an instructional video presented in *PsychoPy* [19]. Participants were instructed to watch the instruction videos first and then choose and arrange the correct five objects, so that they would not have to divide their attention between the videos and the objects. The reconstructed array was photographed for future coding, and the participant proceeded to the next trial. If a participant could not recall all five objects, they were not pressed to guess but their incomplete arrays were accepted as they were. All trials, except in the audio-only condition, were double-blind: neither participants nor experimenters knew which condition was being tested.

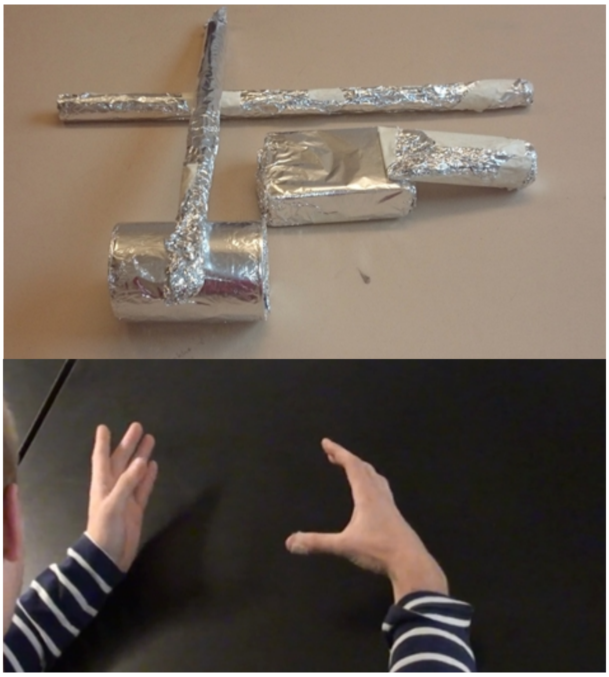


Figure 1. *(Upper) An example of the original formation in trial 6; (Lower) Frame from the (synchronized) instructional video in trial 6, extracted from segment when the instructor describes the position of the rectangular shape.*

After completing six trials, participants filled out a short questionnaire (Table 1) rating the difficulty of the experimental task and the efficiency of the instructor on a visual analogue scale (VAS). They were also encouraged to leave free-text feedback, once after rating the task and once after rating the instructor. Finally, each participant was debriefed and asked whether they had noticed anything strange about the video and sound. If they did not report noticing anything unusual, they were then asked directly whether the video- and soundtracks were synchronized. The entire procedure took 15-20 minutes.

Table 1. *Questionnaire items.*

How difficult was it to understand: (difficult / not difficult) – the instructor's speech? – which shapes to use? – the relations between the shapes? – what to build?
How did you find the instructor: – clear / not clear? – certain / not certain? – professional / not professional?

2.5. Coding

The "furniture" pre-test was coded informally by one of the authors for all participants. All trials were coded independently by two other authors based on an algorithm which awarded points for the correct choice of each object as well as its position in two dimensions, three dimensions, and in relation to the reference object used to describe the location of the object in question. A maximum of 19 points could be awarded for each of the six trials, i.e. maximum 114 points per participant. The coders were (except in the audio-only condition) blind to the experimental condition. Any disagreements in coding were discussed by the two coders, and then either a compromise solution was reached or two different scores were entered in the database and averaged.

2.6. Analysis

All statistical analyses were performed in *R* [20]. Implicit comprehension was operationalized as the total score out of the maximum of 114 on all six experimental trials. The scores from two coders were averaged and rounded to the nearest integer (where different) and modeled with binomial generalized linear mixed models (GLMM) with a random intercept per participant using the *lme4* package [21] and Bayesian modeling with Markov chain Monte Carlo (MCMC) method using *rjags* [22], [23]. Explicit comprehension and satisfaction with the instructor were measured on a VAS and analyzed using ANOVA and *rjags*. Free-text comments were categorized by attitude (neutral/critical) as well as direction (towards the task/the instructor/oneself). Note that commenting was optional and no participant made positive comments.

3. Results

A total of 83 participants (45 females and 38 males) completed the experiment in one of four conditions (Table 2). There were no significant differences between experimental groups in baseline characteristics, such as gender composition ($\chi 2(3, N = 83) = 3.68$, $p = .30$) and the score on the "furniture" pre-test ($\chi 2(6, N = 83) = 2.07$, $p = .91$). ANOVA of self-assessed ability to read maps also failed to discover any effect of condition ($F(3,79) = 0.62$, $p = .60$). Total scores awarded by both coders were very strongly correlated, demonstrating high inter-rater reliability (Spearman's rho: $\rho = .98$). Two participants reported noticing that the audio and video were out of sync; both were in the audio-lag group and both performed extremely well on all trials. Another seven participants (4 in audio-lag and 3 in video-lag condition), when told during debriefing that there might have been a delay, were not sure whether they had noticed it or not; their performance was a bit below average.

Table 2. *Baseline characteristics of study groups*

Group	Number of participants (male / female)	Self-rated ability to read maps ($M \pm SD$)	Pre-test score low/med/high (%)
Sync	20 (12 / 8)	66.7 ± 22.1	5 / 35 / 60
Audio lag 1.5 s	23 (11 / 12)	68.3 ± 24.5	13 / 22 / 65
Video lag 1.5 s	20 (9 / 11)	61.6 ± 23.6	10 / 25 / 65
Audio only	20 (6 / 14)	60.3 ± 20.2	15 / 30 / 55

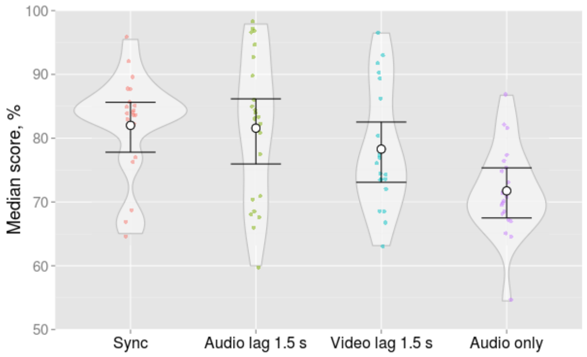


Figure 2. *The distribution of scores for each group (median and 95% credibility intervals).*

3.1. Implicit comprehension

Implicit understanding of the instructional videos was assessed by comparing each reconstructed array with the original and adding up the scores on all trials.

Individual variation of the total score per subject proved to be very considerable, but the overall level of success was high ($M = 78.7\%$, $SD = 10.1\%$). The mean total score per participant in each condition was as follows ($M \pm SD$ as proportion of maximum): sync = 82.4% ± 8.0%, audio lag = 81.9% ± 11.5%, video lag = 78.4% ± 9.7%, audio only = 71.7% ± 7.2% (Figure 2).

According to the MCMC model, there is evidence for higher scores in all three video groups compared to the audio-only group. The most credible difference (median (%) and 95% highest density interval) for sync / audio-lag / video-lag *vs* audio-only conditions is 10.3 [5.0, 16.1], 9.8 [3.3, 16.2] and 6.6 [0.0, 12.3], respectively. In contrast, sync and audio-lag group have essentially the same average scores, while the most credible difference in scores between sync and video-lag groups is only 3.7% [-2.6%, 9.6%]. The video-lag group thus appears to score in between sync and audio, but closer to the former. The difference between all conditions, including audio-only, is small relative to variance within each condition, which translates into low statistical power. A retrospective power analysis shows that we were 84% likely to prove that all 3 video conditions exceed the audio-only condition, 55% likely to prove the difference between the video-lag and audio-only conditions, and only 23% likely to prove the difference between the sync and video-lag conditions.

Naturally, performance on the experimental task may be strongly affected by the individual spatial abilities of each participant, and the effect of condition may depend on these abilities. GLMM models were therefore fitted to investigate possible interactions between condition and each of two measures of underlying spatial ability: (1) the direct question ("How do you evaluate your ability to read maps?") and (2) the score on the "furniture" pre-test, in which the participant had to arrange furniture based on verbal instructions. The interaction between self-rated spatial ability and experimental condition is strong (likelihood ratio test: $L = 14.1$, $df = 3$, $p = .003$). The same holds for the score on the "furniture" pre-test ($L = 15.8$, $df = 3$, $p = .001$). Better results on the pre-test thus predict higher scores on the main task, but primarily in the audio-lag condition.

3.2. Explicit comprehension

Individual scores on the four questions related to the difficulty of the task are strongly correlated (Cronbach's alpha = .86), therefore they were combined and analyzed as a single item, with a significant main effect of condition in ANOVA: $F(3,79) = 12.1$, $p < .001$. The overall rating of task difficulty was higher in the audio-only condition compared to any other condition (the most credible difference is 27% [18%, 36%]). The evidence for any difference between the video conditions is very weak (the highest-density intervals include zero for each comparison). The task was thus judged to be considerably more difficult by participants in the audio-only group, but with no difference between the three video groups (Figure 3).

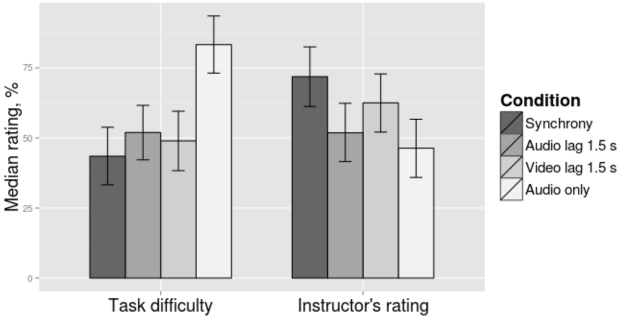


Figure 3. *Subjective ratings of the difficulty of the tasks and satisfaction with the instructor (median and 95% credibility intervals).*

As for the three questions in which participants rated their satisfaction with the instructor, scores on the individual items were also strongly correlated (Cronbach's alpha = .93). These three questions were therefore combined. There is a noticeable main effect of condition on the combined score on these three items ($F(3,78) = 4.6$, $p = .006$). Compared to sync condition, the instructor received lower ratings in the audio-only and audio-lag conditions (the most credible difference is 24.9% [10.8%, 41.2%] and 20.3% [6.1%, 36%], respectively). The uncertainty is high, but it appears that satisfaction with the instructor was highest in the sync condition, lowest in the audio-only condition and intermediate in the audio/video-lag conditions (see Figure 3).

Participants provided in all 77 free-text comments (out of 168 opportunities). As can be seen in table 3, the distribution of comments of different types across conditions was not uniform. Comments directed towards the difficulty of the task (e.g. "that was a lot of information") were rare, and critical comments directed towards the instructor (e.g. "he did not seem to know what he was doing") more frequent in the audio-lag and video-lag conditions. The participants in the audio-only group were more likely to direct criticism towards themselves (e.g. "I had trouble keeping all that information in my head").

Table 3. *Number of free-text responses classified as neutral or critical per group and comment direction.*

Directed towards	Number of comments, neutral / critical			
	Sync N=20	Audio lag N=23	Video lag N=20	Audio only N=20
Task	0 / 2	0 / 1	1 / 0	1 / 3
Instructor	0 / 1	4 / 9	1 / 9	4 / 5
Self	3 / 1	5 / 0	5 / 2	4 / 9

4. Discussion

As Woodall and coauthors long ago pointed out, it is important to establish how closely verbal and nonverbal behaviors are synchronized during communication and describe the nature of this synchronization process, but *"an equally important issue is how it affects communication outcomes such as information exchange and persuasion"* [16]. The latter point has been largely neglected since, but today the ubiquity of digital agents makes this straightforward question of great practical significance: what degree of gesture-speech desynchronization is tolerated before communication breaks down and/or the receiver gets annoyed?

The task used in this study was designed to be solvable only if the audio and video channels are integrated. The fact that scores in the audio-only condition were significantly lower than in the three video conditions (full sync, audio lag 1.5 s, and video lag 1.5 s) indicates that both modalities were needed to solve the task. Our result does not reveal that a delay of 1.5 seconds in either direction prevents the receiver from integrating gestures with speech, despite a weak tendency for lower performance in the video-lag group. Furthermore, compared to the sync condition, the task was rated as considerably more difficult in the audio-only condition but not in the audio/video lag conditions. Not only could the participants integrate gestures and speech despite the large delay, but they did so without experiencing the task as more difficult. However, in contrast to the ratings of the *task*, ratings of the *instructor* were affected by delay, as the instructor in audio-lag condition was rated as worse than in the sync condition and almost as low as in the audio-only condition. More free-text criticism was also directed towards the instructor in both the audio- and video-lag conditions. Unexpectedly, criticism was less likely to be directed towards the instructor when he was not visible, despite the low VAS ratings that he received in this condition.

Clearly, individual variation in spatial abilities may influence the results. Indeed, we discovered a highly significant interaction between both measures of spatial skills (self-evaluated ability to read maps and performance in the "furniture" pre-task) and experimental condition. Participants with good spatial skills in the audio lag group were able to fully compensate for the temporal mismatch, while those with poor spatial skills were unable to compensate and performed worse compared to participants in the sync group. Intriguingly, spatial skills had very little effect on performance in the video-lag condition and none at all in the audio-only and full-sync conditions. Given the small sample sizes, this difference could be spurious, or it could indicate that certain cognitive skills are involved in compensating for the lack of synchrony which are not manifest in other experimental conditions.

On the one hand, it is somewhat surprising that the participants could compensate relatively successfully for such a large delay as 1,500 ms, when previous studies have found that the time window for automatic integration spans no more than a few hundred milliseconds [12],[13]. It is especially impressive when the audio track is advanced relative to the video track - the "atypical" direction, since speech hardly ever precedes gestures in natural conversation [10],[14].

On the other hand, integration of visual and auditory stimuli with very large delays has been reported before. In a study of the McGurk effect, Campbell and Dodd [24] presented participants with short words using audio lags of 400, 800 and 1600 ms. Phoneme identification was optimal in the full sync condition, but even at the longest delay identification was better compared to the audio-only control condition. In a recent series of studies Kirchhof [17] discovered that surprisingly large temporal mismatches of gestures were accepted as natural.

An important question to ask pertains to the mechanisms of cross-modal integration at these longer delays. What exactly happens if gestures and speech are poorly matched temporally and fail to be integrated "automatically" back into a single "idea unit"? An influential position in psychology invokes the notion of "mental models" [25] or "situation models" [26],[27] – holistic representations of the described situation, which are integrated across sentences, modalities, sometimes even languages and multiple documents or conversations. The temporal structure of messages is not always linear. Grammatical rules being what they are, the order of events in a narrative does not always correspond to the order in which they are mentioned in a sentence: for instance, we may say: "Before I opened the door, I had to search for my keys for a few minutes". Seen in this light, a gesture-speech lag of a second or so is a special case of integrating information arriving from different modalities and at different times into a unified situation model. In line with Massaro's "fuzzy logical model of perception" [28], the two modalities will probably be integrated as long as they are perceived as belonging to the same perceptual event. Then again, gesture and speech can hardly be attended to as two completely independent channels. Instead, it seems likely that speech sets up a context for interpreting gestures, and vice versa [29]. This integration may not be automatic, but judging by the rating of task difficulty in the four groups, it requires very little conscious effort.

A limitation of the task used in this study is that both average performance and individual variability were high, making it harder to detect differences between groups. In other words, the auditory channel alone contained enough information for some participants to perform near the ceiling, while others struggled even when presented with unmanipulated videos. As a result, it is hard to be certain whether the tendency for lower scores in video-lag compared to sync condition is an artifact. It would be desirable to try other experimental tasks, in which the informational load of gestures is higher.

Similarly, the tendency for somewhat lower satisfaction with the instructor in the audio/video lag conditions is suggestive, but the evidence is inconclusive. An independent measure of effort could help reveal if this tendency stems from an increased effort manifested as frustration with the instructor without attribution of difficulty to the task itself. Given the high natural variability in gesture-speech temporal coordination, the strokes of the instructor's gestures did not necessarily have a tight temporal coupling with their lexical referents in the original unmanipulated videos. In fact, despite the large temporal offset, around 40% of the gestures in manipulated videos still overlapped with semantically congruent speech (although this effect was balanced between the video-lag and audio-lag conditions). The stimuli also included instances of spoken deictic expressions referencing the gestures ("this", "here"). In these cases instructions were clearly incomplete when the associated gestures were missing in the desynchronized videos. Eliminating congruent overlap and such obvious mismatches by a strict selection of instruction videos from a larger set might reveal effects that our results did not.

In summary, this study investigated whether desynchronized speech and gestures can still communicate task-relevant information. The answer, at least for the task investigated here, is a clear yes. Not only is compensation nearly perfect, but the participants fail to notice a delay of 1.5 seconds in either direction and do not make a conscious effort to integrate desynchronized gestures and speech. Asynchrony may, however, cause the speaker to appear less competent. Many issues, such as the generalizability of this outcome, the nature of integration processes and the cognitive skills involved, await further research.

5. References

[1] Iverson, J. M., and Goldin-Meadow, S., "Why people gesture when they speak", Nature, 396(6708):228-228, 1998.

[2] Wagner, P., Malisz, Z., and Kopp, S., "Gesture and speech in interaction: An overview", Speech Communication, 57:209-232, 2014.

[3] Thompson, L. A., and Massaro, D. W., "Evaluation and integration of speech and pointing gestures during referential understanding", Journal of Experimental Child Psychology, 42(1):144-168, 1986.

[4] Rogers, W. T., "The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances", Human Communication Research, 5(1):54-62, 1978.

[5] Kendon, A., "Gesticulation and speech: Two aspects of the process of utterance", in M. Key [Ed], The Relationship of Verbal and Nonverbal Communication, 207-227, Mouton, 1980.

[6] McNeill, D., "So you think gestures are nonverbal?", Psychological Review, 92(3):350-371, 1985.

[7] McNeill, D., "Hand and mind: What gestures reveal about thought", University of Chicago Press, 1992.

[8] Nobe, S., "Where do most spontaneous representational gestures actually occur with respect to speech?", in D. McNeill [Ed], Language and Gesture, 186-198, Cambridge University Press, 2000.

[9] Loehr, D., "Aspects of rhythm in gesture and speech", Gesture, 7(2):179-214, 2007.

[10] Treffner, P., Peter, M., and Kleidon, M., "Gestures and phases: The dynamics of speech-hand communication", Ecological Psychology, 20(1):32-64, 2008.

[11] Liu, B., Jin, Z., Wang, Z., and Gong, C., "The influence of temporal asynchrony on multisensory integration in the processing of asynchronous audio-visual stimuli of real-world events: an event-related potential study", Neuroscience, 176: 254-264, 2011.

[12] Habets, B., Kita, S., Shao, Z., Özyurek, A., and Hagoort, P., "The role of synchrony and ambiguity in speech–gesture integration during comprehension", Journal of Cognitive Neuroscience, 23(8):1845-1854, 2011.

[13] Obermeier, C., and Gunter, T. C., "Multisensory Integration: The Case of a Time Window of Gesture–Speech Integration", Journal of Cognitive Neuroscience, 27(2):292-307, 2015.

[14] Krahmer, E., and Swerts, M., "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception", Journal of Memory and Language, 57(3):396-414, 2007.

[15] Morrel-Samuels, P., and Krauss, R. M., "Word familiarity predicts temporal asynchrony of hand gestures and speech", Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(3):615-622, 1992.

[16] Woodall, W. G., and Burgoon, J. K., "The effects of nonverbal synchrony on message comprehension and persuasiveness", Journal of Nonverbal Behavior, 5(4):207-223, 1981.

[17] Kirchhof, C., "Desynchronized speech-gesture signals still get the message across", in the 7th International Conference on Multimodality (7ICOM), Hongkong, 2014.

[18] The Swedish Research Council, "God forskningssed", Vetenskapsrådets rapportserie, 1:2011, 2011.

[19] Peirce, J., PsychoPy – Psychophysics software in Python. Journal of Neuroscience Methods, 162(1-2):8-13, 2007.

[20] R Core Team, "R: A language and environment for statistical computing (R version 3.1.3)", [Statistical software], R Foundation for Statistical Computing, 2015. Available: www.R-project.org/

[21] Bates D., Maechler, M., Bolker, B., and Walker, S., "lme4: Linear mixed-effects models using Eigen and S4 (R package version 1.1-7)" [Statistical software], 2014. Available: CRAN.R-project.org/package=lme4.

[22] Plummer, M., "rjags: Bayesian graphical models using MCMC (R package version 3-14)", [Statistical software], 2014. Available: CRAN.R-project.org/package=rjags

[23] Kruschke, J., "Doing Bayesian data analysis: A tutorial introduction with R", Academic Press, 2010.

[24] Campbell, R., and Dodd, B., "Hearing by eye", Quarterly Journal of Experimental Psychology, 32(1):85-99, 1980.

[25] Johnson-Laird, P. N., "Mental models: Towards a cognitive science of language, inference, and consciousness", Harvard University Press, 1983.

[26] Van Dijk, T. A., and Kintsch, W., "Strategies of discourse comprehension", Academic Press, 1983.

[27] Zwaan, R. A., and Radvansky, G. A., "Situation models in language comprehension and memory", Psychological bulletin, 123(2):162-185, 1998.

[28] Massaro, D. W., Cohen, M. M., and Smeele, P. M., "Perception of asynchronous and conflicting visual and auditory speech", The Journal of the Acoustical Society of America, 100(3):1777-1786, 1996.

[29] Kelly, S. D., Barr, D. J., Church, R. B., and Lynch, K., "Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory", Journal of Memory and Language, 40(4):577-592, 1999.

Polyfocal class interactions and teaching gestures. An analysis of nonverbal orchestration

Brahim Azaoui

Aix-Marseille Université, CNRS, LPL UMR 7309 Aix-en-Provence, France
brahim.azaoui@univ-amu.fr

Abstract

While a growing body of research suggests that gestures have an impact in the teaching/learning process, few have explored gestures produced by teachers to understand how instructors cope with the intrinsically polyfocal dimension of class interactions. This paper reports on an empirically grounded account of both how and in what circumstances teachers conduct multimodal orchestration, and the interactional issues it raises. Because it is based on video-recorded corpora of two instructors each teaching both French to native and to non-native students, my study also tackles the issue of the context-sensitivity of teaching gestures.

Index Terms: teaching gestures, two-handedness, co-enunciative ubiquity, context, nonverbal orchestration

1. Theoretical framework

1.1 Teaching gestures

A growing body of research has tackled the topic of teaching gestures in instructional and non-instructional contexts. These studies have mostly shown the impact of teaching gestures in different areas of the learning process. For example, we can consider the role of gesturing in the comprehension of math instructions or math problems ([1], [2], [3], [4]). Alibali et al. [3] for instance provided a math teacher with a tutorial about ways to use gestures in connecting ideas in instruction. The results demonstrate that students benefit more from the teacher who expresses linked ideas using both gestures and speech than from a teacher who does not. In language teaching contexts, a range of research has examined the impact of gestures in L1 or L2 teaching and learning ([5], [6], [7], [8], [9], [10], [11], [12]). In an empirical study Sime [10] sought to understand what learners made of their teachers' gestures. She showed that they made a distinction between relevant and irrelevant gestures among those that their teachers produced, and they were able to attribute the relevance of these nonverbal actions within the learning process as they enhanced comprehension and provide feedbacks. Others have considered more specific aspects, like the role of gestures in memorization ([13], [14], [15]) or error correction ([16], [17]). For example, Tellier [13] experimentally examined the impact of gesture on second language memorization in teaching vocabulary to 5 year-old learners. She showed how the teacher's gestures, and especially their reproduction by the learners helped the latter remember the words they were taught. Muramoto [16] considered the role of gestures in providing error correction so as to contribute to students' successful self-correction. He analyzed the gestures of three instructors in a university Japanese second language classroom and distinguished two sorts of gestures in class: specific language error correction gestures and general foreign language classroom gestures.

Yet, despite this impressive body of research, it seems that few studies have been interested in considering the gestures as a way for teachers to organize class turn-taking and deal with overlapping talks ([18], [19], [20]) rather than a means to enhance learning. Azaoui's empirical study [20] is based on a mimo-gestural analysis of both a corpus of filmed classroom interactions led by the same teacher in two different instructional contexts (French to native students and to non-natives) and video-recordings of students confronted with extracts of lessons they participated in. He sought to understand how, when and why the teacher reacts to the students' disruption of the interactional norms, but also how and why the students break this conversational organization [21]. The results show that the teacher's motivations are twofold: the instructor's verbal and nonverbal actions contribute both to the progress of the lesson plan and the prevention of threats to the students' face [22].

1.2 Classroom polyfocal interactions

Coping with multiple simultaneous actions is the reality of many teachers in classroom. Thus, it seems more accurate to consider classroom interactions as typically "polylogal" [23] (i.e., more than three persons usually speak at the same time; consequently interventions may overlap) - rather than looking at them as if they followed a regular three-part pattern [24]. If "trilogues are potentially more conflicting organizations than dialogue" [25:6] because participants may struggle even more for the floor, one can easily imagine what the situation may be like during polylogues where intrusions and overlapping turns may occur more spontaneously and frequently. In addition, classroom interactions can be said to be polyfocal as several foci of interaction may simultaneously take place [26:66]. Consequently, there is barely a moment when teachers do not produce several gestures at the same time (head/hand gestures, right hand/left hand gestures). So, as much as we can say that students have a polyfocal attention, to the extent that they very rarely "direct their attention in a focal, concentrated way to any single text or medium" (Scallon et al, cited in [27:28]), teachers' attention can also be qualified as being polyfocal. Since they have to manage various actions at the same time, Kress proposed the term "orchestration" to name the "process of assembling/organizing/designing a plurality of signs in different modes into a particular configuration to form a coherent arrangement" [28:162]. If we pay attention to the way this orchestration is conducted, we can notice that it takes various forms and has implications for the interactional process.

These are the issues this paper proposes to tackle. It sets out to provide an empirically grounded account of both how and in what circumstances teachers conduct this orchestration, and what the interactional issues are.

I will first present the methodology of this research. Then, I will examine the results in two separate but complementary sections: I will explore the notion of *two-handedness*, understood as the production of two-handed independent gestures, and that of *co-enunciative ubiquity*, which refers here to the teacher's nonverbal ability to be the co-utterer with at least two students simultaneously.

2. Methodology

2.1 Participants

My research is based on the analysis of two native French secondary school teachers from the South of France (Toulouse and Montpellier). They both teach French to native learners (FL1) and French to non-native students (FL2). The initial idea was to analyze how these teachers dealt with school norms (i.e., linguistic and interactional norms) according to the contexts and students they taught.

The Toulouse teacher's French students were aged 14 whereas the Montpellier teacher's were 11. Both had 28 students per class on average. As for their non-native students, the classes they teach gather students from different origins and ages. In Toulouse, the class consisted of 12 different nationalities. The average age of the non-native students was 12.5 while Montpellier's FL2 class was composed of non natives aged 13 or so who came from 4 different countries.

2.2 The corpora and the coding

To carry out this study the data were gathered empirically ([29], [30], [31]) by filming each teacher in action in her two classes. I recorded some 20 hours of classroom interactions among which 6h30 were fully transcribed and coded using ELAN [32].

It included the transcribing of the speech of the teachers and the students on separate tiers, the annotating of the teachers' gesture dimensions, and the annotating of their mimics. I designed my typology of gesture and mimic dimensions and functions based on various works ([33], [34], [35]).

As far as gestures were concerned, I annotated emblems, deictics, metaphorics, beats, and iconics. As for the facial mimics, I coded the following dimensions: orientation of the gaze, frown, raise eyebrows, smile, nod, tilt. Combinations of two or three of these facial movement dimensions were possible. Following Tellier's typology [35], I considered three main teaching gestures functions: informing, managing and assessing. I adapted the latter considering that it also concerned assessing the way students took the floor in compliance or not with school rules [36].

2.3 The analysis tools

I mostly draw my analysis tools from the talk-in-interaction framework espoused by Kerbrat-Orecchioni [37]. The author emphasizes on the need to analyze interactions by merging theoretical tools proposed by discourse and conversational analysis, which implies calling upon Goffman's interactional approach, ethnography of communication and language act theory. This stance may seem to combine incompatible theories (*e.g.* language act theory and conversational analysis), yet according to the author only the combination of these approaches will facilitate a thorough understanding of the embodied (inter)actions. This approach generated the following results.

3. Results

It is possible to distinguish two aspects of nonverbal orchestration: two-handedness and co-enunciative ubiquity. Both will be studied in the following lines.

3.1 Two-handedness, one mode yet two functions

Two-handedness will not be understood here as the use of the two hands to produce a single gesture serving one of the three previously mentioned functions [38]. Rather, as each hand may generate gestures occurring within separate gesture units, the two hands may produce two different dimensions to serve two independent and complementary functions.

In the first example, the class is talking about the 2012 French elections for presidency. The word "debate" has come up during the discussion and non-native students are trying to define the word. This episode illustrates how, in less than 4 seconds, two-handedness can be used to assess a student's intervention and allocate the next turn to another student:

	Corpus M-FL2	
1	T	debate +++ what does this word mean↑
2	Nolan	I don't know
3	T	you don't know↑
4	Antonio	*I* know
5	T	you know↑ ok we're listening to you
6	Antonio	like uhm:::
7	Nolan	two persons
8	Antonio	the the persons speak
9	T	persons speak↑
10	Nolan	some ++ some some
11	T	you're almost there good you've got it
12	Nolan	**[**some some
13	Antonio	many things uhm:::: one thing X
14	T	but more precisely + go ahead (*to Nolan*)↑
15	Nolan	when ++ two ++ persons speak**]** about a topic
16	T	exactly ++ exactly two persons talking about the SAME topic

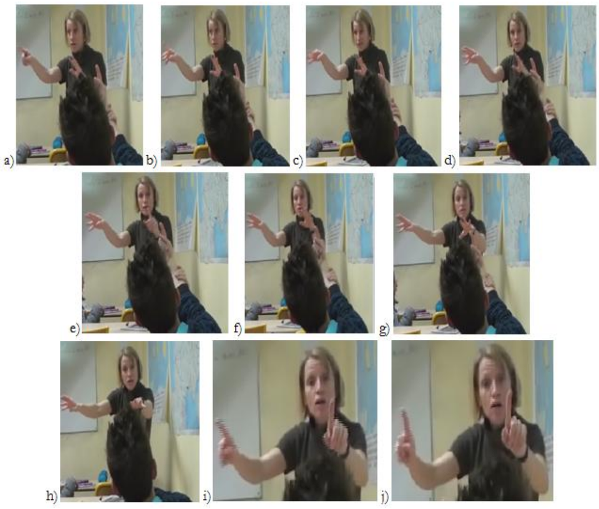


Figure 1: *Two-handedness in FL2 context, turns 12-15.*

Frames *a* to *d* illustrate the teacher producing an emblem with her right hand to assess the intervention of Antonio (turn 13), who is interrupted in turn 10 by Nolan at whom the teacher nevertheless points her left hand to give the floor (frames *e-g*). Interestingly, the teacher keeps her right hand oriented towards Antonio as if not to break the interaction initiated with him. This enables her both to build an interpersonal relationship with the two students and to accomplish shared understanding. She then retracts her right hand to mime the verbal explanation given by Nolan to whom she finally pays full attention as illustrated by the orientation of her head, gaze, body and hands (frames *i-j*).

The second example is extracted from the French to native instructional context. As the teacher is explaining the functioning of end-of-term school reports, a student (Loubna) interrupts her.

	Corpus M-FL1	
1	Youssef	Hum:::: are we not handed over the end-of-term school report after the second term teachers' conference
2	T	No ++ during the meeting the teachers give their opinion + we talk about the student **[**and then//
3	Loubna	there are the class reps, too
4	T	we give
5	Serge	there only is XX
6	T	Hush, will you please not intervene (*to Loubna*) ++++ and the hea**]**d teacher, in other words me, writes down this + the decision + ok

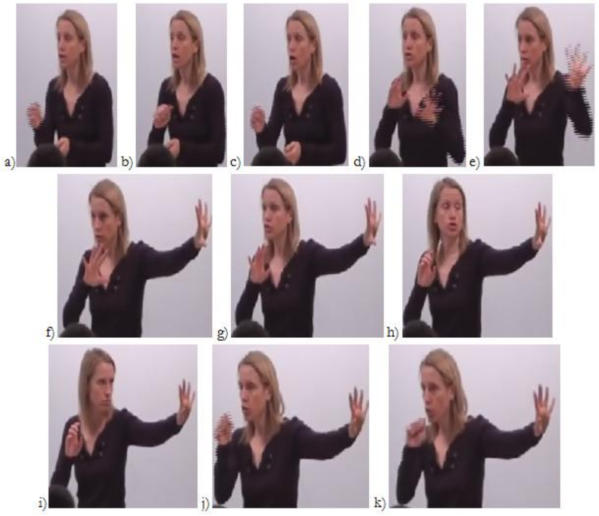


Figure 2: *Two-handedness in FL1 context, turns 2-6.*

Frame *a* shows the teacher producing an iconic gesture that was meant to accompany her verbal explanation now postponed in turn 6 ("write down"). She is interrupted in her verbo-gestural explanation by Loubna, which accounts for the emblem she produces with both hands to ask the student to stop speaking (frames *d* to *g*). This pragmatic function is emphasized by the fixed gaze illustrated in frame *i*. She holds her left arm extended to literally keep the student at bay while she resumes her verbal-gestural explanation where she had previously left it. The two-handedness complementary functions are obvious in frames *j* and *k*: her right hand produces an iconic gesture to inform the students about the functioning of end-of-term school reports, and her left arm prevents Loubna from speaking.

An interview I had with this teacher opens an enhanced window onto this gestural action. She explained how useful this two-handedness was both on a pedagogical level to organize simultaneous interactions and on a more personal psychological perspective since it helped her relieve her voice and the inner turmoil she felt.

3.2 Shift of attention and co-enunciative[1] ubiquity

Nonverbal orchestration is made even more evident when teachers' actions are analyzed in a combined approach of deictic gestures and gaze. In this paragraph I will examine how the interplay of these media enables the teacher to "multiply" herself so as to be the co-enunciator of several students almost simultaneously. This ability, which I termed *co-enunciative ubiquity* [39], is illustrated in the following examples. They will enable me to demonstrate that besides the interpersonal relationship it helps to build, this ability has an impact on the interaction level.

This first extract of class interactions follows an excursion the FL2 class had to the theatre the previous week. The teacher is not pleased with the behavior her students had, and she wants them to reflect over their attitude.

	Corpus T-FL2	
1	T	the problem already happened in class
2	Omar	I know, Miss
3	T	yes
4	Ericka	not quarrel
5	Maria	no right to **[**use the cellphone**]**
6	Omar	XXX

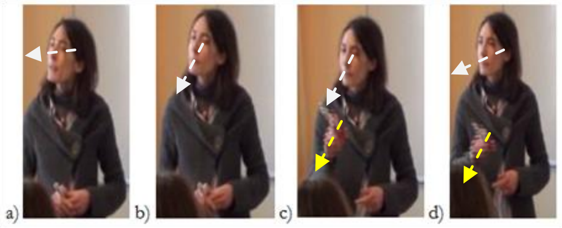


Figure 3: *Co-enunciative ubiquity in FL2 context, turns 2-6.*

Three students speak out almost simultaneously. The instructor's initial gaze orientation (frame *a*) informs us about the attention she pays to the utterance of a student (Omar) seated at the back of the class. At the same time, Ericka's overlapping turn makes the teacher orient her gaze towards her student and produce a deictic gesture to indicate the interest she gives to her idea (frames *b* and *c*). This is confirmed by the superimposed beat gesture (frames *c* and d). Finally, as she retracts her pointing gesture, she briefly looks at Maria, who is acknowledged as a co-participant of the interaction (frame *d*). This description aims to progressively unroll the multimodal teacher's action and to show how this teacher copes with the intrinsic polyfocal and polylogal dimensions of class interactions.

The following example taken from the FL1 class enables us to pursue the demonstration of the teacher's co-enunciative ubiquity and its implications. Here, the teacher is working on a short story about totalitarianism.

[1] The notion "co-enunciative" insists on the simultaneous work of both participants of the interaction [40:44].

First, she asks her students to describe the image they have of the characters in the story. She then overtly allocates the turn to one specific student, as confirmed by the use of the student's name and the orientation of her gaze (frame *a*). An overlapping intervention coming from the left side of the class draws her attention and makes her briefly shift her head and eye orientation towards another student, Albert (frames *b* and *c*).

	Corpus T-FL1	
1	E	so why do you think the character is about fifty years old (*to Pierre*)↑
2	Albert	[he's the average man in the street
3	Pierre	no + I don't know + about fifty or sixty I don't have a clue
4	E	XX ++ yes Albert] a little louder
5	Albert	he's the average man in the street
6	E	right ++ he's the average man in the street

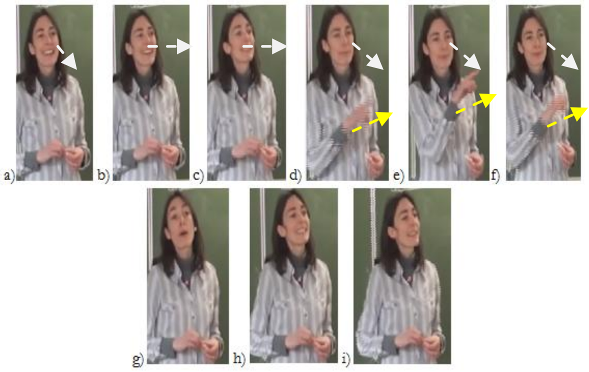


Figure 4: *Shift of attention and co-enunciative ubiquity in FL1 context.*

While considering the frames, it is important to remember that "no one would dispute the close connection between movements of our eyes and shifts of attention" [41:5], no matter how restricted it may be. Posner [42:26] subdivided attention into three separate but interrelated functions: "(a) orienting to sensory events; (b) detecting signals for focal (conscious) processing, and (c) maintaining a vigilant or alert state". The first one is of some particular interest for our understanding of the interaction under study. Indeed, Lamargue-Hamel [43:10] explains that orienting to sensory events is implied in the selection and focalization of relevant pieces of information in a given task. Consequently, it is possible to give the teacher's re-orientation of her gaze and head an intentional purpose that serves her pedagogical interest. It also illustrates the ability to divide her auditory attention: she seems to be constantly filtering external stimuli according to their relevance for the current interaction. Additionally, frames *d*, *e* and *f* illustrate the almost simultaneous combined gesture/gaze disjunction. As her gaze comes back to focusing on Pierre she starts a pointing gesture with her right hand indicating Albert at the back of the class. The beat she produces on her deictic gesture (frame *e*) informs us about the relevance of his intervention.

The first analysis we can make is that this action exemplifies the instructor's ability to pay attention to (at least) two students at the same time. Additionally, the two channels have two separate functions: her gaze has a managing function (attributing the turn) while her pointing gesture assesses Albert's utterance. A second analysis concerns the instructional technique the teacher uses. It corroborates the divided attention we mentioned since the co-enunciative ubiquity she performs helps her select the utterance that best fits her lesson planning. Note that the hand gesture may also serve as a way to "provide the recipients with a 'forward-understanding', i.e., an anticipation, of what will come next" [44:226]. In other words, it anticipates the following exchange with Albert; and the other students are thus informed about the next locus of interest.

This nonverbal action also has consequences on the interactional level. Indeed, research on interaction has often recognized the use of gaze as a means to indicate the ratified interlocutor ([45], [46], [47]). It is here confirmed by the teacher's use of the name Pierre to overtly designate her privileged interlocutor. Yet, the combined analysis of the gesture/gaze disjunction and the teacher's utterance tells us what is really at stake in the extract. An interpretation that can be hypothesized is that this hand gesture/gaze action entails a "communicational trope" [45:92], i.e., the inversion of the hierarchy of the interlocutors. Pierre's utterance loses its interest, the teacher hardly paying attention to the end of his sentence (turn 4). Right from the beginning her attention is polarized by Albert's intervention which is more in compliance with what she wanted her students to understand and keep in mind.

4. Conclusion

To summarize, in this paper I have focused on how teachers resorted to multimodal resources to cope with polyfocal classroom interactions which require organizing turn-taking, informing, and assessing several students simultaneously.

I first explored the production of two-handed independent gestures. The results show that they serve distinctive yet complementary teaching functions: assess verbal proposal and allocate turn, or inform and assess unauthorized intervention. By producing two independent gestures, the teacher is able both to build an interpersonal relationship and progress in her lesson plan. The teacher's comments that I collected during an interview enabled to expand this analysis. They draw our attention to the importance of two-handedness on a more intrapersonal and psychological level.

Secondly, I have examined the nonverbal orchestration a step further by investigating the production of hand gestures in collaboration with gaze orientation. I have paid attention to what I termed co-enunciative ubiquity, i.e., the multimodal ability to manage polyfocal and polylogal class interactions. The interplay of gaze and deictic gestures also served the teacher's intention to have students anticipate the next focus of attention. Additionally, reference to attention theory enabled me to show how this ability attested the fact that the teacher selected the intervention that best suited her pedagogical purpose. This was confirmed by the interactional consequence of this multimodal action, namely a reversal in the hierarchy of the addressed which follows a teaching goal: showing interest to the most appropriate answer.

Interestingly, the results also show that the instructional context has no impact on how the teacher handles this nonverbal orchestration. Two-handedness and co-enunciative ubiquity compose each instructor's "teaching style" ([48], [36]). This term refers to the fact that while some teaching actions may be adapted to the specificity of a given context, others may be recurrent from one pedagogical context to another both in the form they take and in their pedagogical intent. These unvaried actions compose the "teaching style" of some teachers. In this perspective, and as far as our teachers are concerned, no matter the instructional context (FL1 or FL2), there is no difference neither in the way they conduct this orchestration nor in the motivations behind it. I believe these examples of orchestration are not specific to the language teaching classes and may be observed also in other instructional contexts.

Finally, this study corroborates the need to analyze teaching gestures in natural teaching contexts. It enables the opening of an enhanced window onto the complexity of teachers' nonverbal actions.

5. References

[1] Goldin-Meadow S., Nusbaum, H., Kelly, S. D. and Wagner, S, "Explaining math: Gesturing lightens the load", Psychological Science, 12:516-522, 2001.

[2] Alibali, W. M., Young A. G., Crooks, N. M., Yeo A., Wolfgram, M. S., Ledesma, I. M., Nathan, M. J., Breckinridge Church, R and Knuth, E. J., "Students learn more when their teacher has learned to gesture effectively", Gesture, 13(2):210–233, 2013.

[3] Hostetter, A. B., Bieda, K., Alibali, M. W., Nathan, M. and Knuth, E. J. "Don't just tell them, show them! Teachers can intentionally alter their instructional gestures", in Sun, R. [Ed.], Proceedings of the 28th Annual conference of the cognitive science society, 1523-1528, Mahwah, NJ.: Erlbaum, 2006.

[4] Brookes, H., Colletta, J.-M. and Ovendale, A., "Polysigns and information density in teachers' gestures". Actes du colloque international TiGeR, Tilburg, 2013.

[5] Stam, G., "Second language acquisition from a McNeillian perspective", in Duncan, S. D., Cassell, J. and Levy, E. [Eds.], Gesture and the dynamic dimension of language: Essay in honor of David McNeill, 117-124, Amsterdam: John Benjamins Publishing, 2007.

[6] Colletta, J.-M., Le développement de la parole chez l'enfant âgé de 6 à 11 ans. Corps, langage et cognition, Sprimont: Mardaga, 2004.

[7] Lazaraton, A., "Gestures and speech in the vocabulary explanations of one ESL teacher: a microanalytic inquiry", Language learning, 54(1):79-117, 2004.

[8] Gullberg, M., Gesture as a communication strategy in second language discourse. A study of learners of French and Swedish. Malmö: Lund University Press, 1998.

[9] Gullberg, M. and McCafferty, S. G., "Introduction to gesture and SLA: toward an integrated approach", SSLA, 30:133-146, 2008.

[10] Sime, D., "'Because of her gesture, it's easy to understand'. Learners' perception of teachers' gestures in the foreign language class", in McCafferty, S. G. and Stam, G. [Eds.], Gesture: second language acquisition and classroom research 259-279, New-York, NY.: Routledge, 2008.

[11] Tellier, M., "Faire un geste pour l'apprentissage: le geste pédagogique dans l'enseignement précoce", in Corblin, C. and Sauvage, J. [Eds.], L'enseignement des langues vivantes étrangères à l'école. Impact sur le développement de la langue maternelle, 31-54, Paris: L' Harmattan, 2010.

[12] Goodrich, W. and Hudson Kam, CL., "Co-speech gesture as input in verb learning", Developmental Science, 12(1):81-87, 2009.

[13] Tellier, M., "The effect of gestures on second language memorisation by young children", Gesture, 8(2):219-235, 2008.

[14] Allen, L. Q., "The effects of emblematic gestures on the development and access of mental representations of French expressions", The Modern Language Journal, 79:521-529, 1995.

[15] Kelly, S., DeVitt, T. and Esch, M., "Brief training with co-speech gesture lends a hand to word learning in a foreign language", Language and Cognitive Processes, 24(2):313-334 2009.

[16] Muramoto, N., "Gesture in Japanese language instruction : the case of error correction", in Heilenmann, L. K. [Ed.], Research issues and Language Program Direction, 143-175, Boston: Heinle and Heinle, 1999.

[17] Taleghani-Nikazm, C., "Gestures in foreign language classrooms: an empirical analysis of their organization and function", in Bowles, M., Foote, R., Perpiñan, S. and Bhatt, R. [Eds.], Selected proceedings of the 2007 second language research forum, 229-238, Somerville, MA: Cascadilla, 2008.

[18] Kååntå, L., "Pointing, underlining and gaze as resources of instructional action in classroom interaction", Interacting bodies, 2d ISGS Conference proceedings, 2005.

[19] Foester, C., "Et le non-verbal ?" in Dabène, L., Cicurel, F., Lauga-Hamid, M.-C. and Foerster, C. [Eds.], Variations et rituels en classe de langue, 72-93, Paris: Hatier, 1990.

[20] Azaoui, B., "Déritualisation et normes interactionnelles. Entre stratégie pédagogique et revendication identitaire", in Miecznikowski, J., Casoni, M., Christopher, S., Kamber, A., Pandolfi, E. and Rocci, A. [Ed.], Bulletin suisse de linguistique appliquée, 2015 (in press).

[21] Sacks, H., Schegloff, E. A. and Jefferson, G., "A simplest systematics for the organization of turn-taking for conversation", Language, 50(4):696-735, 1974.

[22] Goffman, E., Interaction ritual. Essays on face-to-face, London: Penguin University books, 1972.

[23] Bouchard, R., "L'interaction en classe comme polylogue praxéologique", in Grossman, F. [Ed.], Mélanges en hommage à Michel Dabène, Grenoble: ELLUG, 1998.

[24] Sinclair, J. M. and Coulthard, M., Towards an analysis of discourse: the English used by teachers and pupils. Oxford: Oxford University Press, 1975.

[25] Kerbrat-Orecchioni, C., "A multilevel approach in the study of talk-in-interaction", Pragmatics, 7:1-20, 1997.

[26] Goffman, E. The presentation of self in everyday life, Edinburgh: University of Edinburgh, 1959.

[27] Jones, R., "The problem of context in computer mediated communication", in LeVine, P. and Scollon, R. [Eds.], Discourse and technology: multimodal discourse analysis, 20-33, Washington DC.: Georgetown University Press, 2004.

[28] Kress, G. R., Multimodality: a social semiotic approach to contemporary communication, New-York: Routledge, 2010.

[29] Labov, W., "Some principles of linguistic methodology", Language in society, 1(1):97-120, 1972.

[30] Van Lier, L., "From input to affordance: Social-interactive learning from an ecological perspective", in Lantolf, J. P. [Ed.], Sociocultural theory and second language learning, 254-259, Oxford: Oxford University Press, 2000.

[31] Cambra Giné, M., Une approche ethnographique de la classe de langue. Paris : Didier, 2003.

[32] Sloetjes, H. and Wittenburg, P., "Annotation by category - ELAN and ISO DCR", in Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis S. and Tapias, D. [Eds.], Proceedings of the 6th International conference on language resources and evaluation, 816-820, Marrakech: European Language Resources Association (ELRA), 2008.

[33] Heylen, D., Bevacqua, E., Tellier, M. and Pelachaud, C., "Searching for prototypical facial feedback signals", in Aylett, R, Krenn, B., Pelachaud, C. and Shimodaira, H. [Eds.], Intelligent Virtual Agents. Lecture Notes in Computer Science, 147-153, Berlin : Springer Verlag, 2007.

[34] McNeill, D., Hands and mind: what gestures reveal about thought. Chicago: University of Chicago Press, 1992.

[35] Tellier, M., L'impact du geste pédagogique sur l'enseignement-apprentissage des langues étrangères : Etude sur des enfants de 5 ans. Thèse de doctorat non publiée. Université Paris 7-Denis Diderot, 2006.

[36] Azaoui, B., Coconstruction de normes scolaires et contextes d'enseignement. Une étude multimodale de l'agir professoral. Thèse de doctorat non publiée. Université Paul Valery, Montpellier III, 2014.

[37] Kerbrat-Orecchioni, C., Le discours en interaction. Paris: Armand Colin, 2005.

[38] Tellier, M., Azaoui, B. and Saubesty, J., "Segmentation et annotation du geste : méthodologie pour travailler en équipe", in Braffort, A., Boutora, L. and Sérasset, G. [Eds.], Actes de la conférence conjointe JEP-TALN-RECITAL 2012 : Atelier

Défi GEste Langue des Signes (DEGELS), Grenoble, France, 41-55, Grenoble : AFCP/ATALA, 2012.

[39] Azaoui, B., “Multimodalité des signes et enjeux énonciatifs en classe de FL1/FLS”, in Tellier, M. and Cadet, L. [Eds.], Le corps et la voix de l’enseignant: mise en contexte théorique et pratique, Paris : Editions Maison des Langues, 2014.

[40] Vion, R., La communication verbale. Analyse des interactions, Paris: Hachette, 1992.

[41] Posner, M. I., “Orienting of attention”, The Quarterly Journal of Experimental Psychology, 32(1):3-25, 1980.

[42] Posner, M. I., “The attention system of the human brain”, Annual review of neuroscience, 13:25-42, 1990.

[43] Lamargue-Hamel, D., “Des notions d’attention…”, Rééducation orthophonique, 42(218):7-23, 2004.

[44] Streeck, J. “Ecologies of gesture: Action and interaction”, in Streeck, J. [Ed.], New Adventures in Language and Interaction, 221-240, Amsterdam: John Benjamins, 2010.

[45] Kerbrat-Orecchioni, C., Les interactions verbales, Tome I, Paris: Armand Colin, 1990.

[46] Goodwin, C., Conversational organization. Interaction between speakers and hearers, London: Academic press, 1981.

[47] Goodwin, C., “Transparent vision”, in Ochs, E., Schegloff, E. A. and Thompson, S. A. [Eds.], Interaction and grammar, 370-404, Cambridge: Cambridge University Press, 1996.

[48] Cicurel, F., Les interactions dans l’enseignement des langues. Agir professoral et pratiques de classe, Paris: Didier, 2011.

Symbols used in transcriptions

T	Teacher
↑	upward intonation
underlining	overlapping
++	pause
XX	inaudible utterance
:::::	stretching of sound
//	interruption
[ ]	gesture production

Searching and retrieving multi-levels annotated data

Brigitte Bigi[1], *Jorane Saubesty*[1]

[1]Laboratoire Parole et Langage, CNRS, Aix-Marseille Université
5, avenue Pasteur, 13100 Aix-en-Provence, France
brigitte.bigi@lpl-aix.fr, jorane.saubesty@blri.fr

Abstract

The annotation of recordings is related to many Linguistics subfields as Phonetics, Prosody, Gestures or Discourse... Corpora are annotated with detailed information at various linguistic levels thanks to annotation software. As large multimodal corpora become prevalent, new annotation and analysis requirements are emerging. This paper addresses the problem of exploring annotations in order to extract multimodal data in the linguistic field ranging from general linguistic to domain specific information. The answer choose to fulfill this purpose is a user-oriented approach: the data can be extracted without any specific knowledge or skill. The paper exposes two ways to filter the annotations by a predicative approach: 1/ single filters, i.e. search in one tier depending of the data content, by the extraction of the time values and the duration; 2/ relation filters, i.e. search on annotations of a tier in time-relation with annotations of another one. This system is distributed in SPPAS software, under the terms of a public license.

Index Terms: software, multi-levels annotations, filtering

1. Introduction

When people communicate: gestures, postural shifts, facial expression, backchannel continuers such as “mm-hmm”, spoken turns and many more, all together work in concert to bring about mutual understanding. Annotating recordings of people communicating may therefore involve many Linguistics subfields such as Phonetics, Prosody, Gestures or Discourse... As a consequence, the last ten years or so have witnessed a real increase of linguistic annotated data. Whereas few years ago it was common to formulate linguistic models on the basis of rather limited data, today it is becoming more and more expected for linguists to take into account large quantities of empirical data, often including several hours of recordings. As a consequence, a number of software for the manual annotation of audio and/or video recordings have become available, such as *Anvil* [1], *Elan* [2], *Praat* [3], *Transcriber* [4] or *Exmaralda* [5], to name just some of the most popular tools, all of which are both free and multi-platform. Furthermore, linguists need tools for the automatic annotation, including the alignment of the speech recording with a phonetic transcription of the speech, as SPPAS [6].

As large multimodal corpora become prevalent, new analysis requirements emerge. Multimodal analysis has become a crucial part of research, teaching and practice for a wide range of academic and practical disciplines. The difficulties of multimodal analysis are visible in most of the works that explore this field. Multimodal annotation requires the possibility to encode many different information types, from different domains, with different levels of granularity [7].

”Corpora that include time-based data, such as video and marking gestures, make annotation and analysis of language and behavior much more complex than analysis based solely on text corpora and an audio signal” [8]. Thus, nowadays one of the biggest barriers with which the linguists must cope, is not the storage of data, nor its annotation, but rather *its exploration*. In addition to annotation, some tools provide statistical analysis capabilities. A minimum capability required is to search for annotated entities and their relationships [8]. Generally, different annotation tools are designed and used to annotate the audio and video contents of a corpus that can later be merged in query systems or databases [9]. With the help of multimodal corpora searches, the investigation of the temporal alignment (synchronized co-occurrence, overlap or consecutivity) of gesture and speech has become possible [9]. ”Obviously, the raison d’être of annotation in general is to allow linguists to retrieve all and only all instances of a particular phenomenon” [10].

The question of multi-levels filtering for linguistic annotated resources covers different aspects. It firstly requires a representation framework making it possible to compare, and eventually merge, different annotation schemes from different annotation tools. The basic structures of speech/video annotated data are “tiers” or “tracks” of annotations. Thus, speech/video annotation tools rely on this formalism because the *Tier* representation is appropriate to any multimodal annotated data given its genericity and flexibility and that it simply maps the annotations on the timeline. In the context of such tools, a *Tier* is a series of *Annotation* instances, each one defined by a temporal localization (an interval or a point) and a label. Obviously, due to the diversity of linguistic phenomena, annotation tools lead to a variety of models, theories and formalisms. This diversity results in heterogeneous description formats, each tool developing its own framework. Then, even if some are compatible, none of the annotation tools are directly interoperable, each one using a native format, some of them on top of XML, some others developing an *ad hoc* markup language. The heterogeneity of such annotations has been recognized as a key problem limiting the interoperability and re-usability of Natural Language Processing tools and linguistic data collections.

This paper focuses on the problem of *searching and retrieving data from multi-levels annotated corpora*. After a review of the main tools allowing to built queries in a multimodal annotated corpus, this paper presents the specifications of a software development according to eight criteria it must respect. The system proposed in this paper is a component named DataFilter in SPPAS software [6], described in Section 3. The method to search and retrieve data is based on a predicative approach allowing the definition of 2 types of filters: 1/ single filters, i.e. search in one tier depending of the data content, by the extraction of the time values or the duration (Section 4); 2/ relation filters, i.e. search on annotations of a tier in time-relation with annotations of another one (Section 5). Finally, Section 6 shows with a concrete study the benefit of the proposed software.

2. Background and motivations

A query is a request for a subset of all annotation elements, given some constraint. A query language (QL) is a programming language allowing to write queries. In the context of extracting multi-levels annotated data, multi-levels annotations can quickly become cluttered, so that the user needs query functionality to efficiently find relevant information. The following explores some popular and freely available tools.

Praat allows to paint intervals in green color, labels matching a given pattern with one of the following criteria: is equal to, is not equal to, contains, does not contain, starts with, does not start with, ends with, does not end with, matches a regular expression.

EXAKT (EXMARaLDA Analysis- and Concordance Tool) is the query and analysis tool for EXMARaLDA corpora, and can also be used for corpora created with other tools as Transcriber or Elan. Labels of annotations can be search in the corpus using regular expressions. It allows to save query results (HTML, text) or export them to other applications (e.g. Excel).

Elan proposes an advanced search form. It allows cascading constraints on the labels of the annotations and/or on relations between intervals. The relations are: is inside, overlaps, overlaps only begin time, overlaps only end time, is within...around, is within...around begin time of, is within...around end time of. The result is a list of filtered annotations the user can click on to visualize; it can also be saved as text file.

ANVIL internally maps the user's annotations to a temporary SQL database that is kept in sync at all times. Constraints can be formulated in SQL syntax. Labels of annotations can be queried using regular expressions. ANVIL also implements seven of the Allen relations [11] to compare intervals: equals, before, meets, overlaps, starts, finishes and during. In addition, the user can specify a tolerance limit in seconds. To spare the user from using long and complex SQL expressions, it implements a special syntax to ask for annotations from two tiers that are characterized by a certain temporal relationship.

The ANNIS2 system [12] proposes a query language (QL) including exact and regular expression matching on words forms and annotations, together with complex relations between individual elements, such as all forms of overlapping, contained or adjacent annotation spans, hierarchical dominance (children, ancestors, left- or rightmost child etc.) and more. Alternatively to the QL, data can be accessed using a graphical query builder. The result can be saved as text file or ARFF file.

To sum-up, the previously mentioned annotation tools offer the possibility to search or to retrieve annotations given some constraints. However, none of them fulfills the whole list of the following specifications a system should includes:

- allowing to import multi-levels annotated data from most of the existing annotation tools;
- providing the filtered result in the form of a new annotation tier;
- dealing with interval tiers as well as point tiers;
- allowing to export the filtered tier(s) in most of the existing annotation tools;
- allowing to filter multiple files at once;
- proposing both a scripting language and a Graphical User Interface (GUI);
- being powerful enough to meet the reasonable needs of end-users;
- can be used without requiring any XML-related or QL-related knowledge or skill;

3. DataFilter in SPPAS

The system proposed in this paper is implemented as a component named DataFilter in SPPAS [6], a software for ”Automatic Annotation of Speech” and distributed under the terms of the GNU Public License. It is implemented using the programming language Python. This software fulfills the specifications listed in [13]: it is a linguistic tool, free of charge, ready and easy to use, it runs on any platform and it is easy to install, the maintenance is guaranteed and it is XML-based.

Our proposal is to use the simplest level of representation , which is independent from the constraints of the coding procedure and the tools. Requests are based on the common set of information all tool are currently sharing. Basic operations are proposed and their combination allows the data to be requested, even by non-experts. Such a system fulfills the eight specifications mentioned in Section 2.

The framework implemented in this software to represent multi-levels annotated data is particularly suitable in the context of this paper to compare bounds of intervals or points between the various tiers: SPPAS solves the problem of the imprecision of annotations for each domain. Indeed, it allows to represent a bound as a tuple (M, R), where M is the midpoint value and R is a radius value, i.e. the vagueness of the point, as described in [14]. Consequently, each boundary of the annotations is represented as an uncertain time value: it makes it possible to account explicitly for the imprecision of input data. For example, the radius value can be fixed to 40-80ms in case of Gestures annotations and 5-10ms in case of Phonetics. This representation allows robust comparisons of multi-levels annotations over time. SPPAS also allows annotations to contain more than one label, each one associated with a score: the one with the highest score is considered as the main label, and the others as alternatives. Moreover, labels can be of 3 types: string, number or Boolean.

Actually, it is also quite easy to read some existing annotation file formats and to instantiate them into the SPPAS framework. Among others, it allows to open and save files from *Praat* [3], *Phonedit* [15], *Elan* [2]; *HTK* [16] and *Sclite* [17] and some subtitles formats. It also allows to import data from *Anvil* [1] and *Transcriber* [4].

The common denominator of most of the file formats consists in the basic building blocks (e.g. labels with start and end times, or labels and one time point) plus the additional structural entities (tiers). So, the system proposed in this paper is exploiting only these information: it allows to request all annotations regardless the input file format or the annotation type.

The exploration method is based on the creation of 2 different types of predicates. These latter are then respectively used in 2 types of filters:

1. single filters (Section 4), i.e. search in a tier depending on the data content, the time values or the duration of each annotation;

2. relation filters (Section 5), i.e. search on annotations of a tier in time-relation with annotations of another one.

4. Filtering annotations of a single tier

The main principle here is to create a predicate Sel, or a combination of predicates, that will be used as parameters to create a filter on a tier, named $SingleFilter(predicate, tier)$.

4.1. Filtering on the annotation labels

Pattern selection is an important part to extract data of a corpus and is obviously an important part of any filtering system, as shown in Section 2. Thus, if the label of an annotation is a string, the following predicates are proposed:

- exact match: $Sel(exact = P)$ is true if the label of an annotation strictly corresponds to the pattern P;
- contains: $Sel(contains = P)$ is true if the label of an annotation contains the expected pattern P;
- starts with, $Sel(startswith = P)$ is true if the label of an annotation starts with the expected pattern P;
- ends with, $Sel(endswith = P)$ is true if the label of an annotation ends with the expected pattern P.

All these matches can be reversed to represent respectively: not exactly match, not contains, not starts with or not ends with. Moreover, this pattern matching can be case sensitive or not. For complex search, a selection based on regular expressions is available for advanced users, as $Sel(regexp = R)$, where R is the expected regexp. Moreover, in case of numerical labels, we implemented: $Sel(equal = v)$, $Sel(greater = v)$ and $Sel(lower = v)$, and in case of Boolean: $Sel(bool = v)$. Finally, this pattern matching can be optionally applied either on the label with the highest score, which is the default, or on all labels of an annotation (i.e. the better label and its alternatives).

4.2. Filtering on annotations durations or on a time-range

Another important feature for a filtering system is the possibility to retrieve annotated data of a certain duration, and in a certain range of time in the timeline. Therefore, the same predicate Sel can be used on match duration of an interval, compared to a value v, as follow:

- lower: $Sel(duration\_lt = v)$;
- lower or equal: $Sel(duration\_le = v)$;
- greater: $Sel(duration\_gt = v)$;
- greater or equal: $Sel(duration\_ge = v)$;
- equal: $Sel(duration\_e = v)$;

Search can also starts and ends at specific time values in a tier by using Sel predicate with $begin\_ge$ and end_le.

4.3. Multiple selections

A multiple pattern selection as well as duration or time selections can be expressed with the operator ”|” to represent the logical ”or” and the operator ”&” to represent the logical ”and”, for example:
$Sel(startswith = P_1)$ & $Sel(duration\_gt = v)$.

5. Filtering on relations between two tiers

Regarding the searching , linguists are typically interested in locating patterns on specific tiers, with the possibility to relate different annotations a tier to another. The proposed system offers a powerful way to request/extract data, with the help of Allen’s interval algebra. The main principle here is to create a predicate Rel that will be used as parameter to create a filter on a tier: $RelationFilter(predicate, tier1, tier2)$.

5.1. Framework: Allen’s interval algebra

In 1983 James F. Allen published a paper [11] in which he proposed 13 basic relations between time intervals that are distinct, exhaustive, and qualitative. They are distinct because no pair of definite intervals can be related by more than one of the relationships; exhaustive because any pair of definite intervals are described by one of the relations; qualitative (rather than quantitative) because no numeric time spans are considered. These relations and the operations on them form Allen’s interval algebra.

SPPAS extended Allen’s work to its framework that can handle relationships between intervals with *precise as well as imprecise bounds*. This results in a generalization of Allen’s 13 interval relations that are also applicable when the bounds of the intervals are imprecise. Table 1 indicates the Allen’s relations between TimeInterval $X = [X^-, X^+]$ and $Y = [Y^-, Y^+]$, where X^-, X^+, Y^-, Y^+ are TimePoint instances, as defined in [14]. This generalization preserves the 3 properties of Allen’s original algebra mentioned above.

X relation Y	Description
before	$(X^+ < Y^-)$
after	$(X^- > Y^+)$
meets	$(X^+ = Y^-)$
met by	$(X^- = Y^+)$
overlaps	$(X^- < Y^-) \wedge (X^+ > Y^-) \wedge (X^+ < Y^+)$
overlapped by	$(X^- > Y^-) \wedge (X^- < Y^+) \wedge (X^+ > Y^+)$
starts	$(X^- = Y^-) \wedge (X^+ < Y^+)$
started by	$(X^- = Y^-) \wedge (X^+ > Y^+)$
during	$(X^- > Y^-) \wedge (X^+ < Y^+)$
contains	$(X^- < Y^-) \wedge (X^+ > Y^+)$
finishes	$(X^- > Y^-) \wedge (X^+ = Y^+)$
finished by	$(X^- < Y^-) \wedge (X^+ = Y^+)$
equals	$(X^- = Y^-) \wedge (X^+ = Y^+)$

Table 1: Allen’s relations between two imprecise intervals X, Y

The proposed framework was also developed to include time annotations represented by a single TimePoint (mainly used in the Prosody domain). The relations can be extended to such time representation, as we propose in Table 2 between two TimePoint instances. Tables 3 and 4 show relations between a TimePoint and a TimeInterval. Each table considers all possible relations (each table forms a complete relation system).

X relation Y	Description
before	$(X < Y)$
after	$(X > Y)$
equal	$(X = Y)$

Table 2: Relations between two imprecise points X and Y.

These relations can then be used to search annotations of any kind in time-aligned tiers. It is particularly favorable in the context of multimodal annotations, where annotations are carried out thanks to various annotation tools, each one using its own representation of time. The proposed framework solves this problem in a clear, well-suited and well-defined way.

X relation Y	Description
before	$(X^+ < Y)$
after	$(X^- < Y)$
starts	$(X^- = Y)$
finishes	$(X^+ = Y)$
contains	$(X^- < Y) \wedge (X^+ > Y)$

Table 3: Relations between an imprecise interval X and an imprecise point Y.

X relation Y	Description
before	$(X < Y^-)$
after	$(X > Y^-)$
starts	$(X = Y^-)$
finishes	$(X = Y^-)$
during	$(X > Y^-) \wedge (X < Y^+)$

Table 4: Relations between an imprecise point X and an imprecise interval Y.

5.2. Filtering with time-relations

For the sake of simplicity, only the 13 relations of the Allen's algebra are available in the GUI. We withal implemented in Python the 25 relations proposed by [18] in the INDU model. This model fixes constraints on INtervals with Allen's relations and on DUration - duration are equals, one is less/greater than the other.

Moreover, both our experience while using the proposed system and the user comments and feedback have led us to add the following options:

1. a maximum delay for the relations "before" and "after",
2. a minimum delay for the relations "overlaps" and "overlapped by".

All the above mentioned relations were implemented as predicates. With this proposal, a predicate can be for example $predicate = Rel("overlaps")|Rel("overlappedby")$ to find witch syllables stretch across two words, and then by creating the filter $RelationFilter(predicate, tiersyllables, tiertokens)$.

6. Illustrations

DataFilter of SPPAS has been already used in several studies as to find correlations between speech and gestures [19], to find which gestures are produced while pausing [20] or to extract lexical feedback items [21] just to cite some of them.

While using the GUI, the user starts filtering tiers by running DataFilter and loading files of a corpus. The user selects the tier of each file that will serve as basis, and click on the appropriate "Filter" button (either Single or Relation). The user has then to define the predicates and to apply such filters. The program will generate new tiers with the matching annotations; each one is automatically added to its corresponding file.

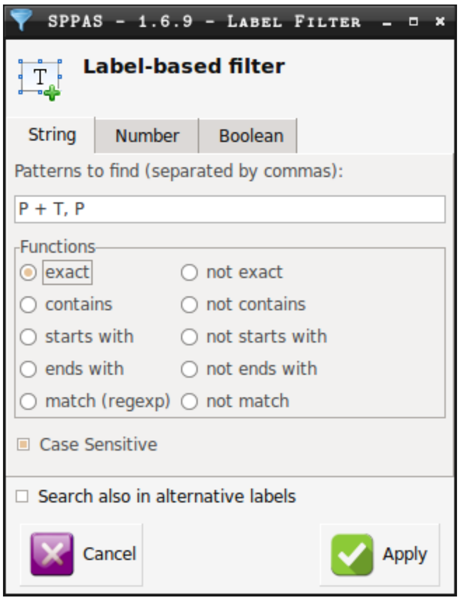
In order to illustrate possible queries using SPPAS, the following request is processed in this section: *What speech and hand gestures the locutor produces right before, during and right after the interlocutor produces multimodal feedbacks versus verbal feedbacks only?*

We performed this request on 6 files of a corpus created by and belonging to the Institut Paoli-Calmettes (Marseille, France). This corpus is an authentic corpus of training sessions for doctors involved in role plays with an actor playing the role of a patient. The corpus is annotated on different levels of granularity. Tiers contain annotations of vocabulary, hand gestures, gaze, among other. In the context of this article, we will consider only 3 tiers:

1. P - Feedback: feedback produced by the patient
2. M - IPUs: speech produced by the doctor and segmented into Inter Pausal-Units
3. M - Dimensions: hand gestures produced by the doctor

To perform the illustration request, the first stage consists in filtering the "P - Feedback" tier of each file to create an intermediate result with a tier containing *head and oral feedback* ("P + T") and *oral feedback* only ("P").

While using the GUI, this predicates are fixed as represented in Figure 1. It allows to enter multiple patterns at the same time (separated by commas) to mention the system to retrieve either one pattern or the other, etc.



Patterns to find (separated by commas):

P + T, P

Functions

● exact ○ not exact
○ contains ○ not contains
○ starts with ○ not starts with
○ ends with ○ not ends with
○ match (regexp) ○ not match

☒ Case Sensitive

☐ Search also in alternative labels

Figure 1: Frame to create a filter on annotation labels. In that case, labels that are exactly matching "P + T" or "P" strings.

So, here the patterns are "P + T, P". Finally, the user has to select the tier name for the result as shown in Figure 2 and must click either to "Apply all" or "Apply any". The user has now one filtered tier by file, each one containing only *oral feedbacks* and *oral and head movements feedbacks*.

To complete the original request, the previous tiers must be unchecked. The user must now find annotations of speech and hand gestures that occur right before, during and right after the feedbacks previously filtered. To do so, the newly filtered tiers must in turn be checked and the user must click on the "RelationFilter" button. Then, he/she selects "M - IPUs" in the "X" windows, and the filtered tier previously created in the "Y" window in the list of proposed tiers, as he/she wants to filter speech. Finally, the Allen's relations must be selected: see a glimpse in Figure 3. Regarding the example, quite every relations are needed. Though, the relations "Before" and "After" must be customized. The user needs to extract IPUs before and after the feedbacks. Customizing the delay allows the user to chose the exact delay between the feedback utterance and the nearer IPUs the user wants to take into consideration. To complete the filtering process, it must be clicked on the ”Apply filter” button and the new resulting tiers are added in the annotation file(s).

In order to answer the question firstly asked, the user must complete the filter loop once again. He/she must click again on the ”Relation Filter” button and select ”M - Dimensions” in the ”X” windows, and the previously filtered tier in the ”Y” window, as the user wants, this time, to filter hand gestures in the list of proposed tiers. Then, the relations must be selected afresh. As the user does not want hand gestures produced out of the IPUs window, the user must check: Starts, Started by, Finishes, Finished by, Contains, During, Overlaps, and Overlapped by. Then, it must be clicked one last time on the ”Apply filter” button and the new resulting tiers are added in the annotation file(s). The last resulting tier therefore contains the annotations of hand gestures produced by the locutor while speaking, right before, during and right after the interlocutor produced *oral* or *oral and head movements* feedback.

The user can keep or delete intermediate tiers and click on the ”Save” button. The files are saved in their original file format and can therefore be opened in the annotation tool used to create the files in the first place. They can also be opened by ”Statistics” component proposed in SPPAS.

7. Conclusions

This paper described a system to filter multi-levels annotations. It is based on the definition of predicates that are used to create filters. These later are applied on tiers of many kind of input files (TextGrid, eaf, trs, csv...). The search process results in a new tier, that can re-filtered and so on. A large list of predicates, applied on a single tier, is available and allows to filter annotations of various label types (string, number, Boolean). The full list of binary relations of the INDU model are also available to filter the annotations of a tier in relation with the annotations of another one. Moreover, this request system can be used either on precise or unprecise annotations. It is available as a Python Application Programming Interface, and with a GUI. Since the proposed system has been developed in a constant exchange with users from various Linguistics fields, we expect it to be especially intuitive.

8. Acknowledgements

This work has been partly carried out thanks to the support of the French state program ORTOLANG (Ref. Nr. ANR-11-EQPX-0032) funded by the “ Investissements d’Avenir ” French Government program, managed by the French National Research Agency (ANR). This work was supported by Brain and Language Research Institute.

9. References

[1] M. Kipp, “Anvil, DFKI, German Research Center for Artificial Intelligence,” http://www.anvil-software.de/, 2011.

[2] Nijmegen: Max Planck Institute for Psycholinguistics, “ELAN - linguistic annotator. language archiving technology portal [computer software],” http://www.lat-mpi.eu/tools/elan/, 2011.

[3] P. Boersma and D. Weenink, “Praat: Doing phonetics by computer. [Computer Software] Amsterdam: Department of Language and Literature, University of Amsterdam. ,” http://www.praat.org/, 2011.

[4] Transcriber, “A tool for segmenting, labeling and transcribing speech. [computer software] paris: DGA,” http://transag.sourceforge.net/, 2011.

[5] T. Schmidt and K. Wörner, “Exmaralda,” in *Handbook on Corpus Phonology*, U. G. Jacques Durand and G. Kristoffersen, Eds. Oxford University Press, 2014, pp. 402–419.

[6] B. Bigi, “SPPAS: a tool for the phonetic segmentations of Speech,” in *The eighth international conference on Language Resources and Evaluation, ISBN 978-2-9517408-7-7*, Istanbul, Turkey, 2012, pp. 1748–1755.

[7] C. Jewitt, “Handbook of multimodal analysis,” *London: Roufledge*, 2009.

[8] T. Bigbee, D. Loehr, and L. Harper, “Emerging requirements for multi-modal annotation and analysis tools.” in *INTERSPEECH*, 2001, pp. 1533–1536.

[9] Á. Abuczki and E. B. Ghazaleh, “An overview of multimodal corpora, annotation tools and schemes,” *Argumentum*, vol. 9, pp. 86–98, 2013.

[10] S. T. Gries and A. L. Berez, “Linguistic annotation in/for corpus linguistics,” *Handbook of Linguistic Annotation. Berlin, New York: Springer. Abgerufen von*, 2015.

[11] J. Allen, “Maintaining knowledge about temporal intervals,” *Communications of the ACM*, pp. 832–843, 1983.

[12] A. Zeldes, A. Lüdeling, J. Ritz, and C. Chiarcos, “ANNIS: a search tool for multi-layer annotated corpora,” 2009.

[13] S. Dipper, M. Götze, and M. Stede, “Simple annotation tools for complex annotation tasks: an evaluation,” in *Proceedings of the LREC Workshop on XML-based richly annotated corpora*, 2004, pp. 54–62.

[14] B. Bigi, T. Watanabe, and L. Prévot, “Representing multimodal linguistics annotated data,” in *9th International conference on Language Resources and Evaluation, ISBN: 978-2-9517408-8-4*, Reykjavik (Iceland), 2014, pp. 3386–3392.

[15] Phonedit, http://www.lpl-aix.fr/~lpldev/phonedit/, 2014.

[16] S. Young and S. Young, “The HTK Hidden Markov Model Toolkit: Design and Philosophy,” *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

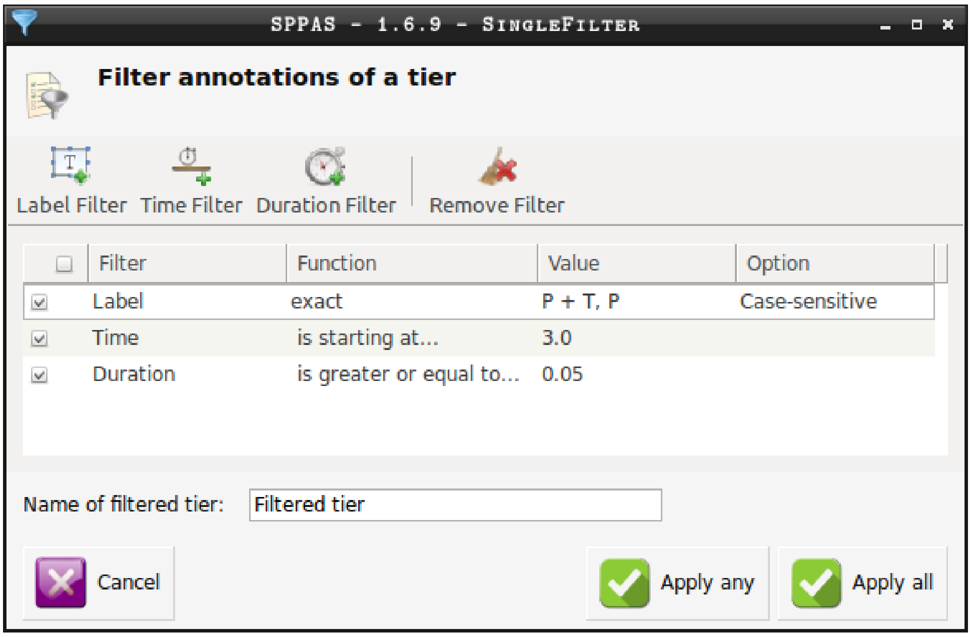
[17] “Speech Recognition Scoring Toolkit, version 2.4.0,” http://www.itl.nist.gov/iad/mig/tools/, 2009.

[18] A. K. Pujari, G. V. Kumari, and A. Sattar, “Indu: An interval & duration network,” in *Advanced Topics in Artificial Intelligence*. Springer, 1999, pp. 291–303.

[19] M. Tellier, G. Stam, and B. Bigi, “Same speech, different gestures?” in *5th International Society for Gesture Studies*, Lund (Sweden), 2012.

[20] ——, “Gesturing while pausing in conversation: Self-oriented or partner-oriented?” in *The combined meeting of the 10th International Gesture Workshop and the 3rd Gesture and Speech in Interaction conference*, Tillburg (The Netherlands), 2013.

[21] L. Prévot, B. Bigi, and R. Bertrand, “A quantitative view of feedback lexical markers in conversational French,” in *Proceedings of 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Metz, France, 2013.



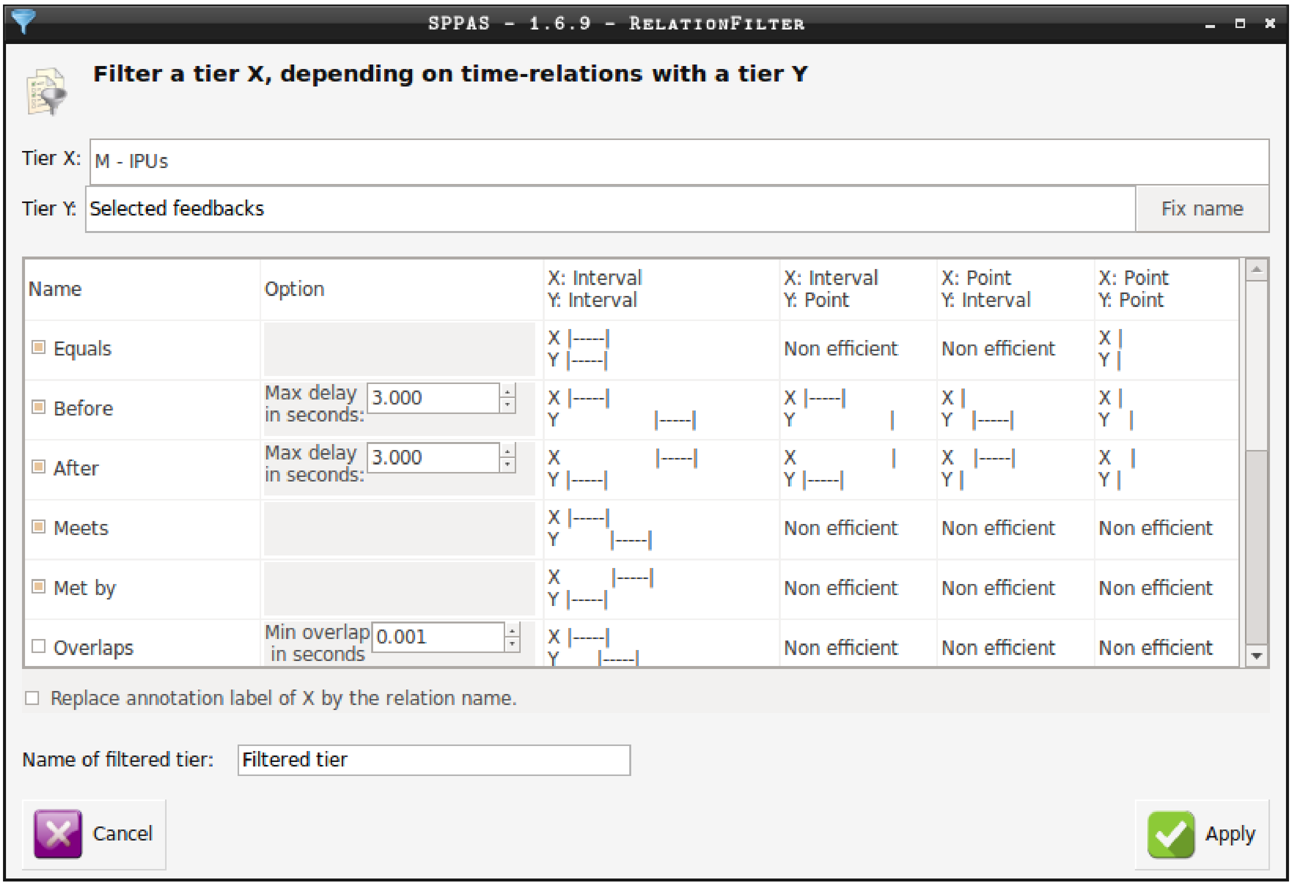
SPPAS - 1.6.9 - SingleFilter

Filter annotations of a tier

☐	Filter	Function	Value	Option
☑	Label	exact	P + T, P	Case-sensitive
☑	Time	is starting at...	3.0	
☑	Duration	is greater or equal to...	0.05	

Name of filtered tier: Filtered tier

Figure 2: Example of the Single filter frame. For the purpose of an exhaustive illustration, 3 predicates are described here 1/ to select patterns, 2/ to create a filter on annotation duration and 3/ on a time-range.



SPPAS - 1.6.9 - RelationFilter

Filter a tier X, depending on time-relations with a tier Y

Tier X: M - IPUs

Tier Y: Selected feedbacks

Name	Option	X: Interval Y: Interval	X: Interval Y: Point	X: Point Y: Interval	X: Point Y: Point
▣ Equals		X |-----| Y |-----|	Non efficient	Non efficient	X | Y |
▣ Before	Max delay in seconds: 3.000	X |-----| Y |-----|	X |-----| Y |	X | Y |-----|	X | Y |
▣ After	Max delay in seconds: 3.000	X |-----| Y |-----|	X | Y |-----|	X |-----| Y |	X | Y |
▣ Meets		X |-----| Y |-----|	Non efficient	Non efficient	Non efficient
▣ Met by		X |-----| Y |-----|	Non efficient	Non efficient	Non efficient
☐ Overlaps	Min overlap in seconds 0.001	X |-----| Y |-----|	Non efficient	Non efficient	Non efficient

☐ Replace annotation label of X by the relation name.

Name of filtered tier: Filtered tier

Figure 3: Frame to create a filter on time-relations

Use of gesture space in gestural accommodation

Charlotte Bouget & Marion Tellier

Aix Marseille Université, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France

charlotte.bouget@gmail.com, marion.tellier@lpl-aix.fr

Abstract

Speakers adapt their speech to their interlocutors and when they talk to an elderly person, they tend to engage in elderspeak.

In this paper, we explore with a new approach how caregivers adapt their use of gesture space in a vocabulary explanation task with older and younger adults. Preliminary results on one caregiver show that she tends to occupy a larger gesture space when speaking to a senior than when addressing a younger partner. Thus, caregivers could spontaneously accommodate their discourse and their gestures to help interlocutors when they have difficulties or when caregivers think seniors have difficulties.

Index Terms: use of gesture space, elderspeak, caregiver, senior, gestural accommodation

1. Introduction

The proportion of the world's population over 60 is projected to double to reach 2 billion in 2050. The number of people aged 80 or older will have quadrupled over the same period[1]. Thus, the amount of research on normal and pathological aging[2] is increasing. Nowadays, studies focus on caregivers[3], their training and their health. That is why we focus our research on communication between caregivers and older adults, assuming that if caregivers adapt their speech verbally when addressing a senior, they should also adapt their gestures, especially in terms of use of gesture space.

1.1. Accommodation theory

Social interactions occur in everyday situations and speakers adapt their discourse depending on their interlocutor. For example, if an adult speaks to a baby, s/he will use a specialized speech called *baby talk* [1]. If s/he speaks to a foreigner, s/he will use a specialized speech called *foreigner talk* [2].

Concerning accommodation with elderly persons, there are different models that attempt to understand how adults adapt their discourse during inter-individual interactions [3]. These models are closely related to social identity theory [3] because speakers adapt their perceptions, their attitudes and their behavior. Thus, communication strategies of speakers change depending on the social representations they have of their addressees.

The Communication Predicament of Aging Model (CPA) is a reference in research on intergenerational communication [4] with a downward spiral. The CPA is a cyclical and patronizing speech that is often produced in response to age stereotypes ; younger adults produce an inadequate accommodation, and older adults cannot answer correctly. Thus, the inadequate response perpetuates negative stereotypes of the elderly. However, two important critics can be addressed to this model. First, older adults can answer appropriately even if adults produce an inadequate accommodation when they speak to them. Secondly, adults have both positive and negative stereotypes of seniors Thus, positive cycle can be achieved and patronizing speech can be reduced.

The Communication Enhancement of Ageing Model (CEA) [5] was developed to provide with a solution to the CPA- model limitations. This model focuses on positive stereotypes of aging. It proposes that when adults assess seniors individually, appropriate communication strategies can be selected and positive stereotypes of aging can be developed and reproduced.

Last, the Age Stereotypes in Interactions Model (ASI) [6] is an extension of the CPA model. Adults can develop stereotypes according to personal characteristics (age, cognitive complexity and quality of contact), interlocutors' characteristics (physiognomic cues to age, personal appearance, physique) and the context of the interaction. These stereotypes (positive or negative) influence beliefs about communication with older adults and negative stereotypes lead to use a specialized speech style.

1.2. Elderspeak

Elderspeak[4] is a particular speech style used by younger adults when addressing older adults [7]. It is characterized by simplified grammar and vocabulary, slower speech, higher pitch, exaggerated intonation, increased loudness, use of repetitions, endearing terms and tag questions [8].

In addition to verbal features, non-verbal characteristics are present and "include gaze, such as a lack of eye contact, eye rolls, or winking; proxemics, such as standing too close to a person or standing over a person who is sitting or lying in bed; facial expressions and gestures, such as frowns, exaggerated smiles, head shakes, shoulder shrugs, hands on hips or crossed arms; and touch, such as patting the older person's head, hand, arm, or shoulder." (p.5, [9]). Even if non-verbal accommodation

[1] World Health Organization, "http://www.who.int/ageing/about/facts/en/"

[2] Older adult : conceptualizations of old age is defined by three specifications : chronological age (measure in years from the date of birth), functional age (psychological state) and social age (reflects the image of people). In this paper we will use "adult" to refer to individuals aged between 18 and 59.

[3] Caregivers can be professional or familial, they help another person in need.

[4] In this study, we employ these terms as synonymous : elderspeak, patronizing speech, secondary use of baby talk and overaccommodation

when interacting with seniors has been studied, the adaptation of coverbal gestures to older partners remains unexplored.

1.3. Gestural accommodation

As far as gestures are concerned, we view the relationship between gesture and speech through a McNeillian perspective [10]. Thus, when adults accommodate their speech to their interlocutor, they should also adapt their gestures. In terms of gestural accommodation, speakers adapt their gesture according to whether they share common knowledge with their addressees or not [11] ; whether speakers see their interlocutors or when they are on the telephone [12, 13] ; when addressing to a human versus a machine [14] ; when they talk to a native partner or a non-native partner [15] or according to their partner's location in space [16].

More specifically, Tellier and Stam [15] have studied gestures produced by future teachers when explaining words to native and non-native partners[5]. They found that future teachers adapted their discourse and their gestures depending on their interlocutor. They analyzed gesture rate, gesture dimensions, duration and the use of gesture space. Results showed that future teachers tended to use gestures that were more illustrative, larger and lasted longer when they explained words to non-native listeners than when they addressed native interlocutors. This adaptation takes place to facilitate the decoding of speech by non-native interlocutors because they may encounter difficulties in oral comprehension. The future teachers adapt their gestures by projecting needs and potential difficulties onto their partner (they can be based on stereotypes of experience of communication with a non-native).

Since seniors interlocutors may also encounter comprehension problems (due to weak audition or slower reaction time for feedback, for instance), our research goal is to find out whether the same gestural adaptation occurs when caregivers address older partners (as opposed to young our middle-aged adults).

In this paper, we focus specifically on the use of gesture space, which has been hardly addressed in terms of gestural accommodation.

McNeill [10] developed a gesture space diagram (Figure 1) based on data collected during a narrative task to analyze where the stroke[6] of gestures was produced. He found that the use of gesture space was different depending on the dimensions of gestures. Iconics are congregating in the *Center-Center* space, metaphorics are produced in the *Center* space, deictics extend into the *Periphery* and beats do not have a specific space.

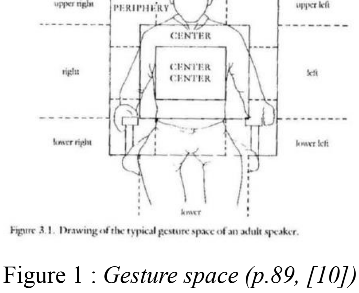

Classification, Transcription, and Distribution
EXTREME PERIPHERY
upper
upper right
PERIPHERY
upper left
CENTER
right
CENTER CENTER
left
lower right
lower left
lower
Figure 3.1. Drawing of the typical gesture space of an adult speaker.

Figure 1 : *Gesture space (p.89, [10])*

Coding the position of gestures in space also indicates the size of gestures. Gesture size varies across cultures [17-19]. In spontaneous situations in everyday environments, Efron found that Italian immigrants used larger gestures than Jewish immigrants [20], and Cavicchio and Kita compared gestures of English and Italian speakers, finding that Italian speakers produced larger gestures than English speakers and that bilinguals' gestures were larger than those of monolinguals [17]. In a natural conversation task, Müller showed that the use of gesture space is spatially more expansive for native Spanish speakers than for German speakers [19].

Asymmetrical interactions can also lead to a change in gesture space. Indeed, in the study by Tellier and Stam [15] mentioned above, future French teachers produced gestures more in the *Center-Center* and the *Center* areas when addressing a native interlocutor and they extended to the *Periphery* and *Extreme Periphery* when addressing non-natives [15].

McNeill's diagram is interesting for coding gesture space but only works for speakers sitting and facing a video camera (which is the display used in McNeill's narration tasks). Therefore, several researchers have suggested variations of the diagram to account for the tridimensional use of space [15, 19] and for standing and moving speakers like teachers in a classroom [20]. Speakers can produce gestures in front of them: "They reach into the space in front of them, move their hands further away from their body, bring them closer to their body or even touch their opponent." (p.20, [21]). So, Tellier and Stam [15] added one category: *arm stretched in front of the speaker* that has been also used by and Tellier, Guardiola & Bigi [22]. Bressem [21] used a more detailed schema with four categories: *speaker's own body, close distance, middle distance and far distance to the body*. He added one interesting particularity; when arms are produced in the back of the body he employed the sign "-".

Moreover, McNeill [10] analyzed where the gesture stroke occurred, but recent studies tend to use the most extreme location of a gesture during any part of the phase [15, 22-24].

In spite of these changes, the main failure of this coding process is that it focuses on the location of the gesture at one specific moment (the stroke or the most extended part) and thus does not take into account the occupation of space. Therefore, a small gesture produced at knee level and one produced above the head will both be coded the same (in extreme periphery) whereas they have different sizes, degrees of visibility and occupation of gesture space.

For this research, we created a new annotation scheme to code the use of gesture space. We tested this method with the corpus described below to find out whether caregivers adapt their gestures when addressing a younger adult *vs* an older adult. If gesture and speech are one system as stated by McNeill [10], then when caregivers adapt their discourse they should also adapt their use of gesture. Since they tend to articulate more and speak louder (as elderspeak is defined), they should also make their gestures more visible to help their older partners by using a larger space.

2. Methodology

2.1. Corpus

To collect data we used a semi-controlled methodological approach which consisted of collecting oral data from different participants who received the same instructions and were tested

[5] This accommodation is named Foreigner Talk.

[6] Stroke : the meaning unit of gesture

under the same conditions. Compared to controlled data, participants had a certain freedom in production. This approach has many benefits, including the ability to compare productions in different conditions and analyze data quantitatively and qualitatively [25].

2.1.1. Participants

There are three types of participants in this study. Seven caregivers (the main participants), seven seniors (healthy and independent) and seven younger adults whose characteristics in terms of age are presented in Table 1.

Table 1 : *Mean age and mean level of education (in years) (standard deviation in parenthesis)*

	Caregivers	Adults	Seniors
Age	50 (8,44)	46,71 (8,92)	85,14 (12,86)
Level of education	12,29 (3,45)	12,86 (1,77)	5,29 (0,76)

2.1.2. Design

In this study, we replicated the protocol used by Tellier and Stam [15] in the *Gesture and Teacher Talk* study (GTT). It consists of a lexical explanation task of 12 French words (3 names, 3 verbs, 3 adjectives and 3 adverbs with two concrete and one abstract concept for each category). Caregivers had to randomly draw the words from a box and successively explain the words to senior and younger partner. The interlocutors had to guess the words.

The order of explanations was counter-balanced ; four caregivers first explained the words to a younger adult and three caregivers first to a senior. Each participant signed an informed consent. Once the participants were installed, we gave them the instructions for the task. There was only one constraint : caregivers were not to use any words from the same word family. They could use any verbal and non-verbal means to explain.

2.2. Coding scheme

Data was coded and analyzed using software called ELAN [26]. We transcribed the speech of all interlocutors, segmented gestures, annotated gesture space, gesture phases, handedness and gesture depth.

2.2.1. Annotation of gestures

Then, we segmented each coverbal gesture produced by caregivers. Gestures begin when they leave their rest position and end when they return to the rest position or when the next gesture is initiated.

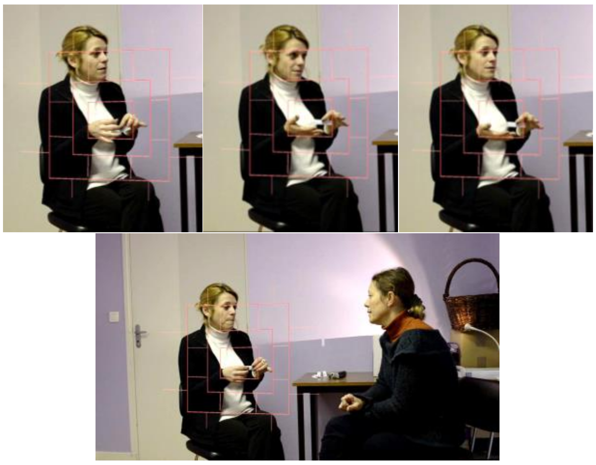
2.2.2. Annotation of gesture space

The goal of this research is to analyze the use of gesture space depending on the interlocutor. To reach that goal, we do not look at where the gesture is produced but how the gesture space is occupied by gesture production. With this method, we do not focus on where gestures are located but how large they are. Thus, the more zones crossed (such as center-center, center, periphery and extreme periphery), the larger occupation of gesture space.

We edited a systematic approach. We added McNeill's diagram [10] on a video. For that, we used photographic editing software (Photofiltre 7) to draw the diagram and we used video editing software (Wondershare Video Editor 3.5.1) to add this drawing to the video. The scheme must be thoroughly placed, and for that we used the same criteria as McNeill [10]. The first square is placed in the center of the speaker. The second is placed at the level of the shoulders and the third is placed at eye level. Some studies simplify this gesture space with only two zones [17] or with the four main zones [15, 22]. However for our research it is important to keep all the zones of this scheme.

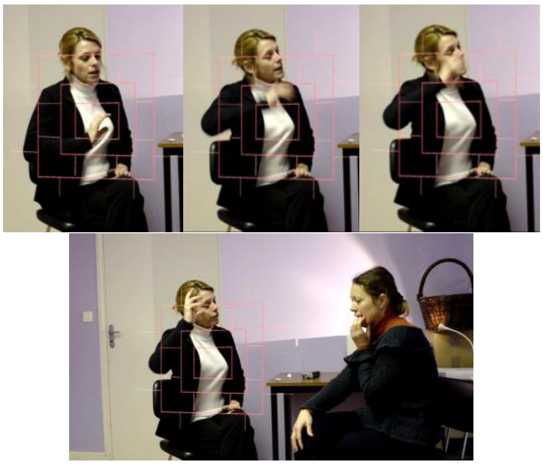
Once the diagram is set on the video, we can import it into ELAN for annotation. The coding of gesture space is simplified thanks to this scheme placed on the video. To code a gesture's occupation of gesture space, we count how many zones are crossed by each gesture, from its beginning to its end.

If a gesture is produced with one hand or two hands in the same zone, we code one zone (Example 1).



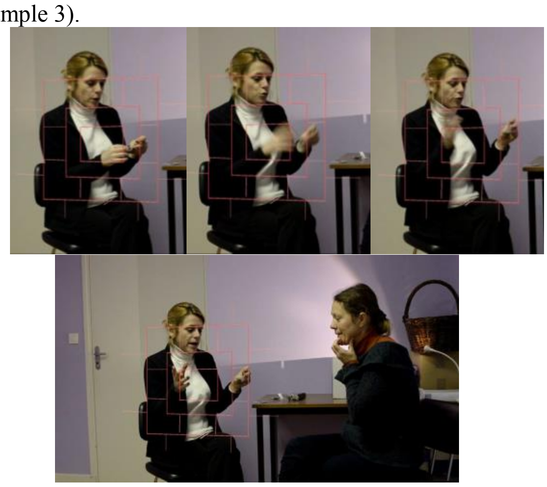
Example 1 *: The left hand holds a paper but does not move. The right hand moves, so we count how many zones are crossed. During the production of this gesture, the hands remain in the same zone.*

If a gesture is produced with one hand that moves through several zones, we count the number of zones including the first one (from where the gesture starts) (Example 2).



Example 2 *: The left hand does not move. The right hand leaves from the middle zone to go up, crossing 4 zones. In this example, the gesture ends when the speaker has her hand in the air, which is a typical phenomenon when there are many consecutive gestures. If the gesture returns to the rest position, we count the number of zones crossed during retraction.*

If a gesture is produced with two hands in different zones, we add up the number of zones crossed by each hand including the starting zone (if both hands are in the same zone at the beginning of the gesture, this zone is only counted once) (Example 3).



Example 3 *: The left hand moves slightly to the left and the right hand draws an arch. If both hands form the same gesture, the first zone is counted only once (3 gesture zones in this example).*

On top of the number of zones, we added three characteristics to the coding of the occupation of gesture space. The first is to indicate if gestures are produced in front of or behind the speaker. We use the sign "+" when the arms are moving in front and "-" when the arms are moving behind the body. This is an important point and must be described as precisely as possible, even though lateral gestures (on a right to left axis) outnumber sagittal gestures (on a front-back axis) [27].

The second characteristic focuses on handedness. We coded whether the gesture was produced with *one hand (OH)* or *both hands (BH)* [28, 29] because as we take into account all the crossed zones, gestures with both hands are larger than gestures with one hand.

The third characteristic deals with gesture phases. We did not segment gesture in phases but we took into account when gestures included a *preparation (P)*, a *retraction (R)* or a *preparation + retraction (P+R)* because gestures with these phases are larger than gestures consisting of only a stroke.

2.3. Research questions and hypothesis

In this study, we hypothesize that gestures produced with older interlocutors will be larger than those addressed to younger adults. Caregivers may adapt their use of gesture space to older partners in order to help them understand speech by making their gestures more visible. We also hypothesize that caregivers will produce more gestures with two hands when addressing older partners to occupy a larger portion of gesture space and to make their gestures more visible. We also predict that there will be more gestures produced in front of the caregiver when addressing to the older adult to reduce space between them [9].

3. Results

3.1. Inter-annotator agreement

Inter-annotator agreement enables one to find out whether several annotators make the same decision in terms of coding and thus reduces subjectivity. To assess reliability, three coders (two experts and one naïve) annotated gesture size. Each annotator had to count how many gesture zones were crossed for each gesture in the same sample of the corpus with the same guidelines. The Fleiss coefficient based on the three annotators was 0.74. This *k*-score is a "substantial agreement" according to Landis & Koch [30]. We are satisfied with this result because the value of annotations is free and this affects the magnitude of the Kappa value. We compared the mean of the three annotators with an ANOVA and found no significant difference between them [F(2,102)=0.342 ; p>0.05]. Thus, the difference between annotators is present but small.

3.2. Results

The results must be considered as preliminary since they only concern one pair of speakers. Results should be confirmed with the analysis of the other dyads of the corpus.

To begin with, it is important to note that there is a difference in terms of duration of explanation across the conditions. The task is easier with a younger adult interlocutor, as in this condition the task was shorter than in the older interlocutor condition.

3.2.1. Verbal characteristics of elderspeak

Concerning verbal accommodation, speech rate is similar with both interlocutors (114 words per minute with the senior partner *vs.* 128 words per minute with the younger adult partner). Thus, the caregiver did not significantly change their habitual speaking speed depending on the interlocutor even if she spoke a bit more slowly with the older interlocutor than with the younger adult interlocutor.

In a descriptive analysis, we focused on different verbal strategies for explaining the same word in both conditions [15]. For instance, the caregiver's speech was more illustrative when addressing to older partner, adding a contextualized example whereas she employed metalanguage with the younger adult partner (Example 4 & 5[7]).

(4) With the younger adult : « alors c'est un verbe / du premier groupe / euh / quand tu as tu as une échelle / et tu veux monter sur un arbre / un synonyme de monter »

(5) With the older adult : « quand tu dois ramasser tes cerises en haut de ton arbre / quand t- tu dis tu dis quoi »

Moreover, the caregiver adjusts her register of language, employing casual language with the younger adult and formal language with the older adult (Example 6 & 7[8]).

(6) With the younger adult : « alors / le contraire de / doucement / euh un bolide il est très »

(7) With the older adult : « c'est comment / c'est euh une voiture / elle peut aller euh très très vite »

[7] (4) so it's a verb / of the first group / uh / when you have you have a ladder / and you want to climb a tree / a synonym of climbing
(5) when you have to gather some cherries at the top of your tree / when y- you say you say what

[8] (6) so / the opposite of / slowly / uh a racing car it is very
(7) how is it / it is uh a car / it can go uh really really fast

3.2.2. Gestural characteristics of elderspeak

Concerning gestural accommodation, the caregiver produced fewer gestures with the younger adult partner (70 gestures) than with the older partner (100 gestures). However gesture rate (the number of gestures produced divided by the number of words) is the same in both conditions (0.16). Concerning gesture duration, results are similar (gesture duration mean was 1.48 seconds with the younger adult partner and 1.66 seconds with the older partner).

In terms of the use of gesture space, Figure 2 shows that gestures were larger with the older partner (μ=2.43, SD=1.48) than with the younger partner (μ=3.19, SD=2.28). We used a Student's t-test and found a significant difference on occupation of gesture space depending on the interlocutors [t=-2.4578 ; df=169 ; p<0.01]. Thus, our hypothesis of larger occupation of gesture space with the older interlocutor is validated for this specific dyad.

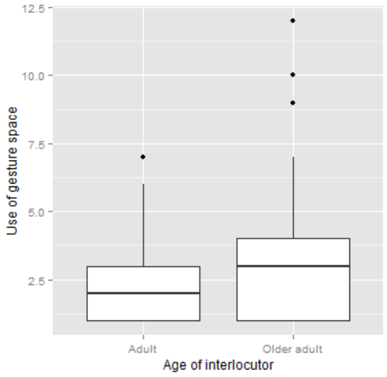


Figure 2 : *Use of gesture space depending on age of the interlocutor*

Moreover, in both conditions, the caregiver used more single-handed gestures than bimanual gestures. This can be attributed to the fact that the caregiver cannot produce gestures with the same hand that is holding her paper. Figure 3 shows that gestures were more bimanual in the senior condition than in the younger adult condition and the use of gesture space was significantly different depending on handedness with a proportion test [χ²(1) = 4.239 ; p <0.05]. As we counted all the crossed zones, gestures made with both hands were significantly larger than those made with one hand. Thus, two-handed gestures were more visible than one-handed gestures even if the opposite can sometimes be observed.

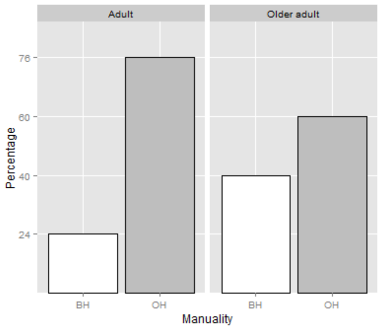


Figure 3 : *Percentage of handedness depending on the interlocutor*

Furthermore, the presence of initial or final phases has an effect on the use of gesture space. Indeed, gestures produced with preparation **and** retraction were larger than gestures produced with preparation **or** retraction phases. However, there is no significant difference concerning utilization on initial or final phases depending on interlocutor (Figure 4).

Concerning gesture depth, there was only one occurrence in the younger adult condition whereas there were 10 occurrences with the older partner. Thus, gestures in front of the speaker were most often used with the senior interlocutor. However, this difference is not significative but only a tendency [χ²(1)= 3.6828 ; p =0.06].

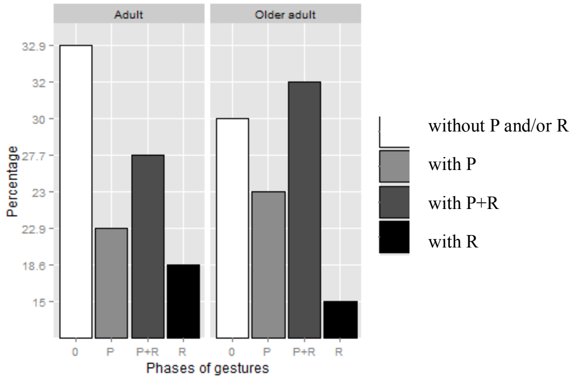


Figure 4 : *Percentage of phases of gestures depending on the interlocutor*

4. Discussion

This study is a preliminary step to rethinking the use of gesture space in terms of occupation of space rather than the location of gesture at one point of its production.

This methodological reflection must continue for two reasons. First, this scheme is used for a semi-controlled approach; the two speakers sit on chairs and cannot move. Thus, it is not adapted for all types of corpora. For example, Azaoui and Denizci [20] readapted the zones defined on McNeill's diagram because they used ecological data of teachers in action, and amplitude is larger when speakers are standing than when they are sitting. Moreover, if speakers move, we must replace the diagram. Secondly, the use of gesture space decreases in consecutive gestures because not all gestures have all the phases; only the stroke is obligatory [10] and gestures with initial and final phases are larger than others.

Another interesting possibility would be to annotate as Tellier & Stam [15] did in terms of location to see if there is any difference between the two approaches of coding the use of gesture space.

Finally, it would be helpful to know whether the older partners are aware of these gestures and if it really helps them to understand speech. For that, we could analyze the gaze of the interlocutor to know if s/he more often looks at gestures when they are larger than when they are small.

5. Conclusions

This study is a methodological reflection on the readaptation McNeill's [10] gesture space diagram to code the size of each gesture and not its location. For that, we created a new way of conceptualizing gesture space and we tested this method on a semi-controlled corpus that involves caregivers in an explanation task of 12 French words to both an older interlocutor and a younger one.

Preliminary results show that there is a difference in the occupation of gesture space when caregivers are addressing an older interlocutor *vs.* a younger one. Gestures are larger with an older younger partner. These results confirm our hypothesis that the caregiver accommodates her use of gesture space depending on the age of her interlocutor. It seems that gestures are more intentionally addressed to the older interlocutor to help when they have difficulties in guessing the word. However, these results are based on just one caregiver. Thus, we must be careful with these results and we need to analyze the other dyads to find out whether these tendencies are hold.

Therefore, to solidify our preliminary results we will code all data in the seven dyads and analyze the size of gestures depending on their dimension, duration and rate. This research has the potential to confirm the results obtained by Tellier & Stam [15] and reinforce the view that accommodation cannot only be defined by verbal and non-verbal characteristics but also in terms of coverbal gestural characteristics.

6. Acknowledgements

We thank Jorane Saubesty and Jérémy Laurent for data coding. We also thank Benjamin Holt for proofreading.

7. References

[1] C.A. Ferguson, "Baby talk as a simplified register," In *Talking to Children*, C.E. Snow & C.A. Ferguson, Eds. Cambridge : Cambridge University Press, 1977, pp. 209-235.

[2] M. Long, "Input, interaction, and second-language acquisition," *Native Language and Foreign Language Acquisition, vol.* 379, pp. 259-278, 1981.

[3] H. Giles *et al.* "Accommodation theory : Communication, context, and consequence," *Contexts of Accommodation : Developments in Apllied Sociolinguistics*, University of Arizona Library, 1991, pp. 1-68.

[4] E.B. Ryan *et al.* "Psycholinguistic and social psychological components of communication by and with the elderly," *Language and Communication,* vol.6, pp. 1-24, 1986.

[5] E.B. Ryan*, et al.* "Communication predicaments of ageing: Patronizing behavior towards older adults," *Journal of Language & Social Psychology,* vol.14, pp.144-166, 1995.

[6] M.L. Hummert, "Stereotypes of the elderly and patronizing speech," In *Interpersonal Communication in Older Adulthood,* M.L. Hummert, J.M. Wiemann, and J.F. Nussbaum, Eds. London, Sage, 1994, pp. 162-184.

[7] G. Cohen and D. Faulkner, "Does 'Elderspeak' work ? The effect of intonation and stress on comprehension and recall of spoken discourse in old age," *Language & Communication*, vol.6, no.1, pp. 91-98, 1986.

[8] E.B. Ryan, "Aging, Identity, Attitudes, and Intergenerational Communication," In *Understanding Communication and Aging: Developing Knowledge and Awareness,* J. Harwood, Eds. Thousand Oaks, CA: Sage Publications, 2007, pp. 73-91.

[9] M.A. Lowery, "*Elderspeak : Helpful or Harmful ? A systematic review of speech to elderly adults*," Ph.D. dissertation, Auburn University, Auburn, 2013.

[10] D. McNeill, "*Hand and Mind : What gestures reveal about thought,* " Chicago : University of Chicago Press, 1992.

[11] A. Galati, S.E. Brennan S.E. "Speakers adapt gestures to addressees' knowledge: implications for models of co-speech gesture," *Language and Cognitive Processes,* 2013.

[12] J, Bavelas, *et al.* "Gesturing on the telephone : Independent effects of dialogue and visibility," *Journal of Memory and Language,* vol.58, pp. 495-520, 2008.

[13] J. Gerwing, M. Allison, "The flexible semantic integration of gestures and words. Comparing face-to-face and telephone dialogues*," Gesture,* vol.11, no.3, pp.308-329, 2011.

[14] L. Mol, *et al*. "The communicative import of gesture: Evidence from a comparative analysis of human-human and human-machine interactions," *Gesture,* vol. 9, no.1, pp.98-127, 2009.

[15] M. Tellier and G. Stam, "Stratégies verbales et gestuelles dans l'explication lexicale d'un verbe d'action," In *Spécificités et diversité des interactions didactiques, V.* Rivière, Eds. Paris: Riveneuve éditions, 2012, pp. 357-374.

[16] A. Özyurek "Do Speakers Design Their Cospeech Gestures for Their Addressees? The Effects of Addressee Location on Representational Gestures," *Journal of Memory and Language,* vol.*46*, no.4, pp. 688-704, 2002.

[17] F. Cavicchio, S. Kita, "*English/Italian Bilinguals Switch Gesture Parameters when they Switch Languages,*" In Proceedings of TIGER: Tilburg Gesture Research Meeting, Tilburg, The Netherlands, 2013.

[18] D. Efron, *"Gesture, race, and culture,"* The Hague: Mouton, 1972. (Original work published as Gesture and environment, 1941.)

[19] C. Müller, "Gesture-space and culture," In *Oralité et gestualité. Interactions et comportement multimodaux dans la communication,* C. Cavé, I. Gauaïtella and S. Santi, Eds. Paris : L'Harmattan, 2001, pp. 565-571.

[20] B. Azaoui and C. Denizci, "*Segmentation du geste pédagogique et redéfinition de l'espace gestuel dans une approche écologique*," Actes, JETOU 2015, Toulouse, France, submitted for publication.

[21] J. Bressem, "*Notating gestures – Proposal for a form based notation system of coverbal gestures,*" unpublished manuscript, 2008.

[22] M. Tellier *et al.* "*Types de gestes et utilisation de l'espace gestuel dans une description spatiale : méthodologie de l'annotation,*" Actes, Atelier DEGELS, 18èmes conférence annuelle Traitement Automatique des Langues Naturelles Montpellier : Université de Montpellier II, 2011, pp. 45-56.

[23] A.K. Kuhlen *et al.* "Gesturing integrates top-down and bottom-up information: Joint effects of speakers' expectations and addressees' feedback," *Langage & Cognition*, vol. 4, pp. 17-41, 2012.

[24] N. Tan *et al.* (2010). "Multi-level Annotations of Nonverbal Behaviors in French Spontaneous Conversation," *In Procedings of Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, M. Kipp, J-C. Martin, P. Paggio and D. Heylen. Workshop held at the 7th International Conference for Language Resources and Evaluation, 17-23 May 2010, Malta, pp. 74-79. [Online]. Available : http://www.multimodal-corpora.org/mmc10.html.

[25] M. Tellier, "Quelques orientations méthodologiques pour étudier la gestuelle dans des corpus spontanés et semi-contrôlés," *Discours* [Online], vol. 15, 2014, Available : http://discours.revues.org/8917 ; DOI : 10.4000/discours.8917

[26] H. Sloetjes and P. Wittenburg, "*Annotation by category – ELAN and ISO DCR,*" Proceedings of the 6th International Conference on Language Resources and Evaluation, 2008.

[27] D. Casasanto and K. Jasmin, "The Hands of Time : Temporal gestures in English speakers," *Cognitive Linguistics,* vol. 23, no. 4, pp. 643-674, 2012.

[28] G.M. Pereira, "Gesture space and gestural handedness," *Saarland Working Papers in Linguistics*, vol. 4, pp. 41-56, 2013.

[29] M. Tellier, B. Azaoui and J. Saubesty, "*Segmentation et annotation du geste : Méthodologie pour travailler en équipe,*" Actes, JEP-TALN-RECITAL 2012 : Atelier DEGELS 2012, Grenoble, France, pp. 41-55, 2012.

[30] J.R. Landis and G.G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics,* vol. 33, no.3, pp. 159-174, 1977.

Shakes, nods, and tilts: Motion-capture data profiles of speakers' and listeners' head gestures

Bela Brenger, Irene Mittelberg

Natural Media Lab, Human Technology Centre, RWTH Aachen University, Germany
brenger@humtec.rwth-aachen.de, mittelberg@humtec.rwth-aachen.de

1. Abstract

Head gestures play an integral role in human communicative action. Speakers regularly employ head movements to convey approval, disagreement, or uncertainty, or to modulate the meaning of their utterances in other ways. Head gestures may also serve as backchanneling signals from the listener to the speaker. Due to their diverse discourse functions, head gestures have been investigated with a variety of foci (e.g., [1], [2], [3], [4]). This paper presents a novel methodology employing motion-capture technology to investigate the forms and functions of head gestures. The focus is on a) the extraction and analysis of specific physical and dynamic features found in head gestures and b) possible differences between speaker and listener head gestures.

1. Introduction

Head gestures are an integral part of communicative action performed by both speakers and listeners. Since they may modulate meaning expressed verbally and serve multiple discourse functions, head gestures have been analyzed with a variety of foci. As Heylen [2] states, "[w]hen one turns to the literature on head movements [...], one is faced with a bewildering list of functions and determinants of all the kinds of head gestures [...]" Despite this broad spectrum, however, head shakes and nods are the most commonly observed, conventionalized head gestures, exhibiting different forms and functions depending on the conventions of a given culture (e.g., [3]; [5]).

Combining qualitative and quantitative research methods, this paper investigates the basic form parameters of head gestures accompanying German discourse. It presents a methodology implementing motion-capture-technology aided extraction and analysis of these physical parameters. Special attention is paid to the difference in communicative action performed by speakers and listeners, as well as to the specific nature of listeners' versus speakers' head gestures. Motion-capture technology has proven apt at capturing and analyzing comparatively small manual movements and head gestures ([6], [7]). Our data set stems from the HumTec Multimodal Speech & Kinetic Action Corpus (MuSKA), consisting of multiple stream recordings (Motion Capture (MoCap), video, audio) of dyadic communicative situations in which participants performed three different tasks designed to engage them in natural conversation.

To account for the kinematics of the head movements and to derive initial motion-capture data profiles for distinct kinds of head gestures, the data analysis included basic physical form parameters such as amplitude, velocity and duration. For this pilot study, only conventional head gestures, namely head *nods*, *shakes* and *tilts* (towards the shoulder), were considered and coded. Speaker and listener gestures were distinguished depending on the local conversational role of person who performed the head gesture in question. Cross-referencing form parameters with the type of head gesture and their characterization as speaker or listener gesture allowed us to establish preliminary profiles for the communicative head action performed by speakers and listeners in dialogic exchanges.

Preliminary results show that both listeners and speakers use head nods frequently, albeit with a stronger predominance in the listener role. Moreover, the assumption of relatively higher complexity of speaker head gestures in comparison to listener head gestures was supported. Listener head gestures showed a shorter mean time of execution compared to speaker gestures; the latter more often consisted of composite gesture events exhibiting sequences of different gesture types. These first findings call for a more in-depth analysis of speaker and listener gestures. Subsequent work shall include the numerical analysis of spatial parameters of head gestures and their semantic and pragmatic relation to both the synchronously produced speech and manual gestures.

The paper begins by reviewing previous research on head gesture, and then presents the research question and methodology developed for the present study. In the final section, the first insights provided by this work are discussed, and avenues for follow-up studies are laid out.

2. Insights from previous research on head gestures

To date, much more research has been done on manual gestures as compared to head gestures. Previous work on head gestures has often focused on the form and function of prototypical or emblematic gestures such as head nods and head shakes. Kendon [3], for instance, suggests that – at least in Western cultures – head shakes seem to be tightly connected to "a 'theme' of *negation"* and modulate the meaning of utterances without being easily translatable into speech. Head nods represent another highly conventionalized and culture-specific communicative practice, which may be used emblematically for 'yes' or to generally express affirmation ([8], [9]). In his paper "Motor signs for 'Yes' and No'", Jakobson [10] put into relief the relation between the central European convention of 'nod-for-yes' and the opposing system used in Bulgaria, where *nods* are associated with negation. Other research into the linguistic functions of head movements has shown various ways in which they may serve deictic expressions, feedback requests, or as modality markers of uncertainty [5].

Head movements have also been ascribed the function of backchanneling: "Backchannel signals were initially identified in Yngve's study of turn-taking and were conceptualized as vocal or gestural expressions of the listener that do not signal his desire or intention to assume the floor" [5].

Motion-capture technology has been previously employed in head gesture research. Utilizing motion-capture aided methods, Kousidis et al. [11] observed a comparatively wide range of gesture inventories in co-speech head gestures produced by speakers in contrast to those performed by listeners. The authors further proposed a more fine-grained differentiation between head gestures produced by participants holding the floor in conversational exchanges and those produced by participants assuming the listener role. In spite of the often found focus on conventionalized and quasi-emblematic uses, Kousidis et al. [11] called attention to the fact that, similarly to manual gestures, head gestures tend to occur in concatenated units with up to 10 individual phases comprising different gesture types. It follows that these composite head gestures should also be treated as complex gesture units [12]. Furthermore, Ishi et al. [13] argue that, at least for Japanese speakers, the incidence rates for head gestures may also vary depending on the social relationship between the dialogue partners.

3. Motivation and research objectives

The aim of this research is to enhance our understanding of the form variants and communicative functions head gestures may exhibit in dialogic exchanges. Combining traditional video annotations with motion-capture data analyses opens up new avenues in gesture research in that we may analyze both single events in great detail (i.e., to the level of the millisecond and millimeter) and also search for patterns emerging from more extensive data sets.

By cross-referencing the temporal and spatial parameters of head motion provided by the numerical motion-capture data with the main types of the head gestures, head *nods*, *shakes* and *tilts*, our aim is to systematically investigate differences in the kinetic parameters of these different communicative behaviors. We are interested in how exactly, for instance, “[h]ead shakes vary in terms of the amplitude [...], in the number of rotations and in the speed [...]. There is no doubt that these variations in performance intersect with and modify the meaning of the gesture“ [3]. In addition, our approach allows us to look for systematic differences in the distribution of these conventionalized head gestures and to correlate their occurrence with the gesturer’s role as assumed in the conversation, that is, to distinguish between speaker and listener gestures. Based on the numerical data, differences in the kinematic characteristics of head gestures can be extracted and related to the conversational role as well, for example pertaining to a shorter/longer duration or lower/higher amplitude of head movements. Drawing on previous work ([11], [6]), our assumption is that *nods* are typical listener gestures, while speakers tend to employ a higher variety of head gestures.

It could further be observed that speakers seemed to cluster gestures and concatenate different types of head gestures into sequences, a practice that is a lot less frequently observed in listener gestures. This again raises the question whether there are systematic differences in the utilization of head gestures depending on the communicative role of the performer. The present study will provide first insights into these issues. We are aware, however, that to arrive at a well-founded conclusion, a larger dataset needs to be be analyzed in follow-up studies. This would also compensate for the idiosyncratic differences that are common and extensive in gestural behavior [7].

4. Methods of data collection and analysis

The video and motion-capture data used for this study stem from the HumTec Multimodal Speech & Kinetic Action Corpus (MuSKA). The corpus consists of recordings (MoCap, video, audio) of dyadic communicative situations in which different tasks were designed to encourage free conversation between the participants. Video and MoCap data were recorded by fourteen infrared cameras in the Vicon Nexus optical Motion Capture System, two Basler high-speed cameras (100Hz), two HD video camera and one SD video camera. Each participant wore a wireless microphone to record audio and a set of thirty-one infrared reflecting markers to track body movement with the help of the MoCap system.

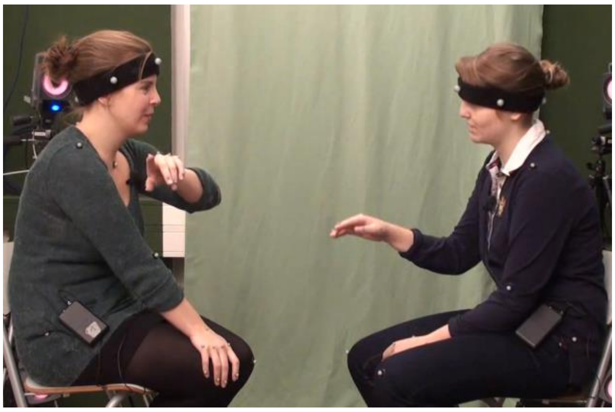


Figure 1: Position of participants in the MoCap volume

For this study, data from four head markers, one neck marker and two shoulder markers were extracted and analyzed. The head markers are connected to a head band for easier application and are aligned with the neck and sternum markers. This is to simplify the calculation of the head’s direction in further studies. The participants were positioned opposite each other with a distance of about 1.2 m between their chairs. This relatively close setup was chosen to also encourage interactive gestures, thus creating a shared gesture space. Generally, the MoCap system delivers more accurate results for a confined region of interest rather than a larger volume. The corpus encompasses recordings of conversations in both American English and German. For the study reported on here only German data were used.

For the first unstructured task, participants were asked to become acquainted with one another, should they not have met before, or to collaboratively remember a shared experience if they knew each other rather well. For the second task, the participants were instructed to collaboratively plan an *Interrail* journey through Europe. The conditions they had to work with were a limited budget, a three-week time limit and a maximum of 5 stays in places of their choice within the given time span. The participants were asked to agree on the itinerary of the trip and to discuss what kind of vacation they would prefer: for example, a sightseeing tour, a beach vacation or a hiking trip. As part of the third task, each participant was shown a short movie that they had to retell to their conversational partner. The data analyzed in this study only stem from the first two tasks.

For analysis, the video data were first viewed and annotated in ELAN for head movements regardless of their communicative function. These annotated gestures were reviewed for their predominant form and categorized as predominant *nods, shakes,* or *tilts.* A gesture was annotated from the beginning of the movement phase until the head stopped. In case of multiple cycles, for example in prolonged head shakes, they were counted as one occurrence of a gesture of the type *head shake*. An analysis of the number of cycles shall be part of further studies. Gestures that did not fit into the three categories were disregarded and will be investigated in a subsequent study. Body posture shifts and self-adaptors were not annotated.

The next step involved dividing all annotated gestures into *speaker* and *listener* gestures. A gesture accompanying ongoing speech production was considered a *speaker gesture*, while *listener gestures* typically were produced by the person not holding the floor. The latter are neither accompanied by speech nor aligned with the onset of speech. The timestamps of the beginning and end of each of these qualifying gestures were then used to identify the periods of interest in the motion-capture data that were subsequently analyzed in terms of the head's six degrees of freedom, velocity and amplitude. These degrees are composed of three axes and two types of movement, translations and rotations on each of them. Figure 2 illustrates the six degrees. Translations along the three translation axes result in the typical directions *forward—backward, up—down, right—left*. The three rotational movements around the respective axis are *pitch*, *yaw* and *roll;* these are equivalent to the movements that are more commonly called *nod*, *shake* and *tilt* in head gestures. To compensate for upper body movements (ventral/dorsal and lateral), e.g. posture shifts or repositioning on the seat, simultaneously to performing a head gesture, the head marker's motion was calculated relative to the motion of the neck marker. To compensate for rotational body movement, the shoulder markers were also used to generate a time dynamic representation of the body's orientation. The shoulder markers build a moving axis throughout the recording for each frame and in combination with the neck marker they constitute a moving coordinate-system. The head movement is always calculated relative to the adapting coordinate system. The amplitude of each gesture equaled the difference between the maximum and minimum coordinates on each axis for a given time period.

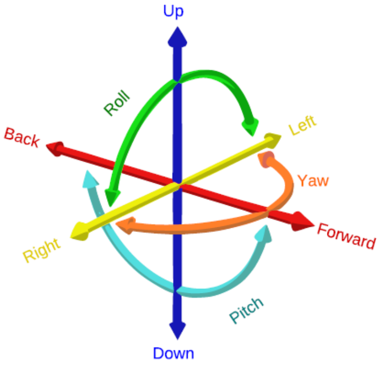


Figure 2: Six degrees of freedom

In addition to the video data, the MoCap system recorded all occurrences of head gestures in high temporal and spatial resolution, allowing for a numerical analysis of the form parameters of single gestures, as well as of recurrent types of gestures. Using the numerical data provided by the MoCap system, all identified gestures were coded in terms of the head's six degrees of freedom, velocity, amplitude and duration of motion. By coding all speaker-turns, we were further able to subdivide the observed behavior into speaker and listener head gestures, thus comparing the respective form parameters and deriving first profiles for each group.

5. Results

For this study, about 30 minutes of dialogue from four conversational tasks were annotated and labeled as described. This resulted in a total of over 740 occurrences of head gestures that qualified for further analysis.

Figures 3, 4 and 5 show examples of each gesture type. In particular, a listener' *head tilt* gesture, a listener *nod* and a *head shake* performed by a speaker. The system draws traces of selected markers for the period in which the gesture occurs.

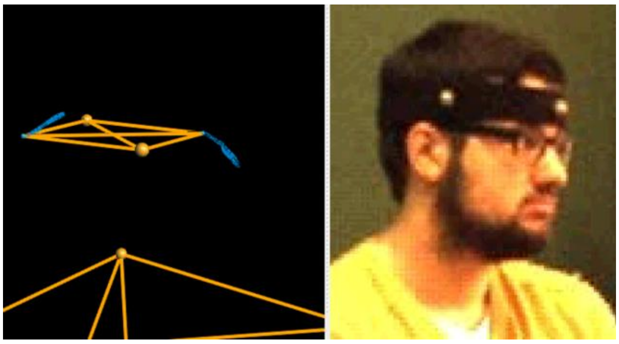


Figure 3: Trace of a *tilt* gesture performed by a listener

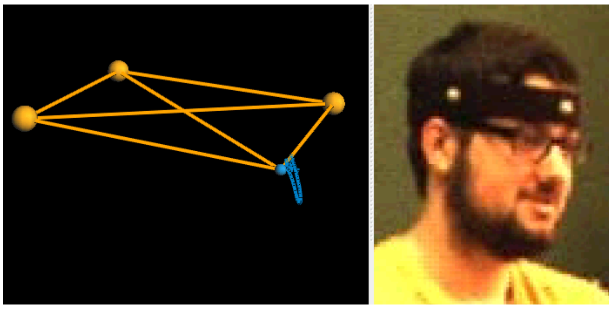


Figure 4: Trace of a *nod* gesture performed by a listener

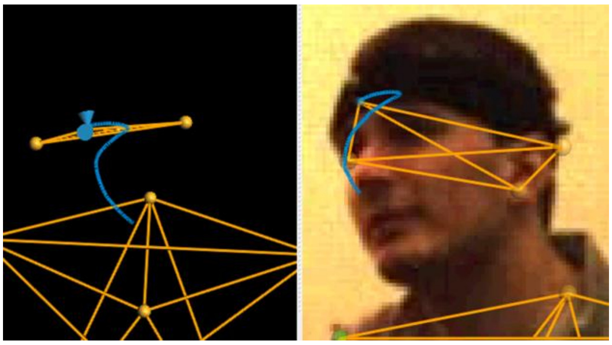


Figure 5: Trace of a *shake* gesture performed by a speaker

Of these head movements 130 were considered listener gestures. A more detailed distribution of head gesture type and mean duration is presented in Table 1 below. Each gesture type was separately categorized as a speaker or a listener gesture. So the total of 611 speaker gestures is subdivided into 327 *nods*, 191 *shakes* and 93 *tilts*. For each subcategory, the mean duration of all occurrences was calculated as well.

Type	Quantity	Mean duration [seconds]
Speaker nods	327	0,78
Speaker shakes	191	0,79
Speaker tilts	93	0,65
Speaker total	**611**	**0,75**
Listener nods	93	0,61
Listener shakes	11	0,62
Listener tilts	26	0,66
Listener total	**130**	**0,62**
Total	**741**	**0,72**

Table 1: Quantity and mean duration of head gestures

Head nods were the most frequent gesture type with a comparably stronger predominance in listener behavior. A distinctive feature of this data set was the low number of head shakes attributed to the listener role. A possible explanation might be that the participants are instructed to get to know each other and to work collaboratively in order to come up with a joint solution to the travel task. In light of the "'theme' of *negation"* that Kendon [3] attributes to head shakes, participants might have been inclined to reduce them to a minimum to avoid slowing down or compromising the ongoing activity.

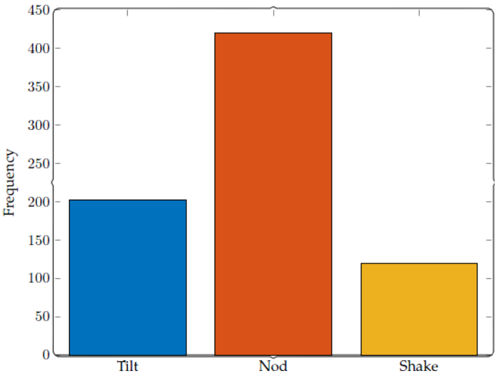


Figure 6: Distribution of gesture types

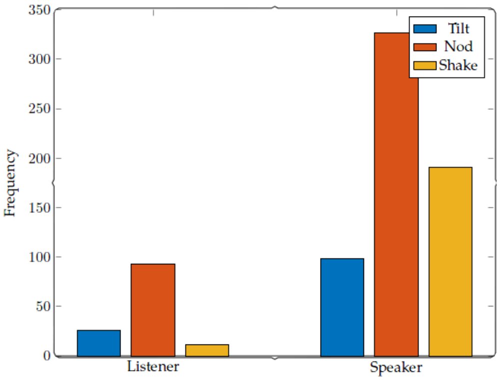


Figure 7: Distribution of listener and speaker gestures

The amplitude of the analyzed gestures varied greatly for all axes and for all gesture types. As such, the question regarding the range of motion of listener head gestures in comparison to speaker head gestures remains open. However, the data did show a tendency towards stronger, longer and more articulated speaker gestures. Figures 8 and 9 show data examples of a multi-cyclic head nod respective head shake performed by a speaker. The data show the velocity of one head marker over the duration of the head gesture. The time resolution is 100 frames per second; the velocity is split into the three axes. Nods and shakes show stronger activation on their primary axis.

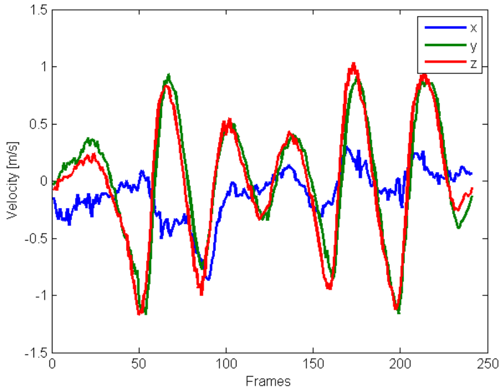


Figure 7: Velocity of cyclic head nod (speaker)

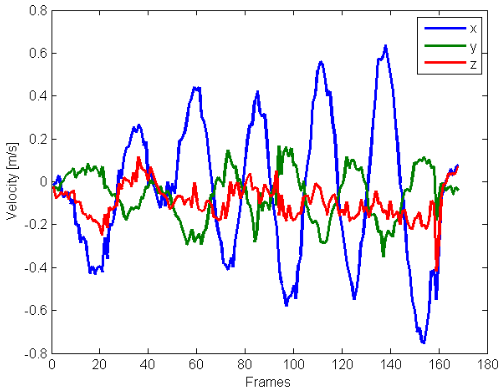


Figure 8: Velocity of cyclic head shake (speaker)

The speakers' tendency towards a more active employment of head gestures also presents itself in the utilization of complex head gesture units [11]. Complex head gesture units are concatenations of two or more gestures without pause. 45 of those units were identified. Even with this small data sample it becomes apparent that these units have a strong predominance in the speaker role. About 85% of these units corresponded with the speaker role, and they were rarely identified in the listener role (see Table 2). This assumption fits with the observed slight tendency toward shorter mean duration in listener gestures, making listener gestures appear overall more subtle and singular in their execution.

Type	Quantity	Percentage
Complex HGU **Speaker**	61	84,7
Complex HGU **Listener**	11	15,3
Total	72	

Table 2: Distribution of complex head gesture units

6. Discussion

Our analysis has revealed differences in listener and speaker gestures not only in terms of their frequency and distribution, but also regarding their manner of execution. However, the results obtained in this pilot study only reveal tendencies. In particular, the strong variation in the range of motion makes it difficult to draw conclusions concerning systematic, form-specific distinctions between head gestures that depend on the performer's conversational role. Specifically, the observed listener and speaker gesture characteristics for frequency and distribution, in addition to the observed relatively higher complexity of speaker head gestures, calls for further in-depth analyses of the internal structure and multiple functions of both speaker and listener gestures. As a first step, the methods should be adapted to account for larger data sets and to tackle the strong variation in the amplitude of the gestures.

Furthermore, the analysis of concatenating head gestures should be extended to sequences, since this phenomenon was observed to show a regular and strong bias towards the speaker role. The segmentation of all the annotated head gestures into different phases that can be analyzed individually might be one approach, as has also proven useful in the case of manual gestures ([14], [15]). This would further improve the integration of concatenated head gestures and enable a more fine-grained analysis of their relation to the synchronously produced speech as well as manual gestures. Moreover, the analysis of a larger data set would enable us to test for statistical significance and allow for the investigation of smaller subsets, such as head shakes and tilts performed by the listener. Finally, extending the numerical analysis of dynamic spatial parameters to head gestures that do not exactly fit the conventional types or consist of combined profiles is a promising avenue for further research in the domain of communicative action performed with the head, arms and torso by both speakers and listeners.

7. Concluding remarks

Motion-capture aided tracking of head gestures and the subsequent generation of head gesture data profiles showed promising results. Firstly, the numerical data reflects the conversational role in which the gesture was uttered. Differences in the data profiles for listener and speaker head gestures occurred systematically and encouraged the separate analysis of these conversational roles. Secondly, the data profiles of gesture types within one conversational role are employable to distinguish singular gesture types.

Through further elaboration of the data extraction and profile generation methods these profiles may be employed for semiautomatic segmentation or structuring of conversations as well as fine-grained qualitative analysis of singular gesture occurrences. However, the overlapping of gesture units, the high variation in amplitude and velocity as well as idiosyncratic gesture styles make a fully automated characterization of these gestures difficult. Moreover, the inventory of head gestures reaches beyond the scope of the simplified selection presented here. The method propsed in this paper can nonetheless be adapted to a more fine-grained analysis with a larger inventory of gesture types and their respective data profiles. The next steps of this research will include the refinement of data profile generation, with a focus on the normalization of the data and the development of methods to analyze larger datasets to enable further statistical analysis.

8. Acknowledgements

This study was funded by the Excellence Initiative of the German Federal and State Governments. We thank the members of the Natural Media Group and Monica Gonzalez-Marquez for valuable input at various stages of this project and the study participants for supporting our research

9. References

[1] Debras, C., Cienki, A. “Some Uses of Head Tilts and Shoulder Shrugs during Human Interaction, and Their Relation to Stancetaking”, ASE International Conference of Social Computing, Amsterdam, Netherlands, 2012.

[2] Heylen, D. “Challenges ahead: head movements and other social acts during conversations” ,45–52, 2005.

[3] Kendon, A. “Some Uses of the Head Shake”, Gesture 2 (2): 147–82, 2002.

[4] Wagner, P., Malisz, Z. and Kopp, S. „Gesture and speech in interaction: An overview”, Speech Communication 57. 209–232, 2014.

[5] McClave, E. Z. “Linguistic Functions of Head Movements in the Context of Speech”, Journal of Pragmatics 32 (7): 855–78, 2000.

[6] Alexanderson, S., House, D. and Beskow, J. “Extracting and Analyzing Head Movements Accompanying Spontaneous Dialogue”, Proceedings of Tilburg Gesture Research Meeting, Tilburg (TiGeR), 2013.

[7] Priesters, M. A., Mittelberg, I. “Individual differences in speakers' gesture spaces: Multi-angle views from a motion-capture study”, Proceedings of the Tilburg Gesture Research Meeting (TiGeR), 2013.

[8] Efron, D. “Gesture and Environment: A Tentative Study of Some of the Spatio-Temporal and “Linguistic” Aspects of the Gestural Behavior of Eastern Jews and Southern Italians in New York City, Living Under Similar as Well as Different Environmental Conditions.” King’s Crown Press, 1941.

[9] Ekman, P., Friesen, W. V. “The Repertoire of Nonverbal Behavior Categories, origins, usage, and coding”, Semiotica (1): 49–98, 1969.

[10] Jakobson, R. “Motor signs for ‘yes’ and ‘no’” Language in Society, 1(01), 91-96, 1972.

[11] Kousidis, S., Malisz, Z., Wagner, P. and Schlangen, D. “Exploring Annotation of Head Gesture Forms in Spontaneous Human Interaction”, Proceedings of the Tilburg Gesture Meeting (TiGeR), 2013.

[12] Buschmeier, H., Malisz, Z., Skubisz, J., Kopp, S. and Wagner, P. “ALICO: a multimodal corpus for the study of active listening”, Proceedings of language resources and evaluation conference, 2014.

[13] Ishi, C. T., Ishiguro, H. and Hagita, N. “Analysis of relationship between head motion events and speech in dialogue conversations”, Speech Communication 57: 233–243, 2014.

[14] Kendon, A. “Gesture: Visual Action as Utterance”, Cambridge: Cambridge University Press, 2004.

[15] Kita, S., Van Gijn, I. and Van der Hulst, H. “Movement phases in signs and co-speech gestures, and their transcription by human coders”, in I. Wachsmuth and M. Fröhlich (Ed), Gesture and Sign Language in Human-Computer Interaction, 23–35, Springer: Berlin Heidelberg, 1998.

Structuring and highlighting speech – Discursive functions of holding away gestures in Savosavo

Jana Bressem[1], *Nicole Stein*[2], *Claudia Wegener*[3]

[1] Faculty of Humanities, Technische Universität Chemnitz, Germany
[2] Faculty of Social and Cultural Sciences, European-University Viadrina, Germany
[3] Faculty of Linguistics and Literary Studies, Universität Bielefeld

jana.bressem@phil.tu-chemnitz.de, n_stein@gmx.de, claudia.wegener@uni-bielefeld.de

Abstract

Based on a study investigating gestures used for the expression of refusal, rejection, exclusion and negation in Savosavo, a Papuan language spoken in Solomon Islands in the Southwest Pacific, the article discusses how a particular type of pragmatic gesture, the *holding away* gesture, may highlight and structure the spoken utterance. It will be shown that the holding away gesture assumes three functions on different levels of discourse: It emphasizes the speaker's focus on the *conclusion and change* of a topic. It highlights the *contrast* between two propositions or emphasizes that the speaker is *inserting* additional information. The article demonstrates that holding away gestures operate on the spoken utterance and take over speech-performative function as they draw attention to the communicative act the speaker is engaged in and, at the same time, make this communicative action visually accessible to the hearer.

Index Terms: multimodality, speech, pragmatic gestures, discourse markers, discourse structure, Savosavo

1. Introduction

Particles fulfill a range of functions in spoken language. Modal particles, such as *denn, halt,* or *eben* in German, for instance, operate on the pragmatic-functional level of the utterance and "integrate utterances into the realm of interaction. [With modal particles], speakers can refer to shared knowledge, to assumptions or expectations of speakers or hearers, a particular reference to a preceding utterance can be marked or the significance that the speakers attest to the utterance can be marked. Modal particles thus modify illocutionary types in particular ways" [1: 2, translation authors]. Furthermore, particles assume a major function in the regulation of interactional processes and display the discursive structure of the utterance. In English, discourse particles or discourse markers, *well, but, unless,* or *then*, for instance, are expressions connecting parts of discourse. Similar to modal particles, they do not express propositional content but rather contribute to the interpretation of the utterance because "they signal a relationship between the segment they introduce, S2, and the prior segment, S1" [2: 950]. They connect messages and may either emphasize contrast (*but*), a quasi-parallel relationship between messages (*furthermore*) or they mark elaborations (*well*) and inferences (*then*). Furthermore, discourse particles may not only connect messages but rather topics and as such are of importance for managing discourse. 'Topic change markers' [2] highlight a thematic excursion or the reintroduction of a previous topic. These functions can, as Schiffrin notes, not only be realized by verbal expressions but also by paralinguistic elements (e.g., prosody) and gestures [3].

Research has shown that gestures with pragmatic functions are able to "relate to features of an utterance's meaning that are not a part of its referential meaning or propositional content" [4]. As such, gestures fulfill performative function by indicating a request, a question or refusal [e.g., 4, 5, 6]. Furthermore, they may "serve in a variety of ways as markers of the illocutionary force of an utterance, as grammatical and semantic operators or as punctuators or parsers of the spoken discourse." [4: 5]. By taking over modal function, gestures indicate the speaker's stance towards the proposition uttered [4-8]. They qualify something as negative, obvious or particularly noteworthy and thus operate on the speaker's own utterance. Accordingly, researchers have argued that such gestures show functional analogies with modal particles [7-9]. However, gestures with pragmatic function may not only be an indication for the speaker's attitude towards the proposition of the utterance but also have the capability of highlighting properties of discourse. By taking over 'parsing' [4] or 'interactive' function [10], gestures contribute to the marking of various aspects of the structure of spoken discourse and provide visible anchor points for connecting or separating parts of discourse [see also 11]. Accordingly, Kendon [12: 248] has discussed pragmatic gestures with discursive function as 'discourse unit markers', highlighting the fact that gestures may be able to "mark discourse units differentially as *topic* in contrast to *comment*" and may serve to "mark discourse units which are 'focal' to the theme or argument of what is being said". In doing so, gestures with pragmatic functions may have the same functions as discourse markers or rising intonation in spoken language [10].

The present article ties in with existing research on the discursive nature of pragmatic gestures. Based on a study investigating gestures used for the expression of refusal, rejection, exclusion and negation in Savosavo, a Papuan language spoken in Solomon Islands in the Southwest Pacific [13, 14], the article discusses how a particular type of gesture, the *holding away* gesture (see Figure 1), may highlight and structure the spoken utterance. The holding away gesture has been discussed in a range of studies on pragmatic gestures. Bressem and Müller [15] present an analysis of the gesture as part of the away family, gestures used by German speakers to express negation, refusal and negative assessment. The authors show that the holding away gesture is used to reject topics of talk, to stop arguments, beliefs or ideas from intruding into the realm of shared conversation and to stop the continuation of unwanted topics. Moreover, it qualifies the rejected topics as unwanted ones.

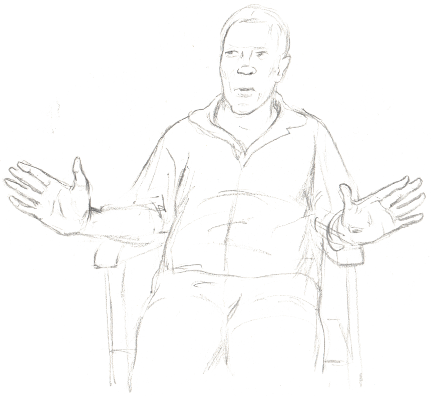


Figure 1: *Holding away gesture in Savosavo*

In a similar vein, Kendon discusses the holding away gestures as part of his account of gestures used by speakers of English and Italian "in contexts where something is being denied, negated, interrupted, or stopped" [4: 248]. With the *Open Hand Prone VP*, the speaker establishes a barrier, pushes back or holds back things moving towards him- or herself. The gesture indicates the speaker's "intent to stop a line of action" [4: 262]. Depending on the position of the hands, the gesture specifies the kind of action to be stopped: 1) close to the body: stopping ones own action, 2) in front of the body: stopping the action of the speaker and the interlocutor, 3) movement towards the interlocutor: stopping the action of the interlocutor. Also for speakers of English, Harrison identifies different variants of the gesture by which speakers may refuse or interrupt themselves or others (*PVraise*), express positive evaluation, apology or negation (*PVoscillate, PVhorizontal*) [16]. For speakers of French, the gesture is also documented as carrying the semantics of rejection and being used by speakers to actively refuse something [17: 200].

Research thus demonstrates that the holding away gesture is characterized by a variety of forms and functions across different Indo-European languages. However, these studies have primarily concentrated on its performative or modal use. The gestures' relevance for marking various aspects of the structure of spoken discourse has not yet been addressed in detail. The present article aims to fill this gap by presenting a first analysis of the discursive function of holding away gestures in Savosavo.

2. Savosavo language

Savosavo is the easternmost of only four (at best distantly related) non-Austronesian (Papuan) languages spoken among more than 70 Austronesian languages in Solomon Islands. The Savosavo speech community comprises about 3,500 people living on Savo Island, a small volcanic island approximately 35km northwest of the capital Honiara.

3. Database and methods

The holding away gestures were identified in a corpus consisting of 68 hours of video recordings from 84 different speakers (52 male, 32 female), ranging in age from about 20 to about 85, collected during Wegener's PhD fieldwork and the Savosavo Documentation Project (see [13] and the project website http://dobes.mpi.nl/projects/savosavo/ for more detail). It is stored in the DoBeS archive at the Max Planck Institute for Psycholinguistics in Nijmegen, and can be accessed under https://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI55379 9%23. For the analysis of the holding away gestures, 6 hours of video recordings from the total of 68 hours of video recordings were chosen, consisting of mostly narratives, some procedural texts as well as a few interviews. The corpus comprises monologic, dyadic as well as group constellations of altogether 14 male speakers ranging in age from 39 to about 80. Altogether, 56 instances of the holding away gesture were identified. The holding away gestures were analyzed within a form-based linguistic approach also adopted for analyses of holding away gestures in German [15]. Accordingly, the analysis of the sweeping and holding away gestures in Savosavo consisted of a 4-step procedure [18]. The gestures were first annotated and coded in their form. Subsequently, the gestures were analyzed in relation to the verbal utterance. Here the gestures' meaning and function was examined with respect to the sequential, syntactic, semantic as well as pragmatic information given by speech but also by semantic and pragmatic information conveyed by adjacent gestures. In a next step, the analysis of the local context, i.e. the interactive environment of a gesture, was combined with an analysis of its context-of-use, the broader discursive situation in which a recurrent gesture occurs [4, 19]. The determination of the contexts-of-use built the basis for the fourth step, i.e. the distributional analysis of the gestures, the identification of gestural variants and the detection of a systematic correlation of context-of-use and variations of form [20]. The gesture annotation was either incorporated into existing ELAN files with morpho-syntactic annotations [13] or new ELAN files were set up. In the latter case, morpho-syntactic annotations for Savosavo were later added at and around those points in time where the gestures under investigation occurred. The distributional analysis was done using an Excel data basis.

The analysis of the gestures in relation with speech and the determination of the different contexts-of-use were conducted in collaboration with a native speaker of Savosavo, because non-linguistic context, such as background information on cultural, geographic, historical and other specific aspects of the life on Savo, is crucial to the understanding of speech and gestures. Moreover, in particular for the analysis of gestures with pragmatic functions, native competence of the language is indispensible in order to catch all of the gesture's relevance and function for expressing the illocution of the utterance.

According to this procedure, different context variants of the holding away gesture and, in particular, specific functions of the holding away gestures for highlighting and structuring discourse were identified.

4. Holding away gestures in Savosavo

The holding away gesture in Savosavo is characterized by a particular formational core that is kept stable across speakers and contexts-of-use: The (lax) flat hand(s) with the palm oriented vertically away from the speaker's body are held in the center of the gesture space. This formational core can be varied, so that the hands may be moved away from the speakers body (cf. [4]) or moved downwards (see example 1, 2). The palm of the hands may be oriented diagonally downwards and the hands can be positioned in different regions of the gesture space (see [14] for more details). In accordance with existing research we assume that the formational core of the holding away gesture is derived from an underlying everyday action, such as the action of holding or pushing away an object, stopping a door from smashing into the face, or an unwanted person from intruding into the personal space. The vertically oriented hand(s) create a blockage, which either keeps objects from moving closer or pushes them away [15]. As a result, annoying or otherwise unwanted objects are hindered from entering the space around the body. This effect of action is semanticized in the holding away gesture: Something wanting to intrude has been or is being kept away from intrusion. As such, the gesture is used to “reject topics of talk, to stop arguments, beliefs, or ideas from intruding into the realm of shared conversation, to stop the continuation of unwanted topics” [15: 1598].

We documented 56 holding away gestures, which are used in 3 different contexts-of-use (see Table 1): explanation (34, 61%), request (20, 36%), and description (2, 3%). In descriptions, speakers describe the characteristics and processes of (historical) events, fishing techniques or rituals, for instance. In explanations, speakers add one or more statements to clarify or explain something (e.g., a particular cultural aspect potentially unknown to a foreigner) or to give a reason or justification for an action (e.g., the end of a war or the duration of a particular event). In the context-of-use ‘request’, speakers fulfill the speech act of asking for something. Here, the gestures function as ‘performatives’ as they “aim at a regulation of the behavior of others” and ‘perform’ the illocutionary force of an utterance [8].

Context-of-use	Function of gesture		Number of instances		
explanation	*speech-performative*	*topic shift*	17	34	n=56
		contrast	10		
		insert	5		
	abstract-referential		2		
request	*performative*			20	
description	*speech-performative*	*topic shift*	1	2	
	abstract-referential		1		

Table 1: *Overview of contexts and functions of holding away gesture*

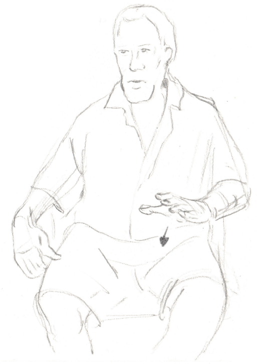
In the examples from the context-of-use ‘request’, gestures are executed in temporal overlap with speech and request others to *stay* in a particular place (e.g., “don't you come ashore here” ak_biti_630) or are used as an *appeasement* (e.g., “I am not harming anyone” ap_cs_kabulabu_552). When used without speech, the holding away gesture requests someone to be *quiet*, to *stop* an ongoing action (e.g., talking while someone else is talking), or to *keep* someone from starting an action (e.g., to give further information on a topic) [for more detail see 14].

As shown in Table 1, the holding away gesture is most common in the context-of-use ‘explanation’. 34 instances of the gestures are used when speakers provide explanatory statements or justify actions or events. In 2 instances, speakers employed the gestures to enact the stopping of events or actions that are in progress or are about to start. However, the majority of holding away gestures takes over speech-performative, discursive function. We will discuss this use in detail in the following section.

5. Structuring and highlighting discourse

94% of the gestures in the context-of-use ‘explanation’ (32 instances) fulfill speech-performative function and thus act upon the speaker’s own utterance [8: 1544]. In these cases, “gestures are aligned with what the speaker is presently doing, and convey something about it” [21: 74]. They display the communicative act of the speaker and visualize the structure of the spoken utterance. In our corpus, holding away gestures take over three different functions for marking aspects of the spoken discourse: They mark a *conclusion and change* of topic, highlight the *contrast* between two propositions or emphasize that the speaker is *inserting* additional information.

In the first example, we see an instance in which the holding away gesture visually marks the conclusion of one topic, and, at the same time, marks the change to another topic. While talking about the last war on Savo and an important warrior, speaker DE explains the Sepe dance, which was inspired by this warrior and is performed on the island of Savo. After having finished describing the dance, its characteristics and explaining who performs the dance, the speaker utters “that is the Sepe dance” and at the same time produces a holding away gesture encompassing almost the whole phrase (see example 1). Afterwards, he continues his narration with another aspect of the story about the last war on Savo. In this example, speech and gesture work together in marking the closing of a topic and indicate that the speaker’s explanation about the Sepe dance has come to an end. The vertically oriented hand, which is movement downwards with a short accentuated movement, sets up a barrier in front of the speaker’s body, blocking any requests for further explanations of the topic of the Sepe dance. The gesture takes over meta-communicative function by operating on the concurrent speech and displaying the communicative act of the speaker, namely his intention to end the story of the Sepe dance and his goal to move on to a different aspect of the overall topic.



(1)	*Lole*	***lo***	***Sepena.***
	lo=le	*lo*	*Sepe=na*
	3SG.M=EMPH.3SG.M	DET.SG.M	Sepe=NOM
	PP	ART	N
		G1	G1

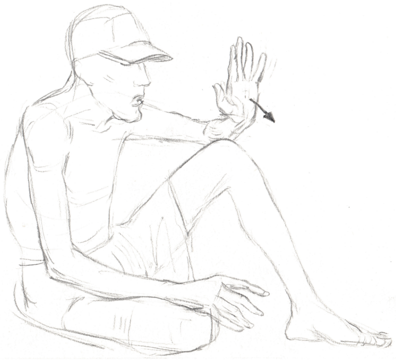
“That is the Sepe dance.” (de_torolala_425)

G1: The left flat hand, palm oriented diagonally vertically away from the speaker’s body, is moved downwards in the lower center of the gesture space.

Example 1: *Holding away gesture highlighting the conclusion and change of topic.*

In doing so, the gesture takes over a similar function as observed for discourse markers in spoken languages: The gesture functions as a topic-relating discourse marker [2]: Through the holding away gesture, the topic of the present utterance (the Sepe dance) and the topic of the following utterance (last war on Savo) are set in relation. The gesture helps to structure the discourse in terms of topic management. This is an interesting difference to studies of other languages, which usually show how pragmatic gestures operate on the topic-comment structure of one utterance (e.g., [12]). In our corpus, the holding away gesture does not indicate the topic or comment portion of one particular utterance, but rather sets two different discourse topics in relation, marking the change from one topic to another. In this and other examples, when speakers use the holding away gestures with the function of indicating a change of topic, it is accompanied by a closing statement on the present topic (e.g., “that is the Sepe dance”, “that is what they say” si_kuarao_1532, “that is a different story” jn_lotu_103) before picking up another topic.

A second function can be observed in the following example, in which the gesture does not function as a topic-relating discourse marker, but focuses on the message and is used to set up a contrast between two propositions. In example 2, speaker PNG talks about the length of the Second World War in Solomon Islands. He counts the years during which the fighting went on and concludes that it was only three years. While uttering “only for three years”, the speaker performs a one handed holding away gesture by which he sets up a visual barrier blocking off any objection from his interlocutors and metaphorically holds away possible arguments or counter-examples meant to contradict his explanation. Here again, the gesture operates on the speaker’s own utterance, yet this time it indicates that the speaker is setting up a contrast between his utterance and a contradicting alternative: The gesture establishes a contrast between the actual duration of the Second World War in Solomon Islands mentioned by speaker PNG and a potentially expected longer duration as compared to other countries, for instance. The gesture operates on the message of the utterance and not, as in example 1, on the topic.



(2)

Omalo	*gneqai*	*ata;*	***kede***
oma=lo	*gneqa-i*	*ata*	*kode*
no=3SG.M.NOM	be.long-FIN	here	only.NSG
NEG=PP	V	LOC	QUAN
			G1

ighia	*eleghoghalalo*	*te*
ighiva	*elegho=gha=la=lo*	*te*
three	year=PL=LOC=3SG.M.NOM	EMPH
QUAN	N=PP	PA
G1		

ata	*palei.*
ata	*pale-i*
here	stay-FIN
LOC	V

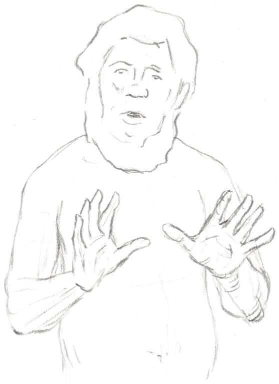
"It wasn't long here, only for three years it stayed here.” (png_WWII_1_628)

G1: The left flat hand, palm oriented diagonally vertically away from the speaker’s body, is moved downward in the upper center of the gesture space.

Example 2: *Holding away gesture setting up a contrast between propositions*

In other examples of this kind in our corpus, speakers set up a contrast between a fishing taboo mentioned in the present utterance and other potential fishing taboos (“The only taboo is that which I said earlier, stepping over the string and (all) that.” si_kurao_746) or between different types of custom money owned by people of different status (“not the custom money that the young people or the normal people would own, the important people only” ap_seka_547). In all cases, the holding away gesture seems to show a functional analogy to contrastive discourse markers in spoken languages by which an “explicit message of [an utterance] is in contrast with an […] implied message [of another utterance]” [2: 947].

In example 3, we see the third discursive function of the holding away gestures documented in our corpus. Here, the gesture indicates that the speaker is departing from his main story line and is inserting additional information.



(3)

Pozogho	***dol**oghu*	*pai kia*
pozogho	*dolo-ghu*	*pai kia*
basically	be.friend-NMLZ	or.maybe
ADV	N	CONJ
	G1	

zughuzughu	*abagnighu*
zughu~zughu	*abagni-ghu*
NMLZ~disagree	argue-NMLZ
N	N

"basically, peace, or otherwise disagreement and arguments(, or otherwise anything)” (jn_lotu_349)

G1: Both hands, palm oriented vertically away from the speaker’s body, are moved downwards in the center of the gesture space.

Example 3: *Holding away gesture setting up a contrast between propositions*

Speaker JN tells the story of the first arrival of missionaries on Savo Island and describes how a group of elderly women communicates with two missionaries. As neither of the groups speaks the language of the other, the elderly women and the missionaries communicated by using their hands. After having uttered “because of that they only used their hands to make signs”, the speaker inserts some further information, explaining what could have been the topic of their conversation. While saying “basically, peace, or otherwise disagreement and arguments, or otherwise anything, only with the hands did they talk about it on that day”, he produces a holding away gesture in temporal overlap with “peace”. Here, the hands visually mark the point in time where the additional information is added. After having uttered “peace”, speaker JN lists some further topics of talk (disagreement, arguments). By being executed in temporal overlap with the first item listed, the holding away gesture highlights the part of the utterance inserting additional information and thus visually foregrounds the insertion. In spoken English, for instance, discourse markers such as *furthermore, in addition* or *namely*, highlight that the present utterance is "adding yet one more item to a list of conditions specified by the preceding discourse" [2: 948]. Considering example 3, a similar function can be attested to the holding away gesture. Here, the two vertically oriented hands visually mark the point in time where additional information is given to provide some further elaboration on the possible topics discussed by the women and the missionaries. In another example in our corpus, the gesture is used when a speaker talks about magic and adds an aside, specifying a particular type of magic ("vele magic, that custom thing, vele magic they took" png_WWII_3_1616).

All of the discussed examples above illustrate that the holding away gesture is able to operate on the level of the message, when setting up a *contrast* or *inserting* information. Yet it can also be used as a topic-relating discourse marker when emphasizing the speaker's focus on the *conclusion* of a topic and the subsequent topic *change*. By doing so, holding away gestures relate discourse segments and do not contribute to the propositional meaning of either segment. Rather, they operate on the pragmatics of the spoken utterance by embodying communicative actions and discourse structure. The holding away gesture displays the communicative act the speaker is engaged in and, at the same time, provides a clue to the listener on how to treat the respective information and to refrain from possible counter arguments. The meaning that is expressed by the gestures is thus mainly a procedural one, specifying how segments of an utterance are to be interpreted relative to the each other. Following Kendon, it can be concluded that pragmatic gestures, or in the present case, holding away gestures "appear to serve as if they are labels for segments or units within a discourse, thereby indicating the part these units play within the discourse structure" [12: 264] for the speaker and the hearer.

6. Conclusion

Based on an analysis of a particular type of pragmatic gesture used by speakers of Savosavo, the article elaborated on the relevance of pragmatic gestures for highlighting and structuring discourse. Taking up Fraser's pragmatic classification of discourse markers, it was shown that the holding away gesture assumes a diverse function on different levels of spoken discourse structure in Savosavo. The gesture may operate on the level of the message of the utterance or it puts topics of different utterances in relation to each other. By doing so, holding away gestures act on the spoken utterance and take over speech-performative function as they highlight the communicative act the speaker is engaged in and make this communicative action visually accessible for the hearer. Holding away gestures with discursive function thus take over particular communicative relevance as they not only regulate discourse but also clarify discourse structures for speaker and hearers by drawing attention to speech act sequences, cohesion and thematic relations.

Taking up the analysis presented in this article, a comparison of the functions identified for the holding away gestures in Savosavo with other languages would be particularly interesting for gaining further insights into the nature of the holding away gestures, pragmatic gestures in general and their discursive potential. Regarding performative functions of the holding away gestures, a cross-cultural and cross-linguistic distribution can be identified. Speakers of Savosavo use the gestures in a very similar way as speakers of German, English, or French, for example. Their formational features as well as their semantic and pragmatic characteristics match those described by other researchers (see [4, 15-17]). The documented forms, meanings, and functions thus seem not to be restricted to their use in Indo-European languages but might have a rather wide cross-linguistic and cross-cultural distribution [see 14 for more detail]. Investigating the discursive function of the holding away gestures across a range of different languages would provide a further puzzle piece for language specific or possible universal functions of pragmatic gestures. Examining the relevance of gestures for discourse structure thus poses an interesting field of research by which further insights into the nature of pragmatic gestures can be gained and, furthermore, on the relevance of gestures for establishing multimodal utterances.

7. Acknowledgments

We are grateful to the Savo speech community for supporting this research. We would also like to thank Edmond Gagavo for his assistance in analyzing the gestures and Mathias Roloff for providing the drawings (www.mathiasroloff.de). This research was supported with a grant from the Volkswagen Foundation (DOBES – Documentation of Endangered Languages). For further information on the Savosavo DOBES project see http://dobes.mpi.nl/projects/savosavo/.

8. References

[1] M. Thurmair, *Modalpartikeln und ihre Kombinationen* vol. 223: Walter de Gruyter, 1989.

[2] B. Fraser, "What are discourse markers?," *Journal of pragmatics,* vol. 31, pp. 931-952, 1999.

[3] D. Schiffrin, *Discourse markers*: Cambridge University Press, 1987.

[4] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press., 2004.

[5] C. Müller, *Redebegleitende Gesten: Kulturgeschichte, Theorie, Sprachvergleich*. Berlin: Arnold Spitz, 1998.

[6] J. Streeck, *Gesturecraft. The Manu-facture of Meaning*. Amsterdam: John Benjamins, 2009.

[7] C. Müller and G. Speckmann, "Gestos con una valoración negativa en la conversación cubana," *DeSignis,* vol. 3, pp. 91-103, 2002.

[8] S. Teßendorf, "Pragmatic and metaphoric gestures– combining functional with cognitive approaches in the analysis of the "brushing aside gesture"," in *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction (Handbook of Linguistics and Communication Science 38.2.)*, C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and J. Bressem, Eds., ed Berlin/ Boston: De Gruyter Mouton, 2014, pp. 1540-1558.

[9] C. Müller, "Wie Gesten bedeuten. Eine kognitiv-linguistische und sequenzanalytische Perspektive," *Sprache und Literatur,* vol. 41, pp. 37-68, 2010.

[10] J. Bavelas Beavin, N. Chovil, D. A. Lawrie, and A. Wade, "Interactive Gestures," *Discourse Processes,* vol. 15, pp. 469-489, 1992.

[11] D. McNeill, *Hand and Mind. What Gestures Reveal About Thought*. Chicago: University of Chicago Press, 1992.

[12] A. Kendon, "Gestures as illocutionary and discourse structure markers in Southern Italian conversation," *Journal of Pragmatics,* vol. 23, pp. 247-279, 1995.

[13] C. Wegener, *A Grammar of Savosavo*. Berlin; Boston: de Gruyter, 2012.

[14] J. Bressem, N. Stein, and C. Wegener, "Refusing, excluding, and negating manually: Recurrent gestures in Savosavo," in preparation.

[15] J. Bressem and C. Müller, "The family of Away gestures: Negation, refusal, and negative assessment," in *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction (Handbooks of Linguistics and Communication Science 38.2.)*, C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and J. Bressem, Eds., ed Berlin/Boston: De Gruyter Mouton, 2014, pp. 1592-1604.

[16] S. Harrison, "Grammar, Gesture, and Cognition: The Case of Negation in English," Ph.D PhD Thesis, Université Michel de Montaigne, Bourdeaux 3, 2009.

[17] G. Calbris, *Elements of Meaning in Gesture*. Amsterdam: John Benjamins Publishing Company, 2011.

[18] J. Bressem, S. H. Ladewig, and C. Müller, "Linguistic Annotation System for Gestures (LASG)," in *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction. (Handbooks of Linguistics and Communication Science 38.1.)*, C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and S. Teßendorf, Eds., ed Berlin/ Boston: De Gruyter Mouton, 2013, pp. 1098-1125.

[19] S. H. Ladewig, "Beschreiben, suchen und auffordern – Varianten einer rekurrenten Geste," *Sprache und Literatur,* vol. 41, pp. 89-111, 2010.

[20] S. H. Ladewig, "Recurrent gestures," in *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction (Handbooks of Linguistics and Communication Science 38.2.)*, C. Müller, E. Fricke, A. Cienki, S. H. Ladewig, D. McNeill, and J. Bressem, Eds., ed Berlin/ Boston: De Gruyter Mouton, 2014, pp. 1558-1574.

[21] J. Streeck, "Pragmatic aspects of gesture," in *International Encyclopedia of Languages and Linguistics*, J. Mey, Ed., ed Oxford: Elsevier, 2005, pp. 275-299.

The social nature of cognitive–semiotic processes in the semantic expansion of gestural forms

Heather Brookes

University of Cape Town

Abstract

This paper gives a social-semiotic account of five gesture families. It examines semantic expansion and conventionalization in the form-meaning relations of gesture families in the gestural repertoire of speakers of urban varieties of Zulu and South Sotho. Gestural forms vary in the extent to which they undergo semantic extension and conventionalization. Gestures that depict concrete objects have limited related semantic possibilities. Where conventionalization of gestures occurs, this process is motivated by the interaction of both visual cognitive image scheme and social communicative needs. In the case of imagistic schema, these can be expanded based on underlying metaphorical abstract semantic cores. However, these expansions are not only cognitively motivated. They are culturally shaped by norms and values underlying physical and social conventions as well as communicative expressive needs. The paper argues for a socio-semiotic framework for the cross-linguistic analysis of gesture families.

Index Terms: gesture family, semiotic, semantic expansion, metaphor, conventionalization, socio-cognitive

1. Introduction

The analysis and classification of gestures has been a frequent topic in the gesture literature [1]. Various differences have been emphasized. For example, from a functional perspective, pragmatic gestures that convey speech acts or mark discourse, are distinguished from representational gestures that express content [2]. Another important distinction has been made between co-speech gestures and gestures that can convey meaning independently of speech. This distinction is connected to the notion that some gestures are spontaneous, idiosyncratic and improvisatory while others are highly conventionalized [3].

As the number of studies of spontaneous gesture use has increased, these distinctions appear less clear-cut. For example, representational gestures can function pragmatically and pragmatic gestures can convey propositional content [4]. Although some co-speech gestures may appear to be improvisational, the gestural forms re-occur with similar or related meanings and functions and therefore must have underlying cultural conventions governing their use [5], [6]. Similarly, highly conventionalized gestures such as quotable gestures or emblems co-occur with speech and can function like co-speech gestures [7].

A semiotic approach to the analysis of gesture provides an alternative starting point for the analysis and classification of gestures. It takes a gesture's core kinesic feature(s) and examines how various components such as form, location, movement, and combination of body parts vary from one context of use to another and how these features express variations in meaning and function [8]. Accordingly, several scholars have proposed categorizing gestural forms into gesture families [8]. A gestural family consists of different iterations of a common core gestural form and meaning. The core form expresses related meanings based on its physical variation (i.e. location or movement) and spoken verbal context [5].

The semiotic nature of gestural forms and their meanings have largely been explained in terms of cognitive and embodied motivations of gestural production [9]. Speakers map abstract ideas onto the physical domain (shape, movement and location), and these metaphorical mappings are conceptual metaphors grounded in our physical experience of the world. For example, the open hand supine gesture is grounded in the fundamental physical actions of giving and receiving. It occurs in many cultural groups and can convey many different meanings and functions [1], [10].

But one gestural form may not have the same metaphorical meaning across cultures nor express the same number of functions and meanings [11]. We can see how cognitive metaphors are produced through the mapping of abstract concepts onto the physical through visual cognitive schema. These schema may sometimes be common across cultures because of similar embodied experiences of the physical world. If there are differences, the source of this variation may lie in the socio-cultural aspects underlying gestural production? Much of the semantic analysis of gesture describes the semiotic motivation of gesture as an internal process of the mind based in physical experience through the body. However, sense-making is not only an internal cognitive process, but a process that occurs in social interaction where both thought and socio-cultural values and behaviors impact each other.

In this paper, I analyse and compare five gesture forms and their families from the repertoire of gestures in use among urban Zulu and South Sotho speakers in Johannesburg. I examine: 1) how the referent is depicted in gestural form, 2) its analogical literal or metaphorical character, 3) the way it is used with speech, 4) whether it has related established gestural polysigns[8]/messages, and 5) the number of possible meanings each gesture family conveys. I demonstrate how socio-cultural metaphors contribute to the semantic productivity and conventionality of gestural forms. I argue that visual cognitive image schema are not only cognitively motivated but also shaped by social and cultural communicative needs. In the light of these findings, I explore 1) the usefulness of using the concept of gesture families to account for the full repertoire of gestures; 2) the conventional rather than the improvisatory nature of co-speech gesture; and 3) the relationship between recurrent gestures and emblems.

2. Gesture families

Urban Bantu language speakers, in and around greater Johannesburg, have a large repertoire of conventional gestural forms. These are used both independently and with speech. These gestures have been documented and some have been analysed based on elicited and filmed data in spontaneous contexts of use [12], [13], [7], [14], [15]. The analysis presented here is a first attempt to examine these gestures from a semiotic perspective using a gesture family framework. The gesture families selected represent different sizes of gesture families based on the extent of their semantic repertoire and degree of conventionalization.

2.1. The *sleep* gesture

We begin with the gesture for *sleep* in which the palm of the hand is held parallel towards one side of the head with the head and hand slightly inclined to the side. In some instances, only the head is tilted to the side and the hand is not used. The form is visually analogous to the position of the head resting on something when sleeping. In this sense, it is metonymic in that it represents one aspect of the act of sleeping (See Figure 1).

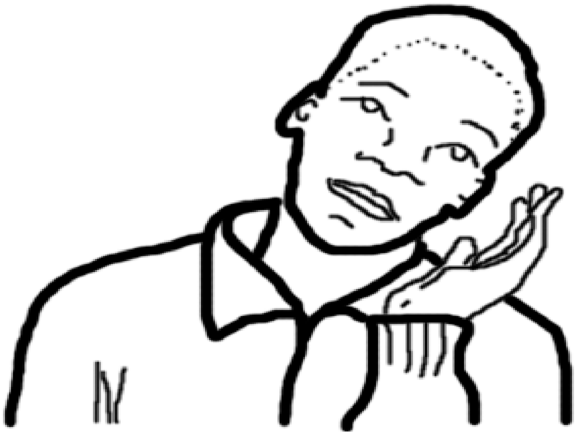


Figure 1: The sleep gesture.

The *sleep* gesture occurs in every day talk with its spoken equivalent or synonyms thereof. A commonly observed use is for someone to ask where a person is and for the interlocutor to answer in South Sotho, *O robetse* 'He's sleeping' with a tilt of the head to the side on the word sleeping. When used without speech, it either gives information that someone is sleeping or asks if someone is sleeping. Its performative function as a statement/comment or question can be deduced from context. For example, a person walks down the road with his friend and points to someone sleeping on the sidewalk and does the sleeping gesture.

The gesture has no variation in form other than the optional use of the hand. The use or non-use of the hand does not change the meaning, and there are no additional meanings that the gesture conveys. It makes literal reference to the action of sleeping and does not have any additional meanings or functions. The *sleep* gesture can be considered to be a gesture family with only one member.

2.2. The *money* gesture

If a gesture represents an object or action that plays a prominent social role in every day life, we often find that speakers use the gesture with related spoken concepts. An example in this community is the gesture for *money* in which upturned tips of thumb and first two fingers are held and sometimes rubbed together (see Figure 2). The gesture is visually analogous to holding or showing money.

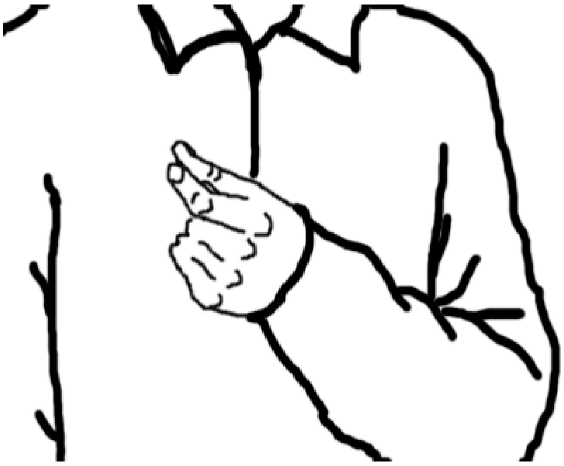


Figure 2: The money gesture.

Speakers use this gesture in similar ways to the *sleep* gesture. Independently of speech, it can convey a request, an offer or express a comment about a person's financial state, but this interpretation depends on context. With speech, speakers may use it while describing a person who is rich, to comment on how much money a person might have, to express that something costs a lot of money or they spent a lot of money, to ask how much a person has or how much they owe and to request money [see [7] for examples].

Unlike the *sleep* gesture, it can occur with many spoken synonyms and related concepts to do with money. It appears that the *money* gesture co-occurs with a wider range of meanings in conjunction with speech because of its significance in every day life. However, it there is no distinct variation in form that equates to a different meaning. For example repeatedly rubbing forefinger and thumb together does not necessarily mean 'very rich.' It could convey the intensity of a request for money. There is no physical distinction that makes an established difference in meaning.

2.3. The *talk* gesture

While gestures for objects and actions like *sleep* and *money* have a limited semantic range and set of communicative acts that depend on context for their interpretation, some gestural families have an established related gesture - similar in form, but with stabilized inflections usually in the movement of the stroke and/or the orientation or positioning of the hand - that is an established message. One example is the gesture for *talk*, in which thumb and extended abducted fingers make an opening and closing motion in front of the mouth. The gestural form is visually metonymic depicting the movement of the mouth. See Figure 3.

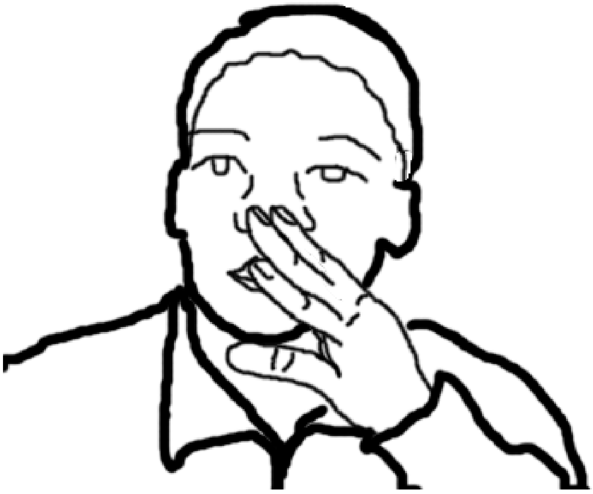


Figure 3: The talk gesture.

Similar to the money gesture, the *talk* gesture can occur with related spoken topics. It can also convey different messages without speech but these depend on the context such as 'Let's talk,' 'They're gossiping,' 'Talk quietly' and 'Talk louder.*'* In the latter two cases, the opening between the thumb and finger may, but not always, be smaller or wider. However, the gesture conveys an established message when speakers increase the amplitude between the fingers and thumb to the maximum as they open and close them, to express the established spoken gloss, *O na wede wede* 'You talk too much.'

Like the *money* gesture, it appears to accompany a range of spoken meanings all connected to the notion of 'talk' because of the significance of 'talk' and related activities in every day life. At the same time, a particular variation in form has become an established comment/insult. The 'talk too much' gesture can be understood in terms of Calbris' [8] concept of the polysign with two components, the movement of the thumb and fingers in front of the mouth and the widening of the movement that combine to form an established sign associated with a specific spoken phrase. Here we have a gesture family with at least two established related forms and perhaps two slightly less well-established variations in *talk quietly* and *talk louder*.

2.4. The *child* gesture

Another gesture that has a related established polysign is the gesture for *child* in which the fingers and thumb of an upturned hand touch at the tips in a *finger bunch* (see Figure 4).

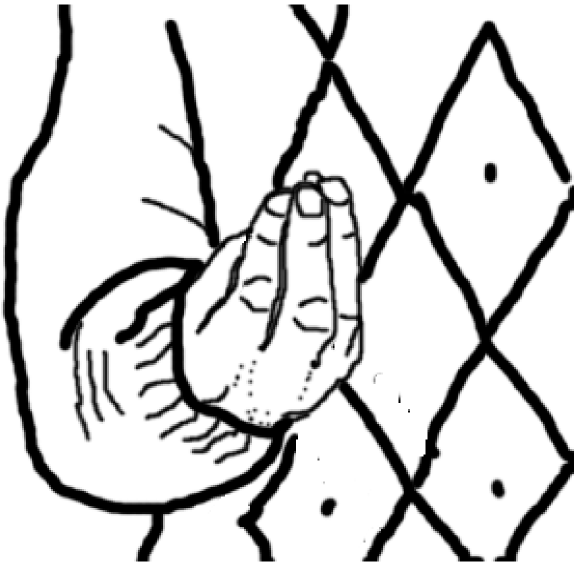


Figure 4: The child gesture.

Speakers usually gloss the form as *child* and commonly use it with speech to indicate a child's age by holding it out to the side to indicate the child's height from the ground without expressing this information in speech. Speakers also use it when talking about a sibling to indicate whether the brother is older or younger. The gesturer positions the hand (held out to the side) in relation to the self either below head height to indicate a younger brother or sister or above head height for an older sibling again without expressing this information in speech. However, it has a related quotable form. When held out in front of the speaker at stomach level and moved sideways back and forth it is an established and recognizable insult meaning *You're a small boy* in other words, you are as ineffectual or useless as a small boy.

The finger bunch no longer analogically depicts the referent directly. One can surmise that the finger bunch could be depicting something small, and therefore we can say the form is metonymic and abstract representing a key characteristic of a childhood. At the same time it is metaphorical in that childhood is being depicted in terms of size. Alternatively, it could be suggesting the 'essence' or 'core' of humanity, or out of childhood comes adulthood. This interpretation could be plausible especially in the light of the taboo on using a flat hand parallel to the ground to depict a person's height. A flat hand can only be used to show the size of an animal and it is taboo to use it to show a person's size.

It has a related established gestural form and meaning that involves three combined physical components, the bunched fingers upward, in front of the stomach, with lateral transverse movements. These components make up an established polysign involving two analogical links, a form shape sign for smallness and a movement sign. Transverse movements have been noted in the Open Hand Supine gesture in this context when two hands are held with palms up and moved laterally across one another to show something is lacking, there is nothing to hold or receive.

The same gestural form, prominent among Italian speakers, has a different set of meanings and functions based on how it is metaphorically understood in that speech community. While there may be some semantic similarity in the physical form depicting the 'extraction of the core or essence' or essential core equals small versus child equals small, its semantic application is quite different. Among Italians, its underlying form-meaning relation allows it to have multiple pragmatic functions in relation to speech [1]. Among Bantu language speakers, it represents a concrete object. The possibility of ‘personhood’ is there, but the gesture does not co-occur with many concepts related to that notion. Neither does the gesture occur with concepts related to ‘essence’ or ‘smallness.’ Its core form-meaning relation is with ‘child.’

2.5. The *clever* gesture

The last gesture family to be presented here is the gesture for *clever* ‘streetwise.’ Its core form involves pointing the extended index and fourth finger towards the eyes of the speaker (See Figure 5). Its core meaning relates to ‘seeing.’ Analogically it metonymically depicts ‘seeing’ by pointing to the eyes. However, the ‘seeing’ is metaphorical as Cienki [16] points out suggesting that with the *clever* gesture ‘seeing’ metaphorically equals ‘knowing.’ In this cultural context, the particular notion of ‘seeing’ is related to being open to the new, forward looking, progressive and urban. The *clever* gesture can be understood in contrast to the gesture for ‘a stupid’ in which the flat palm is drawn diagonally across the face to show ‘not seeing,’ sight being cut off or a person with a closed mentality. The common spoken word with this gesture is *bari* ‘a stupid/backward/rural person (country bumpkin)’. It comes from the old Afrikaans word *baar* meaning ‘raw native.’ The *clever* gesture is culturally metaphorical in that it connotes ‘seeing’ in terms of ‘knowing’ in the urban environment. It describes a person who is alert, streetwise, urban and progressive/modern encapsulated in the term *clever* that does not mean intelligent in the local spoken varieties but ‘streetsmart and city slick.’

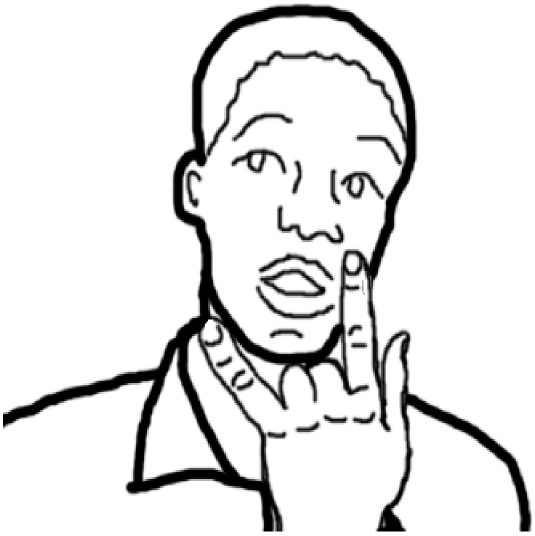


Figure 5: The clever gesture.

This gesture can be used in conjunction with speech with its spoken equivalent and synonyms thereof as well as with other related words and phrases that describe the characteristics of what constitutes ‘a clever’ or metaphorically a ‘seeing/knowing person’ such as being witty, entertaining, verbally skillful, sophisticated, urban and able to be ahead of everyone else as well as thwarting the system.

This gestural form also expresses a range of established meanings independently of speech that are all related to the concept of ‘seeing/perceiving.’ The basic form is combined with other physical components to make different established polysigns. When directed towards the eyes, the gesture expresses ‘look/see.’ If the eyes are open wide, then the gesture is a warning to “Watch out.” When there is a movement of the hand with first and fourth fingers extended diagonally up and down across the face, it expresses that someone is *clever* ‘streetwise and city slick.’ If this movement is combined with wide-open eyes and/or vigorous movements of large amplitude of the hand, then the person is extremely streetwise. However, if the gesture is done with minimal amplitude of the stroke and eyes are wide open, the gesture means the person is a crook. If the extended index and fourth finger are held towards the interlocutor or up in the air it expresses the meaning ‘I see you’ which is a common greeting. If the index and fourth finger are held close against the body in a particular direction, the meaning is a warning that someone in a certain location (opposite to where the two fingers are pointing – in other words the direction that the person is looking) is watching the person to who you are making the gesture.

In this case, we see that the although there is a literal use of the gesture as in ‘see,’ its metaphorical nature and the cultural meaning of the metaphor underlie the gesture’s polysemy allowing it to generate many different but related meanings that have become established polysigns and emblems.

3. Discussion

Gestural forms that have a limited semantic range, in other words, they express a single meaning, are often literal metonymic depictions of every day objects and actions. These gestures are visually analogous to their concrete referents. Where objects or actions play a greater social role, their gestural representations often occur with semantically related spoken words and phrases. Thus the gesture may express different but related meanings determined either by speech or context, but not from the form (or a well-formed physical distinction) of the gesture. Where a particular phrase or message such as a common state or an insult becomes socially established by frequent use and consequently contiguity of spoken phrase and distinct well-formed gestural components, an established gestural message or emblem results. Where the gestural form becomes metaphorical, it can generate a number of related or polysemous meanings. If the metaphor is grounded in socio-cultural and historical concerns, it can generate more meanings in which the components of the gesture become contiguous and established signs result. The extent to which the analogic components become contiguous and result in an established form-meaning relation depends on the extent to which they are needed and used among a group of speakers. What appears to extend and conventionalize a gesture family is the combination of both potential conceptual and sociocultural metaphor.

In previous work on gesture families (for example [1], [4], [10]), the semantic theme of the gesture family is abstract. Here I argue that there is a concrete literal meaning to the core form of all gesture families that then becomes abstracted through metaphorical processes. The extent to which a form generates variation and abstraction depends on the conceptual potential of the gestural form and social communicative needs. In other speech communities, a gesture for ‘look/see’ may not transform into the metaphor of knowing and the abstract notion of ‘streetwise knowledge’ for example. Speakers could have extended the *sleep* gesture to mean dull and boring.

Instead the *stupid* gesture is used with metaphorical phrases such as *bekalele* 'sleeping [dull and boring]' to talk about a person who lacks the communicative skills to be entertaining and therefore *a clever* 'streetwise.'

4. Conclusions

Some gesture forms have more iterations than others. A gesture family can consist of a single gestural expression/meaning. Gestural families vary in their semantic possibilities based on the analogic and metaphorical nature of their gestural forms and their social significance. Within a gesture family, some iterations are less well established than others. In other words, they do not conform to well-formedness [11]. Sometimes a particular iteration of a gesture comes to have an established meaning or expression based on communicative needs and frequency of use. Metaphoric processes provide the mechanism by which gestures have the possibility of expanding semantically but these expansions are shaped by sociocultural concerns that determine semantic productivity and emblematic establishment. While the idea that metaphor is a primarily a cognitive phenomenon and thought is grounded in embodied experience, socio-cultural notions and communicative requirements shape how visually embodied concepts are mapped onto the physical gestural domain.

Using the concept of 'gesture family' allows a coherent account of a speech community's gestural repertoire and also allows for more systematic and empirically grounded cross-linguistic comparisons. There appears to be continuum of both semantic and functional expansion and conventionalization within each gesture family so that the same core form can on some occasions be pragmatic and on others representational, on some occasions less context dependent and others quite well established. Some of these iterations may involve changes and combinations in the physical shape, location and movement of the core gestural form.

Finally, in analyzing the core form of a gesture as part of a gesture family, the term *recurrent* gesture has been introduced to describe the discovery that many co-speech gestures have features such as location and movement that demonstrate an underlying cognitive and cultural conventionality [4], [5], [6]. With the ability to capture co-speech gestures on video and build up a database of the in situ uses of particular gestural forms, we see that co-speech gesturing is less idiosyncratic and improvisatory that first thought. Recurrent co-speech gestures share similar functional and structural characteristics to emblems/quotable gestures. Perhaps an emblem can be considered as one step further along the continuum towards iconization within a gesture family based on social circumstances that involve either practical or abstract ideological concerns.

5. Acknowledgements

The writing and presentation of this paper is supported by the SARChI research chair of R. Mesthrie (No. 64805, Migration, Language and Social Change) and the University Research Committee of the University of Cape Town.

6. References

[1] Kendon, A., Gesture: Visual Action as Utterance, Cambridge, 2004.

[2] Colletta, J. M., Pellenq, C. and Guidetti, M., "Age related changes in co-speech gesture and narratives. Evidence from French children and adults", Speech Communication., 52:565–576, 2010.

[3] McNeill, D., Hand and Mind, Chicago, 1992.

[4] Ladewig, S. H., "The cyclic gesture", in C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill and J. Bressem [Eds], Body, Language. Communication. An International Handbook on Multimodality in Human Interaction. (Handbooks of Linguistics and Communication Science 38.2.), 1605-1618, De Gruyter Mouton, 2014.

[5] Bressem, J. and Muller, C., "A repertoire of German recurrent gestures with pragmatic functions", in C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill and J. Bressem [Eds], Body, Language. Communication. An International Handbook on Multimodality in Human Interaction. (Handbooks of Linguistics and Communication Science 38.2.), 1575-1591, De Gruyter Mouton, 2014.

[6] Ladewig, S. H., "Recurrent gestures", in C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill and J. Bressem [Eds], Body, Language. Communication. An International Handbook on Multimodality in Human Interaction. (Handbooks of Linguistics and Communication Science 38.2.), 1558-1574, De Gruyter Mouton, 2014.

[7] Brookes, H. J., "What Gestures Do: Some Communicative Functions of Quotable Gestures in Conversations among Black Urban South Africans", Journal of Pragmatics, 37:2044-2085, 2005.

[8] Calbris, G., Elements of Meaning in Gesture, John Benjamins, 2011.

[9] Cienki, A. and Müller C., Metaphor and Gesture, John Benjamins, 2008.

[10] Müller, C., "Forms and uses of the Palm Up Open Hand. A case of a gesture family?", in C. Müller and R. Posner [Eds], Semantics and Pragmatics of Everyday Gestures, 233-256, Weidler, 2004.

[11] McNeill, D., "The Emblem as Metaphor", in M. Seyfeddininpur and M. Gullberg [Eds], From Gesture in Conversation to Visible Action in Utterance, John Benjamins, 75-94, 2014.

[12] Brookes, H. J., "O clever 'He's streetwise.' When gestures become quotable: the case of the clever gesture", Gesture 1(2):167–184, 2001.

[13] Brookes, H. J., "A repertoire of South African quotable gestures." Journal of Linguistic Anthropology, 14(2):186–224, 2004.

[14] Brookes, H.J., "Amangama amathathu 'The three letters.' The emergence of a quotable gesture (emblem)", Gesture, 11(2):194–218, 2011.

[15] Brookes, H. J., 2014 in M. Seyfeddininpur and M. Gullberg [Eds], From Gesture in Conversation to Visible Action in Utterance, John Benjamins, 75-94, 2014.

[16] Cienki, A., "Why study metaphor and gesture?", in A. Cienki and C. Müller [Eds], Metaphor and Gesture, 5-25, John Benjamins, 2008.

The relation between global pitch range and gestures in a story-telling task

M. Grazia Busà [1], Sara Brugnerotto [1]

[1] DiSLL, University of Padova, Padova, Italy

Mariagrazia.Busa@unipd.it, Sara.Brugnerotto@studenti.unipd.it

Abstract

Anecdotal evidence suggests that both pitch range and gestures contribute to the perception of speakers' liveliness in speech. However, the relation between speakers' pitch range and gestures has received little attention. It is possible that variations in pitch range might be accompanied by variations in gestures, and vice versa. In second language speech, the relation between pitch range and gestures might also be affected by speakers' difficulty in speaking the L2. In this pilot study we compare global pitch range and gesture rate in the speech of 3 native Italian speakers, telling the same story once in Italian and twice in English as part of an in-class oral presentation task. The hypothesis tested is that contextual factors, such as speakers' nervousness with the task, cause speakers to use narrow pitch range and limited gestures; a greater ease with the task, due to its repetition, cause speakers to use a wider pitch range and more gestures. This experimental hypothesis is partially confirmed by the results of this study.

Index Terms: pitch range variation, gesture rate, story telling, English L2, Italian L1

1. Introduction

One of the goals of public speaking classes is to teach students to use a 'lively' voice when delivering a speech. This means that students should speak with a voice that varies in intonation, rhythm and volume. This is because by varying intonation, rhythm and volume speakers can emphasize important points of their discourse and deemphasize others, and thus help listeners follow the information flow. In other words, variation in speech helps listeners maintain their focus on the speaker's message and not wander away [1, 2].

In addition to voice, public speaking classes emphasize the importance of body language in discourse: students are told to maintain an open body position and to use gaze and gestures to highlight parts of speech. This contributes to maintaining the listeners' attention by providing them with a visual channel, in addition to the audio channel, that helps them follow the information flow.

For second-language learners, speaking in public involves planning thoughts, discourse structure and words, together with intonation and gestures, in a language that is not their own. This results in a very heavy cognitive load that may impair one or all levels of output: linguistic, prosodic, and gestural. As a result, second-language learners' delivery of speeches in public may appear incongruent or tedious, with an effect on the successful outcome of their presentations. However, in L2 as in L1, performance can be improved through preparation and rehearsal, which can contribute to reducing the contextual factors, such as nervousness, that affect speakers' congruence and delivery.

The worldwide success of public speaking classes shows that students can –in fact– learn to modify their voice and body language habits in discourse, and give oral presentations that are effective in holding the audience' attention.

However, though the dynamics of successful speaking attract the interest of many, there is a lack of scientific research focusing on the quantitative measurements of performance.

This paper reports on a preliminary study aimed at investigating how contextual effects, such as nervousness for a speech delivery, may affect speakers' use of pitch range and gestures. This is done by presenting an investigation of the global pitch range and gestural characteristics of 3 Italian speakers of English engaged in a story-telling task in Italian and English.

2. Pitch range, gestures and common ground

It is known that in most languages meaning and emphasis are created by means of variations of the fundamental frequency (or F_0) of the human voice. The range over which these variations may occur is called pitch (or F_0) range. Typically, a voice that is heavily inflected, that is, has a wide pitch range, will sound animated; a voice that has a narrow pitch range will sound monotone. Thus, pitch range has been used as a measure of speaker's perceived liveliness [1, 2, 3] –though the use and interpretation of pitch range may vary depending on language [3, 4, 5] and sociocultural/ sociophonetic factors [6].

It has been suggested that L2 speech may be characterised by limited pitch variation and a narrower pitch range than L1 speech [1, 2, 3, 4, 8, 9, 10, 11]. It is possible, in fact, that prosodic information is processed differently by native and non-native speakers because of their different levels of competence in the L1/L2. For example, as suggested by [7], non-native speakers may rely more on segmental, as opposed to prosodic, information to get their meanings across, given the fact that they lack the amount of extra-linguistic knowledge that native speakers can rely on when communicating. Differences in pitch range in L1 and L2 may also be more conspicuous in particular speaking styles, such as formal presentations [1, 2, 12], during which non-native speakers may be particularly focussed on getting their meanings across, at the expense of prosody.

A framework for measuring global pitch range cross-linguistically was first established by Ladd [13], then elaborated by Patterson [14], and finally by Mennen et al. [3; 4]. Within this framework, a number of measures are used to quantify differences in pitch level (i.e., the speaker's overall pitch height or register) and pitch span (i.e., the speaker's range of frequencies in a speech sample). These include F_0 max, min, mean and median, as well as linguistic measures, linked to specific linguistically-defined landmarks in the F0 contour.

A different measure of pitch range was used by Hincks [1, 2] to compare speakers' liveliness over long stretches of speech. Hincks looked at the normalized standard deviation of F_0, and found that a value of pitch variation, which she called pitch variation quotient (PVQ), strongly correlates with perceived speakers' liveliness, though only weakly with speakers' proficiency level. Pitch variation appeared to be a stronger perceptual cue to liveliness in male speech than in female speech. She concluded that pitch variation may not be the *only* measure of speakers' liveliness (rhythm and intensity being also measures of liveliness), but it is certainly an important one.

Research has shown that speech and gestures are interconnected [e.g., 15, 16]. According to McNeil [17, 18], speech and gestures are synchronous at the semantic level, as they are co-expressive of the same underlying meaning; at the pragmatic level, as they co-occur to express the same pragmatic function; and at the phonological level, as gestures are temporally coordinated with the phonology of the utterances.

A number of studies have examined the relationship of prosody and gestures, focussing in particular on the investigation of the temporal alignment of gestures with prosodic prominence [e.g., 19, 20, 21, 22, 23]. Evidence has been found that gestures are coordinated with prosodic stress, but there is little consensus as to how exactly gestures are aligned with prominent parts of speech [e.g., 24, 25, 26, 27, 28, 29]. Beat gestures might have a stronger influence on speech production than representational gestures [30]. It is possible that some gestures have an effect on the perception of speech prominence. For example, the realization of a visual beat in association with a prosodically prominent word has an effect on the acoustic realization of the word, and causes that word to be perceived as more prominent than the neighboring words [30].

While research has focussed on the synchronization of gestures with prosodic prominence, the relationship between speakers' global pitch range and gestures has received little attention. Anecdotal evidence suggests that there might be a relation between the amount of pitch variation in speakers' speech and the extent to which speakers gesture when they speak. In fact, it is highly likely that speakers convey paralinguistic meanings through their voices as well as through their gestures.

Co-speech gestures seem to fulfill a number of functions, and may in fact be multifunctional [reviewed in 31, 32, 33]. Gestures have been shown to facilitate speakers' cognitive processes during speech production; for example, they seem to help speakers conceptualize, retrieve lexical items, manage cognitive loads, organize information into syntactic constituents. Gestures also seem to be planned and produced with the addressee's needs in mind, and so play a role in communication. For example, speakers produce more and larger gestures when they see their interlocutor(s), than when they do not (e.g., when they are talking over the phone) [34]. Speakers' gestures are also affected by common ground, that is the amount of knowledge that is shared between the participants in a spoken interaction. It has been shown that assuming common ground causes speakers to use less words in their narratives than when no common ground can be assumed (because in the first case speakers can rely on their interlocutors to understand implicit references); on the other hand, common ground produces an increase in the use and extent of gestures during speech, possibly to enhance communication with the interlocutors [31, 32, 33]. Finally, gestures may be constrained also by contextual factors, accounting for individual differences, speakers' emotional involvement, etc. These, however, are still largely unexplored.

In L2 communication, L1 gestures appear to have an effect on L2 gestures at all stages of language development. In fact, L2 acquisition is characterized by processes of transfer and interference of gestures from the L1 to the L1 that should be studied, together with verbal language, as part of the interlanguage [35, 36, 37, 38, 39, 40, 41].

Some studies suggest that bilingual speakers might gesture more than monolingual speakers because gesturing helps them formulate their spoken message and is a way to compensate for the reduced proficiency in their L2 [42]. In addition, speakers with low levels of competence might use more L1-specific gestures than speakers with higher levels of competence [40]. L2 speakers' greater use of gestures than L1 speakers might be explained on cognitive grounds, that is, due to the cognitive complexity that speaking a foreign language requires [43].

However, studies do not support unambiguously the idea that bilinguals use more gestures than monolinguals. Other factors besides reduced proficiency in the L2 may account for the differences between the use of gestures in L1 and L2. Communication and contextual factors might affect gesture use in L2 speakers as they do in L1 speakers. For example, common ground might have an effect on L2 speakers' gestures and lead to increased gesturing that is unrelated to L2 speakers' proficiency level [31, 32, 33]. Contextual factors such as task expressiveness, nervousness, as well individual factors might also affect L2 speakers' gestures. Nicoladis et al. [44] examined the relationship between gesture use, L2 proficiency level and task complexity in a story recall task. They found only weak evidence supporting the idea that increased task complexity leads to increased gesture use, and suggest that gesture use might also be related to expressivity, as well to the speaker's gender.

What happens when L2 speakers speak in front of an audience? A number of factors may determine how L2 speakers' use their voice and gestures in a public presentation. Public speaking training classes insist that speakers can improve their non-verbal communication skills by learning the basics and rehearsing before they give their speech in public. It is assumed that rehearsal may help the speaker lessen the tension, sound and look less stiff, more natural during the presentation, and be more pleasant for the listener to hear. For L2 speakers, reducing the tension may significantly impact on the verbal and non-verbal production in L2, and bring about an improvement in both.

There is little scientific research to support the beliefs and assumptions of public-speaking training classes. To fill this gap, this paper reports a preliminary study of students' non-verbal behavior in a presentation in front of a class. The study is part of an investigation aimed at understanding speakers' use of voice and body language in public speaking as well as how non-verbal communication can be enhanced though formal instruction. The study examines the pitch range and gestural characteristics of 3 Italian speakers of English engaged in a story-telling task in Italian and English. The hypothesis tested is that contextual factors such as nervousness or performance anxiety will cause speakers to use narrow pitch range and reduced gesturing; greater ease with the task (because of rehearsal and/or greater familiarity with the task) will cause speakers to use wider pitch range and more gesturing.

3. Experiment

To test the experimental hypothesis, this study compares the pitch variation quotient (PVQ) [2] and the overall number of gestures of three Italian speakers telling the same story, once in Italian and twice in English, as part of an in-class oral presentation task.

3.1. Subjects, Method and Materials

The subjects were part of a larger group of (10) subjects who took part in the experiment. They were all English L2 learners, participating in a public-speaking class, master-degree level, taught by the first author. All subjects were female, mean age 22.75, speakers of Italian L1 and students at the University of Padova, with a competence of English at the B1 level of the CEFR. The data of the remaining 7 subjects are under analysis.

The speakers had to tell the class a fable, Aesop's "The Fox and The Crow", that they had previously read at home. The speakers told the story a first time in Italian, and right afterwards in English. They then repeated the story in English a second time a week later. Thus, the first time the speakers told the fable in Italian and English they had little time to prepare for the task; the second time they had much more time to prepare the story at home before repeating it in class. The speakers were video-recorded by the teacher. Each recording lasted about 90-120 seconds.

The three data sets will be referred to as Italian (=Italian L1); English 1 (=English, repetition at time 1) and English 2 (=English, repetition at time 2).

Out of the whole material, the authors selected 10 utterances that were used by all the subjects telling the fable. In these utterances the concepts expressed were the same, though the words and type of sentences used by the speakers were different. The purpose of selecting only the utterances that were used by all speakers was to compare, for any given utterance, the possible co-occurrence of one or more gesture. The selected utterances are reported in Table 1.

N.	Utterance
1	Once upon a time
2	It was flying around
3	On the shelf of a window
4	It flew down
5	It picked up the cheese
6	It went to the top of the tree
7	The crow opened its beak
8	The cheese fell to the ground
9	The fox caught it
10	It ran away

Table 1: List of utterances selected for the analysis.

3.2. Data Analysis

The audio signal was extracted from the videos using the AVC software (available at http://www.any-video-converter.com/). The audio signal was imported in Praat (www.praat.org), and pitch was measured setting the pitch floor to 75 Hz, and the ceiling to 500 Hz (since all the speakers were female). The boundaries of the selected utterances in the audio files were marked on a text grid. To calculate the PVQ, following a procedure indicated in [2], the pitch listings were extracted from each audio file, the outliers were removed, mean and standard deviation were calculated, and the data were normalized dividing the standard deviation of F_0 by the mean. This procedure was carried out on both the whole audio files and the selected utterances. The statistical significance of the results was tested with one-way ANOVAs with task as a factor, and post-hoc Tukey HSD tests.

The audio signal was then imported in Elan (https://tla.mpi.nl/tools/tla-tools/elan/). An analysis was carried out to annotate each gesture co-occurring with the selected utterances in the three data sets (Italian, English 1 and English 2). At this preliminary stage of analysis, the aim was only to get a total count of the gestures, per speaker and data set, so as to verify if there exists any relation between the variation in the speakers' PVQ and their overall gestures. Because of this, for this analysis, we grouped together all iconic and non-iconic gestures. An analysis of the speakers' gestures classified by type will be carried out in the next phase of the study.

Gesture rate was calculated for each data set following a procedure used in Nicoladis et al. [44]. Gesture rate is a measure of the percentage of word tokens accompanied by gestures, and is calculated by dividing the number of gestures by the total number of words multiplied by a hundred. The use of this measure controls for individual differences in speech.

To calculate the gesture rate for this analysis we counted all the words used in the selected utterances for each speaker. Speakers' disfluencies, repetitions and corrections were computed as part of the total number of words. However, they were also counted separately, as they may reflect grammatical or lexical difficulties that speakers may tend to compensate with their gestures.

4. Results

4.1. Pitch Variation

Tables 2 and 3 show the PVQ data for the three speakers, as calculated, respectively, for the whole story and the selected utterances.

Table 2 shows that all speakers vary their pitch more in the English 2 task than in English 1 or Italian. Interestingly, for all speakers the PVQ of Italian is comparable to the PVQ of English 1, showing that, at time 1, the speakers did not use a very varied pitch in English or Italian. This difference is greater for speaker C than for A or B.

At the ANOVA test, the difference in pitch values in the three tasks was highly significant for all speakers: for speaker A: F(2, 21421) = 337.06, p <.0001 –though the difference between PVQ in Italian and English 1 was not significant at a Tukey HSD test; for speaker B: F(2, 17022) = 936.12, p <.0001; for speaker C: F(2, 24426) = 1724.9, p <.0001.

PVQ - story	Italian	English 1	English 2
Speaker A	0.17	0.18	0.21
Speaker B	0.22	0.23	0.24
Speaker C	0.20	0.20	0.26

Table 2: Pitch variation quotient for the three speakers in the entire story in Italian, English 1 (repetition at time 1) and English 2 (repetition at time 2).

PVQ - utterances	Italian	English 1	English 2
Speaker A	0.18	0.20	0.18
Speaker B	0.22	0.22	0.24
Speaker C	0.23	0.19	0.25

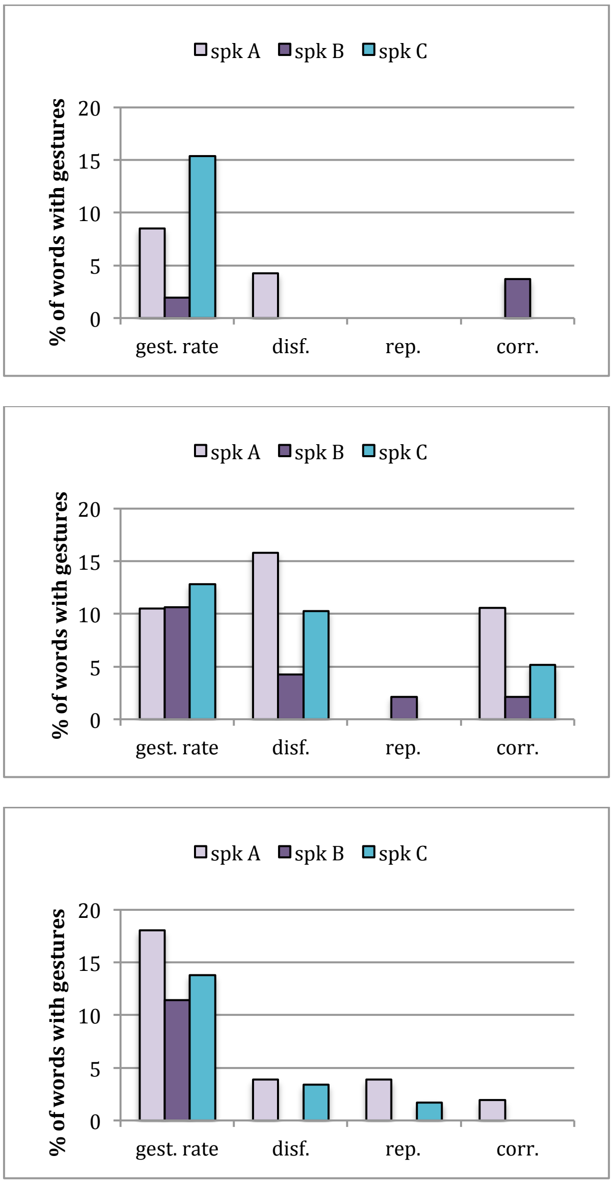
Table 3: Pitch variation quotient for the three speakers in the selected utterances in Italian, English 1 (repetition at time 1) and English 2 (repetition at time 2).

Table 3 shows the PVQ data for the utterances only. Speaker A appears to vary her mean pitch more in English 1 than in the other two data sets, but the difference in PVQ in the three data sets is not significant at the ANOVA test. Speaker B varies her mean pitch more in English 2 than in Italian and English 1 [F(2,22) = 11.73, p = 0.000341], with a difference between Italian and English 1 that was not significant at the post-hoc Tukey test. Speaker C has higher mean pitch values in Italian and English 2 than in English 1, but the difference between the three data sets is not significant at the ANOVA test.

4.2. Gesture rate

Figure 1-3 show the gesture rate and percentages of disfluencies, repetitions and corrections for the three speakers in Italian, English 1 and English 2, respectively.

The data show that for two speakers gesture rate increases from Italian to English 1 to English 2; for the third speaker gesture rate is highest in Italian, and then slightly higher in English 2 than in English 1. Disfluencies and corrections are most frequent in English 1, but they occur, for two of the speakers, also in English 2; two speakers show some disfluencies and corrections also in Italian.



Figures 1-3. Gesture Rate, Disfluencies, Repetitions and Corrections in Italian (top), English 1 (center), and English 2 (bottom).

To test the correlation of the present data with the data on the pitch variation we ran Spearman correlation tests, but they did not yield positive correlations, probably because of the limited data provided. However, the data show some trends. Overall, speaker C and A gesture more than speaker B. Speaker C has the highest gesture rate and PVQ in Italian; her gesture rate decreases in English 1 to rise slightly in English 2; her PVQ also decreases in English 1 to rise considerably in English 2. This speaker also has the highest percentage of disfluencies and corrections in the data sets. Speakers A and B show a considerable increase in gesture rate from Italian to English 2. For speaker A, this increase in gesturing cannot be clearly linked to her (non significant) variations in PVQ in the three tasks; however, this speaker shows a high percentage of difluencies, especially in English 1, which might be related to the increase in gesture rate and requires further investigation. Speaker B has the lowest gesture rate in Italian; this rate increases in English 1 and English 2; in English 2 she has shows an increase in PVQ.

5. Discussion and Conclusion

This study is a preliminary investigation of the relationship between speakers' global pitch range and gestures, based on the assumption that their combined effect might contribute to the perception of speakers' liveliness in speech. The data from this study allow us to draw only tentative conclusions, which await confirmation in future studies.

Global pitch range and gesture rate were compared in the speech of 3 native Italian speakers. The speakers told the same story in Italian and in English and then, a week later, in English again. The presentations were part of the students' activities in a public-speaking class.

The analysis shows that when the speakers repeated the story in English the second time their pitch was more varied than when they told the story in Italian and/or English the first time. This is interesting since speakers are expected to show a wider variation in pitch in their native language and not in the L2 –as reviewed in § 1, L2 speech tends to be characterised by limited pitch variation and a narrower pitch range than L1 speech. It is possible that the speakers used a wider pitch range in the second repetition in English due to stylistic and contextual factors. That is, they had more time to prepare, put a greater effort in performing well, had less tension in accomplishing the task, etc. It can be hypothesized that knowing the task, being able to prepare and rehearse for it creates the conditions for sounding more lively in speech. However, we realize that to really evaluate the impact of rehearsal on global pitch range, the experimental design needs to include also a second repetition of the story in Italian. This would allow us to compare the students' performances in the second repetition in Italian and English, and see how pitch range changes with respect to the first repetition in both languages. This will be done in future work.

The gesture data show, as expected, individual differences in the use of gestures. The three speakers show quite different gesture rates in Italian. Also, for speakers A and B gesture rate is lowest in Italian, increases in the first repetition in English, and is highest in the second repetition. For speaker C gesture rate is highest in Italian, it is lowest in the first repetition in English, and rises again in the second repetition in English. Speakers A and B's increased gesture rate in the first repetition in English can be explained on both cognitive and communicative grounds [31, 32, 33]. The speakers may gesture more in English than in Italian because gestures help them tell the story in English L2, which is a complex cognitive activity. At the same time, the speakers may gesture more in English than in Italian because they are adapting their gestures to addressees with whom they share common ground: the speakers are telling the story in front of the class, and the class has heard the story before. Speaker C's lowest gesture rate for the first repetition in English cannot be attributed simply to cognitive or communicative factors –which would both lead to increased gesturing. Contextual or individual factors, such as the speaker's tension for the task, might have affected her gestures.

Finally, the data show that, in general, speakers' wider pitch co-occur with higher gesture rate, providing preliminary support to our hypothesis.

This study has some obvious limitations, which will be corrected in its continuation. One relevant aspect that this study does not tackle concerns the nature of the gestures produced by the speakers. Future work might show that, for example, L2 speakers produce more deictic gestures in L2 than in L1, as has been shown in much previous research [e.g. 45]. The use of iconic gestures in this task is also worth investigating. Classifying the types of gestures produced by the speakers is indeed important for drawing conclusions in this type of study.

The investigation will be expanded with the addition of more subjects as well as the analysis of other acoustic parameters that might contribute to the perception of speakers' liveliness. Also, the subjects will be tested a second time also in Italian to obtain data that are comparable with second repetition in English.

In spite of its limitations, we believe that this study shows that investigating the relation between global pitch range and gestures in first and second language speech is worth pursuing.

6. References

[1] Hincks, R., "Processing the prosody of oral presentations", Proc. InSTIL/ICALL2004 – NLP and speech technologies in advanced language learning systems, 63-69, 2004.

[2] Hincks, R., "Measuring liveliness in presentation speech", Proc. INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, 01/2005.

[3] Mennen, I., Schaeffler, F. and Docherty, G., A methodological study into the linguistic dimensions of pitch range differences between German and English. Proc. IV Conference on Speech Prosody, University of Campinas, 527-530, 2008.

[4] Mennen, I., Schaeffler, F. and Docherty, G., "Cross-language differences in fundamental frequency range: A comparison of English and German, Journ. of the Acoustical Society of America, 131(3): 2249-2260, 2012.

[5] Graham, C., "Revisiting f0 range production in Japanese-English simultaneous bilinguals", UC Berkeley Phonology Lab Annual Report, 110-125, 2013.

[6] van Bezooijen, R., "Sociocultural aspects of pitch differences between Japanese and Dutch women", Language and Speech, 38(3): 253-256, 1995.

[7] Jenkins, J., "A sociolinguistically-based, empirically-researched pronunciation syllabus for English as an International Language", Applied Linguistics, 23(1): 83-103, 2002.

[8] Aoyama, K. and Guion, S. G., "Prosody in second language acquisition: An acoustic analysis on duration and F_0 range", in O.-S. Bohn and M. J. Munro [Eds], The Role of Language Experience in Second-Language Speech Learning. In Honor of James Emil Flege, 281-297 Amsterdam, John Benjamins, 2007.

[9] Pickering, L., "The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse", English for Specific Purposes, 23: 19-43, 2004.

[10] Traunmüller, H. and Eriksson, A. "The perceptual evaluation of F_0 excursions in speech as evidenced in liveliness estimations", Journ. of the Acoustical Society of America, 97: 1905-1915, 1995.

[11] Ullakonoja, R., "Comparison of pitch range in Finnish (L1) and Russian (L2)", Proc. 16th ICPhS, 1701-1704, 2007.

[12] Johns-Lewis, C., "Prosodic differentiation of discourse modes", in C. Johns-Lewis, [Ed.], Intonation in discourse, 199-220, Breckenham, Kent: Croom Helm, 1986.

[13] Ladd, D. R., Intonational phonology, Cambridge: Cambridge University Press, 1996.

[14] Patterson, D., A linguistic approach to pitch range modeling, PhD dissertation, Univ. of Edinburgh, 2000.

[15] Goldin-Meadow, S., Hearing gesture: How our hands help us think, Cambridge, MA: Harvard University Press, 2003.

[16] Kendon, A., Gesture: visible action as utterance, Cambridge: Cambridge University Press, 2004.

[17] McNeill, D., Hand and mind: What gestures reveal about thought, Chicago: Chicago University Press, 1992.

[18] McNeill, D., Gesture and thought, Chicago: University of Chicago Press, 2005.

[19] Birdwhistell, R. L., Kinesics and context: Essays on body motion communication, Philadelphia: University of Pennsylvania Press, 1970.

[20] Kendon, A., "Gesticulation and speech: Two aspects of the process of utterance", in M. R. Key [Ed.], The Relationship of Verbal and Nonverbal Communication, The Hague: Mouton, 207-227, 1980.

[21] Bull, P. and Connelly, G., "Body movement and emphasis in speech", Journal of Nonverbal Behavior, 9: 169-187, 1985.

[22] Loehr, D. P. Gesture and intonation, Doctoral dissertation, Georgetown University, Dissertation Abstracts International, 65 06, 2180, UMI No. 3137056, 2004.

[23] Esteve-Giberta, N. and Prieto, P., "Prosodic structure shapes the temporal realization of intonation and manual gesture movements", Journal of Speech, Language, and Hearing Research, 56: 850–864, 2013.

[24] Rochet-Capellan, A., Laboissierre, R., Galvan A. and Schwartz, J. "The speech focus position effect on jaw-finger coordination in a pointing task", Journal of Speech, Language, and Hearing Research, 51: 1507-1521, 2008.

[25] de Ruiter, J. P. "The production of gesture and speech", in D. McNeill [Ed.], Language and Gesture, 284-311, Cambridge University Press, Cambridge, 2000.

[26] McClave, E., "Pitch and manual gestures", Journal of Psycholinguistic Research, 27: 69-89, 1998.

[27] Rusiewicz, H. L., "Synchronization of prosodic stress and gesture: A dynamic systems perspective", Proceedings of the 2nd Conference on Gesture and Speech in Interaction (GESPIN 2011), Bielefeld, Germany, 2011.

[28] Roustan, B., and Dohen, M., "Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus", Proceedings of Speech Prosody 2010, 100110, 1-4, 2010. Retrieved from www. speechprosody2010.illinois.edu/papers/100110.pdf, 2010.

[29] Leonard, T., and Cummins, F., "Temporal alignment of gesture and speech", in Proceedings of the Gesture and

Speech in Interaction (GESPIN 2009) Conference, Poznan, Poland, 2009.

[30] Krahmer, E. and Swerts, M., "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception", Journal of Memory and Language, 57, 396-414, 2007.

[31] Holler, J. and Wilkin, K., "Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task", Language and Cognitive Processes, 24(2): 267-289, 2009.

[32] Holler, J., Tutton, M. and Wilkin, K., "Co-speech gestures in the process of meaning coordination", in Proceedings of the 2nd Conference on Gesture and Speech in Interaction (GESPIN 2011), Bielefeld, Germany, 2011.

[33] Galati, A. and Brennan, S., "Speakers adapt gestures to addressees' knowledge: Implications for models of co-speech gesture", Language, Cognition and Neuroscience, 29(4): 435-451, 2014.

[34] Mol, L., Krahmer, E., Maes, A. and Swerts, M., "Seeing and being seen: The effects on gesture production", Journal of Computer-Mediated Communication, 17(1): 77-100, 2011.

[35] Gullberg, M., "Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon)", International Review of Applied Linguistics, 44(2): 103-124, 2006.

[36] Pika, S., Nicoladis, E. and Marentette, P. F. "A cross-cultural study on the use of gestures: Evidence for cross-linguistic transfer?", Bilingualism: Language and Cognition, 9: 319-327, 2006.

[37] Brown, A., "Gesture viewpoint in Japanese and English: Cross-linguistic interactions between two languages in one speaker", Gesture, 8(2): 256-276, 2008.

[38] Brown, A. and Gullberg, M., "Bidirectional crosslinguistic influence in L1-L2 encoding of manner in speech and gesture", Studies in Second Language Acquisition, 30(2): 225-251, 2008.

[39] Ortega, L., Understanding second language acquisition, London: Hodder Education, 2009.

[40] Nicoladis, E., "The effect of bilingualism on the use of manual gestures", Applied Psycholinguistics, 28: 441-454, 2007.

[41] Cavicchio, F., and Kita, S., "Bilinguals switch gesture production parameters when they switch languages", Proceedings Tilburg Gesture Research Meeting (TIGeR) 2013, 2013. Retrieved from: http://tiger.uvt.nl/list-of-accepted-papers.html.

[42] Nicoladis, E., Pika, S. and Marentette, P., "Do French-English bilingual children gesture more than monolingual children?", Journal of Psycholinguistic Research, 38 (6): 573-585, 2009.

[43] Kita, S., "How representational gestures help speaking", in D. McNeill [Ed.], Language and gesture, Cambridge: Cambridge University Press, 162-185, 2000.

[44] Nicoladis, E., Pika, S., Yin, H. and Marentette, P., "Gesture use in story recall by Chinese–English bilinguals", Applied Psycholinguistics 28(4): 721-735, 2007.

[45] Sherman, J. and Nicoladis, E., "Gestures by advanced Spanish-English second-language learners", Gesture, 4: 143-156, 2004.

Hand gestures and speech impairments in spoken and sung modalities in people with Alzheimer's disease

Diane Caussade [1,2,3], *Fanny Gaubert* [4], *Maud Sérieux* [4], *Nathalie Henrich-Bernardoni* [1,2], *Jean-Marc Colletta* [3], *Nathalie Vallée* [1,2]

[1] Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
[2] CNRS, GIPSA-Lab, F-38000 Grenoble, France
[3] LIDILEM, Université Grenoble 3, Saint Martin d'Hères, France
[4] Centre de formation en orthophonie, ISTR, Université Claude Bernard Lyon 1, Lyon, France
diane.caussade@gipsa-lab.fr

Abstract

In Alzheimer's disease (AD), studies on language production do not treat aspects of speech and hand gestures in a concomitant way. However, many studies describe either apraxia of speech, or orofacial apraxia, or upper limb apraxia, or aphasia. This paper reports an original protocol exploiting speech, singing and hand gestures to evaluate the correlation between upper limb and speech apraxia in spoken and sung modalities in 4 AD patients paired with 4 control participants. We did not evidence any speech apraxia in our AD patient population, unlike upper limb apraxia. However significant differences were observed on productions of hand gestures and speech between the patients and the control participants. Regarding patients, the movement, configuration and orientation of hand gestures were slightly altered. The hand gestures alteration seemed to depend on their value but not on the spoken *vs*. sung modality. The simultaneous repetition of connected hand gestures affected also both vocal and speech productions. More specifically, hand gestures seemed to impact the production of speech. The modality (spoken *vs*. sung) also seemed to influence speech productions at different degrees: patients made more errors in singing, and the more with connected hand gestures showing a double task effect.

Index Terms: Alzheimer's disease, gesture, speech apraxia, upper limb apraxia, voice quality, singing, speech

1. Introduction

According to the World Health Organization (WHO), Alzheimer's disease (AD) is the most frequent cause of neurocognitive disorder [21]. This neurodegenerative disease includes symptoms, such as amnesia, agnosia, attention disorders, apraxia, aphasia and dysphonia [17, 21, 22, 23, 26], which impact communication. This paper focuses on aphasia and apraxia in AD. Aphasia consists in the impairment of perception and production of language. It has been the main focus of most studies on communication disorders in AD as aphasia is easier to spot than apraxia [1, 26]. Apraxia is an impairment in the ability to program motor execution, like articulatory or upper limb movements [17, 22]. Speech and orofacial apraxia are part of articulatory movements impairment. Speech apraxia is a programming disorder of articulatory gestures used to produce phonemes [17]. While orofacial apraxia is a type of ideomotor apraxia in which the impairment concerns voluntary non-verbal movements of the face, lips and tongue [18]. Both speech and orofacial apraxia are often described in the semiology of AD [17]. Yet the study of apraxia has often been neglected [17, 22]. As for upper limb apraxia, it is defined as an impairment of non-verbal movements of the upper limbs, and notably hand gestures [22]. Thus, studies on bimodal language production of people with AD (such as [2], [21] and [23]) underlined speech and upper limb apraxia, but not in a concomitant way. However, the multimodal nature of communication has been widely reported (e.g. [6], [8] and [25]). An argument in favor of ontogenetic links between hand gestures and speech is the fact that around 12 months old babies begin to use pointing, which announces the emerging of first words, and then of syntax [6]. Later, between 2 and 5 years, children produce iconic gestures together with speech [13]. Recent studies [16] show ontogenetic links between music and language, which could explain the impact of music, and notably singing, on people with AD, in particular on attention, communication and motor disorders ([5], [10]). In this context, comparing communicative productions in spoken and sung modalities could help better understand the underlying effects of music on communicative productions of people with AD.

The aim of this study is to investigate communication impairments of persons with AD. Here, supported by the results obtained from a previous case study ([3], [4]), we assumed that the communication impairments would include a concomitant upper limb and speech apraxia, and a deterioration of hand gestures quality and of speech to a degree depending on the modality. We also hypothesized that deictic gestures would be better preserved than iconic ones. As deictic gestures develop first in speech ontogeny ([8], [9]), they could be better anchored than iconic ones. In view of these elements, the developed protocol is presented below.

2. Methodology

2.1. Experimental design

An original experimental protocol, approved by Grenoble CERNI ethic committee (*Comité d'Ethique pour les Recherches Non Interventionnelles*, 24/09/2013), was designed to study aphasia, speech and upper limb apraxia in a repetition task. This protocol was first tested and improved through a pilot study ([3], [4]).

Participants were asked to repeat 8 nursery rhymes composed of 6 sentences of 8 syllables each. Nursery rhymes were divided into spoken and sung modalities equally. In each kind of modality, two nursery rhymes were completed with four iconic and two deictic gestures each. The experimental protocol was completed with several clinical tests in order to evaluate speech and orofacial apraxia (which may impact speech), and upper limb apraxia (which may impact upper limb gestures, such as hand gestures).

Speech apraxia was evaluated by means of the MT86 clinical protocol [11], which consists in repeating words, pseudowords and sentences presented by the experimenter to the participants.

The MBLF (Motricity Bucco-Linguo-Facial) software was adapted to test the orofacial motricity [7]. Orofacial praxis, which may have an impact on speech production, was tested, namely of lips, tongue, cheeks and mandible. Instructions were given orally to the participants, then the articulatory gestures were presented. Upper limb apraxia was evaluated by the Mahieux's battery [18]. This battery includes three subtests consisting of the production of symbolic and mimetic gestures on verbal instructions by the experimenter, and abstract gestures on imitation of the ones produced by the experimenter. Finally, the NSE (*Niveau Socio-Educatif*) test was used to evaluate the participants' socio-educational level, as its impact on the results to the MMSE (Mini-mental state examination) has been proven [12]. Those tests are independent variables that would help to verify if our results are coherent with normalized tests, and to discuss the results obtained in the nursery-rhymes repetition task.

2.2. Data collection

All the recordings were performed by the same experimenter at the participants' home, using two camcorders (front and profile views), and a lapel microphone. The experimenter and the participant sat face to face on chairs, with a free space between them. In order to avoid the experimenter to converge phonetically with the participant and to minimize variation, the stimuli were preliminary recorded by the experimenter, played on a laptop and then repeated by the experimenter to the participant.

2.3. Data analysis

2.3.1. Evaluation of cognitive impairment

The score to the MMSE was calculated on a 30 points scale. A score greater or equal to 27 points indicates a normal cognition. Below this, scores can indicate mild (19-24 over 30), moderate (10-18 over 30) or severe (≤9 over 30) cognitive impairment. The MMSE was evaluated by our hospital partner (Dr. Olivier Moreaud's team, Department of Neurology, Grenoble Hospital) for the patients group. For the control participants, the MMSE was evaluated by the experimenter.

2.3.2. Evaluation of socio-educational level

The score to the NSE test was calculated as 1 point for no diploma, 2 points for a secondary school level, 3 points for a graduation level and 4 for higher education.

2.3.3. Evaluation of upper limb apraxia

In the Mahieux's battery [18], the scores for symbolic gestures and abstract gestures were calculated on a 1-point scale: 1 point when the gesture was recognizable and 0 when it was not. The score for mimetic gestures was calculated on a 2-point scale: 2 points for a normal realization of the gesture, 1 point for a persisting one-side body assimilation to the object, and 0 point for a false gesture or a bimanual body assimilation to the object. The performances of symbolic gestures were considered as abnormal when more than four out of five gestures were improperly executed (score of 4/5); for the abstract gestures, when six gestures out of eight were not well reproduced (score of 6/8); for the mimetic gestures, when eight out of ten gestures were not well produced (score of 8/10).

2.3.4. Evaluation of orofacial apraxia

The orofacial gestures were observed through the video recordings of the MBLF repetitions to calculate a score on a 3-point scale: 3 points for normal gesture, 2 points for an ample yet unmaintained gesture, 1 point for a flicker of contraction and 0 for an absence of contraction [7].

2.3.5. Evaluation of speech apraxia

The apraxia of speech was evaluated thanks to the words, pseudowords and sentences produced by the participants in the MT86 test, which were annotated using Praat© software in order to fill the scoring table in the most accurate way [11].

2.3.6. Analysis of hand gestures quality

The 8 deictic and 16 iconic gestures produced during the nursery-rhymes repetition task were annotated via ELAN© (EUDICO Linguistic Annotator software). Four criteria were selected as essential to determinate the hand gestures quality score, namely: emplacement, movement, configuration and orientation of the gesture. For each of them, a 2-point scale was used: 2 points for identical repetition, 1 point for non-identical repetition and 0 point for no repetition. As four criteria were evaluated on these 2-point scales, the total score was calculated on 8 points.

2.3.7. Analysis of speech

For the nursery-rhymes repetition task, the participants' speech production was annotated and analyzed with Praat©. Their errors were identified and classified in substitutions, omissions, or additions of phonemes and words, autocorrections, trials and repetitions of words.

2.3.8. Statistical analysis

Statistical significance was tested by means of the analysis software R[1]. For assessing differences between patients and controls, the Welch two-sample T-test was applied.

2.4. Participants

Eight right-handed French-native female speakers participated to this study (see Table 1). Four speakers were diagnosed with AD by our hospital partner. Their MMS score was comprised between 19 and 24 over 30, which corresponds to a mild cognitive impairment (mean score 21.7). These patients with AD were paired by age and socio-educational level to four control participants, which did not have a cognitive impairment according to their MMS score between 28 and 30 (mean score 29.5).

Code	Type	Age	MMS	NSE
pf1	patient	67	20	2
pf2	patient	70	24	2
pf3	patient	67	24	4
pf4	patient	81	19	2
cf1	control	62	28	3
cf2	control	63	30	4
cf3	control	67	30	4
cf4	control	77	30	4

Table 1: *Description of the tested population.*

[1] http://www.r-project.org/

The mean ages of the patients and the control participants were of 71 and 67, respectively. The age difference between the two groups was not statistically significant (t=-0.8379, p=0.43), while the difference in MMSE score was significant (p<0.01). Professional musicians were excluded from the trial. Socio-educational level, as evaluated by the NSE test, ranged from 2 to 4 for the speakers, with a mean score of 3.1 for the controls, and 2.5 for the patients. This difference was not statistically significant (p=0.08), which suggests the control group could serve as a reference for the patients.

3. Results

3.1. Production of hand gestures

3.1.1. Upper limb apraxia

The evaluation of upper limb apraxia, and more specifically of hand gestures, using the Mahieux's battery, is reported Table 2. The mean score was found to be higher in the control group (20.7/23) than in the patient group (15.50/23). However, the difference was not statistically significant (t(3.4)=2, p=0.1) because of group heterogeneity. Thus, only the performances of symbolic gestures of one patient (pf1) were considered as abnormal, all the other participants of the study produced correctly the symbolic gestures. For the mimetic and abstract gestures, the performances of the control group were evaluated as normal, while the productions of three out of four patients were out of the norm. One of the patients (pf3) obtained a high score, similar to the ones of the control group, and even better than some of the control participants. For both groups, symbolic and mimetic gestures were reproduced more successfully than abstract gestures, which could be due to the fact symbolic and mimetic gestures were produced on verbal instructions and more linked to language than abstract gestures [19].

Code	Total /23	Symbolic /5	Mimetic /10	Abstract /8
pf1	17	4	08	5
pf2	09	5	03	1
pf3	21	5	10	6
pf4	15	5	08	2
cf1	21	5	09	7
cf2	21	5	10	6
cf3	19	5	09	6
cf4	22	5	10	7

Table 2: *Mahieux test results for each participant.*

3.1.2. Hand gestures quality in nursery-rhymes repetition task

The patients with AD were able to repeat all the 8 deictic and 16 iconic gestures of the nursery-rhymes repetition task with good quality. For all criteria together, there was no task effect on the capacity to repeat hand gestures, although patients had a lowest mean score (6.5) than control participants (7.6).
For each criteria the quality was slightly lower for the patients, as assessed by the quality of hand emplacement (mean score 1.8/2), hand movement (mean score 1.5/2), hand configuration (mean score 1.5/2), and hand orientation (mean score 1.6/2), than for the control group's hand emplacement (mean score 2/2), hand movement (mean score 1.9/2), hand configuration (mean score 1.8/2), and hand orientation (mean score 1.9/2). Statistical differences were found between the two groups for the four quality criteria: '*movement*', '*configuration*', and '*orientation*'. Thus, patients only drafted the movement and produced their configuration was often lacking accuracy. About orientation, the patients' errors concerned mainly deictic gestures. One criterion had a celling effect: the '*emplacement*' (p < 0.01). Figures 1, 2 and 3 illustrate the quality scores measured for the three criteria '*movement*', '*configuration*' and '*orientation*'. Whereas the average score was high for both controls and patients, heterogeneity was observed within the patients, unlike the controls. A slight effect of the modality was seen for the '*movement*' and '*orientation*' criteria, however not significant.

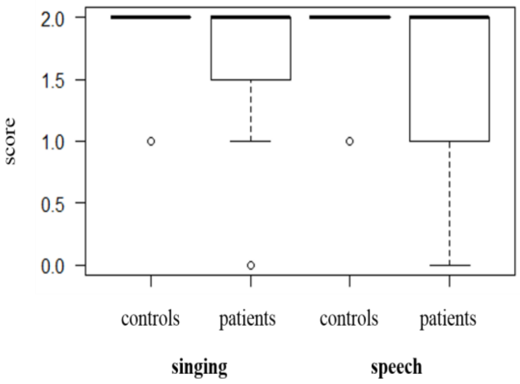


Figure 1: *Movement quality score in singing and speech.*

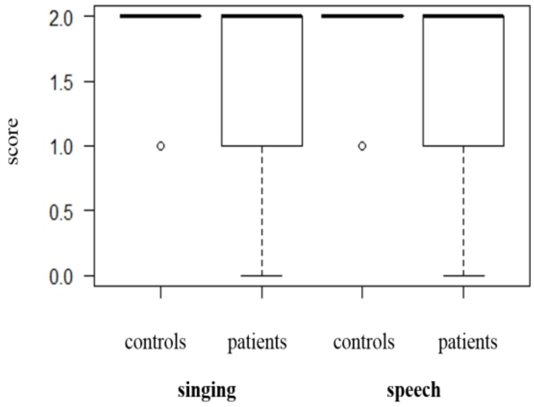


Figure 2: *Configuration quality score in singing and speech.*

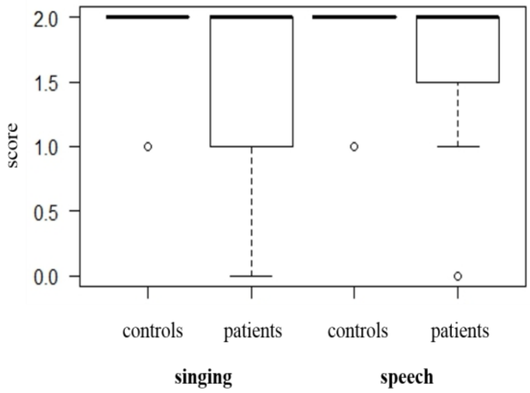


Figure 3: *Orientation quality score in singing and speech.*

Concerning the type of gestures, the patients had no difficulty to repeat accurately deictic gestures, as illustrated in Figure 4. The repetition of iconic gestures was more difficult for both groups. The patients' scores were lower and more heterogeneous than the scores obtained by the control participants.

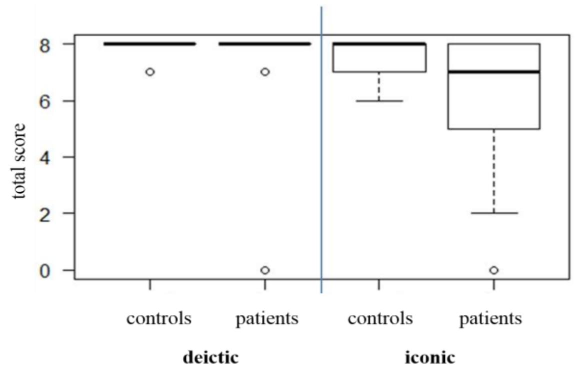


Figure 4: *Total score quality for deictic and iconic gestures.*

3.2. Speech production

3.2.1. *Orofacial apraxia*

Code	Total	Lips	Tongue	Cheeks & mandible
	/96	/27	/39	/30
pf1	72	24	33	15
pf2	76	18	34	24
pf3	96	27	39	30
pf4	74	24	30	18
cf1	68	25	31	12
cf2	96	27	39	30
cf3	96	27	39	30
cf4	93	27	38	28

Table 3: *MBLF test results for each participants.*

Patients did not show an orofacial apraxia, as assessed by the MBLF test (see Table 3). Though they obtained a lower average score (79.5/96) than controls (88.2/96), the difference was not statistically significant (t(5.7)=0.9, p=0.35). This can be explained by the scores of one of the controls (cf1) who got a lower score than all the participants. Also, one of the patients (pf3) got a ceiling score (96/96), which may be due to the fact that she practiced diction in acting classes for ten years.

3.2.2. *Speech apraxia*

Patients made more errors (25.5) than controls (27.7) in repeating words and pseudowords in the MT86 test, yet with no statistically significant difference (t(5.5)=1.5, p=0.17). All participants made the same types of errors, namely: phonemic substitutions first (e.g. /biʃu/ for /biʒu/ *'bijou'*), then phonemic omissions (e.g. /ɛ̃stʁytœʁ/ for /ɛ̃stʁyktœʁ/ *'instructeur'*) and phonemic additions (e.g/kɑ̃pandj/ for /kɑ̃paɲ/ '*campagne'*).

3.2.3. *Speech errors in nursery-rhymes repetition tasks*

Patients made more errors (204) in repeating the nursery rhymes than the control participants (49). Regarding phonemic and word errors, a significant difference (p<0.001) between patients' productions and the productions of the control participants was found. Patients made 61 phonemic errors, 83 word errors, and 60 other types of errors, when controls made 24 phonemic errors, 10 word errors and 4 other types of errors. As shown in Figure 5, patients' phonemic errors concerned first phonemic substitutions (45/61), omissions (13/61) and additions (3/61). In comparison, the control participants made firstly phonemic substitutions (13/24), omissions (11/24), and no additions. Most phonemic substitutions corresponded to a devoicing of consonants in a cluster (e.g. /plykʁɑ̃/ for /plykʁɑ̃d/ *'plus grand'*) or in an intervocalic context (e.g. /dapɔr/ for /dabɔr/ *'d'abord'*). Phonemic omissions concerned the last segment of a consonant cluster, most of the time the fricative (e.g. lapy/ for /laply/ 'la plus'). Phonemic additions consisted in the pronunciation of [ə] before a vowel (e.g. /tomatəavek/ for /tomatavek/ '*tomates avec'*) or in the addition of a consonant before another consonant (e.g. /paʁt/ for /pat/ *'pâte'*).

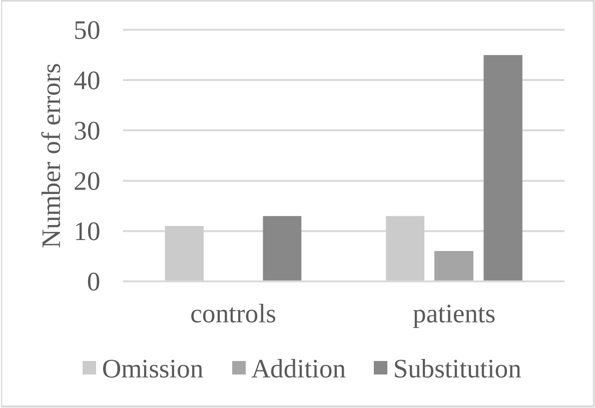


Figure 5: *Number of segmental errors in function of type.*

As shown in Figure 6, word errors made by the patients concerned first substitutions of word (41/83), before additions (23/83) and omissions (19/83). Words substitutions consisted in using synonyms (e.g. /swa/ for (mwa/ *'moi'*), or in suppressing or adding segments, for example prefixes (e.g. /depoʒe/ for /poʒe/). The controls produced only 10 errors of words: 5 omissions, 4 substitutions and a single addition.

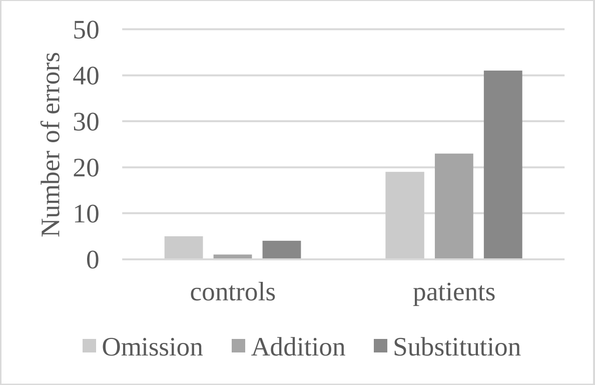


Figure 6: *Number of word errors in function of type.*

As shown in Figure 7, other types of errors were also observed (60 for patients and 4 for control participants): 22/60 tryouts (compared to 2/4 ones for control participants), 19/60 autocorrections (compared to 2/4 ones for control participants), and 19/60 repetitions for patients (no repetition for controls).

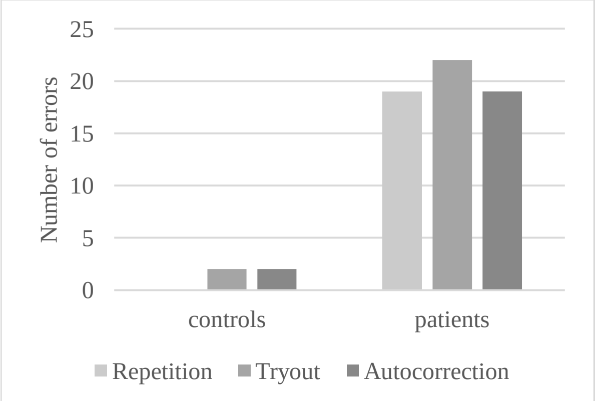


Figure 7: *Number of other types of errors.*

Moreover, for each type of errors, the patients score was higher in repeating the nursery rhymes with gestures (116/204) than those without gestures (88/204), as shown in Figure 8. This task effect, however less strong, was observed for the control participants as well, who made 13/24 errors in repeating the nursery rhymes with gestures and 11/24 errors in repeating the ones without gestures. Figure 8 also shows the effect of spoken-sung modality: for each type of errors, the patients made more mistakes while singing the nursery rhymes (114/204) than while speaking them (90/204). In comparison, the control participants made slightly more errors in singing (13/24) than in speech (11/24). The patients produced more errors in sung nursery rhymes with gestures (60/204), in spoken nursery rhymes with gestures (56/204) and in sung nursery rhymes without gestures (54/204), than in spoken nursery rhymes without gestures (34/204). The control participants made more errors in the spoken nursery rhymes with gestures and the sung nursery rhymes without gestures (7/24), and fewer in spoken nursery rhymes without gestures (4/24). As we noted, both groups made more mistakes in the nursery rhymes with connected gestures and the fewest in the spoken modality without gestures.

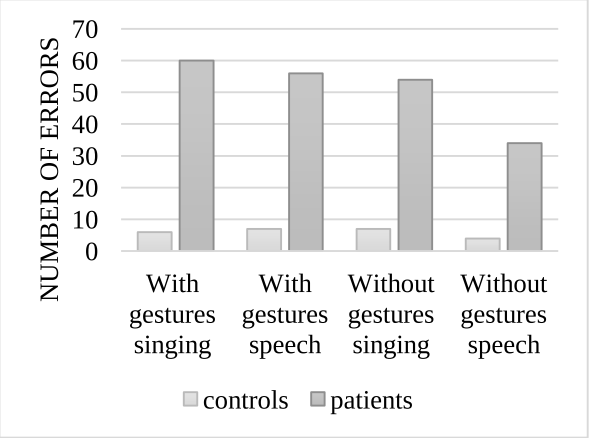


Figure 8: *Number of errors in function of hand gestures production.*

4. Discussion

No orofacial nor speech apraxia was evidenced by the following logopedic tests ([7], [11]), as the differences between patients and controls' scores were not significant. However the nursery-rhymes repetition task showed slight differences between patients and controls oral production. Mahieux's battery evidenced an upper limb apraxia, which can explain the results obtained at the analysis of hand gestures quality in the nursery-rhymes repetition task.

Concerning hand gestures quality, significant differences appeared on hand movement, hand configuration and hand orientation between patients and control participants. The fact that patients drafted the movement and that the configuration was produced with less accuracy could be an early sign of an upper limb apraxia. About the fact the errors produced by patients concerned mainly the deictic gestures could be explained by a decentering disorder typical of AD [23]. Although deictic gestures were more easily reproduced than iconic ones, which could be explained by the fact pointing develops before iconic gestures in speech ontogeny, and would be more anchored cognitively. Concerning the modality, neither speech nor singing did alter significantly the quality of hand gestures, which is not in line with our hypothesis.

An individual behavior can be pointed out: the patient (pf4) with the lowest score in the MMSE (19/30) and the Mahieux test score also had the lowest score to the nursery-rhymes repetition task, suggesting that cognitive impairment and upper limb apraxia could have an impact on the quality of voluntary hand gestures execution.

Our study showed hand gestures execution affects oral productions of AD patients. This phenomenon could be explained by a double task effect due to a cognitive overload, more important for people with AD who suffer of divided attention disorders. Those results are in contradiction with the positive effect of gestures on spontaneous speech production, notably on lexical retrieval ([6], [13]), which can be explained by the fact this study is based on a controlled task of repeated speech. Those results give rise to the automatic-voluntary dissociation [16], which besides is used to evaluate apraxia.

Moreover, a modality effect was observed for both groups: the sung modality with connected hand gestures was the task with most errors, and the speech modality without connected hand gestures was the task with fewer errors. In effect, the nursery rhymes were not known by the participants nor learnt prior to the repetition tasks; they were only presented once by the experimenter, and then repeated by the participant. Those results are opposed to the ones obtained in different previous works ([5], [10]), that only studied the impact of well-known songs, which implicate long-term memory and not working memory as in repeating tasks of unknown songs.

Regarding more precisely the errors made by the patients in the nursery-rhymes repetition task, they are in line with the MT86 scores, and with the literature on AD's oral communication impairments ([1], [15], [17] and [22]). Thus, words' omissions concur with word finding, the first disorder described in apraxia in AD. Words' substitutions could be caused by verbal paraphasia, and repetitions by palilalia. Concerning segment or prefixe additions, they could correspond to a verbal paraphasia. Phonemic omissions could be due to a phonemic disintegration or to a simplification process, as they concerned mainly liquids in clusters. Phonemic substitutions could be a consequence of a phonemic paraphasia, or of a dysphonia, which is coherent with literature on AD [22], as most of phonemic substitutions produced by patients corresponded to a devoicing of consonants in intervocalic context. As the MBLF score's difference between patient and control groups was not found to be significant in our study, errors made by patients could not be attributed to an orofacial apraxia. In particular, the patient pf3 had even a MBLF score similar to the control ones, but still made significantly more phonemic errors than the controls.

These preliminary results call for further exploration on a larger population, to avoid, or at least minimize, the effect of interpersonal variability. This study is still in progress, in particular with persons with AD showing more severe cognitive and hand gestures impairments, in order to keep investigating multimodal communication disorders at different stages of the disease. In summary, however no concomitant upper limb, speech and orofacial apraxia was preliminary evaluated thanks to the following logopedic tests ([7], [11], and [18]), an impact of hand gestures execution on oral repetitions was found in this study. This motor phenomena could be an argument in favor of a co-expressivity between hand gestures and speech that would go further than the semiotic dimension put forward by McNeill by involving an articulatory link [19].

5. References

[1] Arroyo-Anllo, E.M., Lorber, M., Rigaleau, F. & Gil, R., "Verbal fluency in Alzheimer's disease and Aphasia", Demencia, 11: 5-18, 2012.

[2] Carlomagno, S., Pandolfi, M., Marini, A., Di Iasi, G. & Crisitilli, C., "Coverbal gestures in Alzheimer's type dementia", Cortex, 41: 535-546, 2005.

[3] Caussade, D., Vallee, V., Henrich, N. & Colletta, J.-M., "Coordination/synchronisation gestes-voix dans la démence de type Alzheimer : un protocole expérimental original utilisant la parole et le chant", Journées de Phonétique Clinique 2013, Lièges, 2013.

[4] Caussade, D., "Mise au point et évaluation d'un protocole expérimental pour l'étude de la coordination/synchronisation gestes-voix dans la démence de type Alzheimer : évolution dans la parole et dans le chant", Mémoire de Master 2 recherche, Université Stendhal, Saint-Martin-d'Hères, 2013.

[5] Charrière, C. & Bally, M.-S., "Evaluation des capacités de communication par le biais de la chanson dans la maladie d'Alzheimer ; aspects scientifiques et rééducatifs", in P. Gatignol, La voix dans tous ses maux, Ortho Édition, 2009.

[6] Colletta, J.-M. & Guidetti, M. (eds.), "Gesture and Multimodal Development", Amsterdam, Benjamins Current Topics series, 2012.

[7] Gatignol, P. & Lannadere, E., "MBLF", Adeprio logiciels, 2011.

[8] Goldin-Meadow, S. & Butcher, C., "Gesture and the transition from one- to two-word speech: When hand and mouth come together", In D., McNeill (ed.): Language and gesture, Cambridge University Press, New York: 235-257, 2000.

[9] Gonseth, C., "Multimodalité de la communication langagière humaine : interaction geste/parole et encodage de la distance dans le pointage", Thèse de doctorat en Sciences Cognitives, Université Stendhal, Grenoble, 2013.

[10] Göttel, E., Brown, S. & Ekmn, S.-L., "Influence of caregiver singing and background music on posture, movement, and sensory awareness in dementia care" International Psychogeriatrics: 411-430, 2003.

[11] Joanette, Y., Nespoulous, J.-L. & Roch Lecours, A., "MT 86 – Protocole Montréal-Toulouse d'examen linguistique de l'aphasie", Ortho Edition, Paris, 1998.

[12] Kalafat, M., Hugonot-Diener, L. & Poitrenaud, J., "Standardisation et étalonnage français du 'Mini Mental State' (MMS) version GRECO", Revue de Neuropsychologie, 13 (2): 209-236, 2003.

[13] Kendon, A., "On the origins of modern gesture studies", In S. K. Duncan, J. Cassell & E. T. Levy (eds.), Gesture and the dynamic dimension of language, John Benjamins, Amsterdam: 13-28, 2007.

[14] Lee, H., "Vieillissement normal et maladie d'Alzheimer : analyse comparative de la narration semi-dirigée au niveau lexical", JéTou, 2011.

[15] Lefebvre, L. & Pruvost, C., "Langage et nombre chez des sujets atteints de démence de type Alzheimer : aspects syntaxiques et lexico-sémantiques", Glossa, 108: 30-52, 2010.

[16] Lennart Wallin, N., Merker, B. & Brown, S.(dir.), *The origins of music*, Cambridge, MA, MIT Press, 2012.

[17] Luchesi Cera, M., Zazo Ortiz, K., Ferreira Bertolucci, P. & Cianciarullo Minett, T., "Speech and orofacial apraxias in Alzheimer's disease", International Psychogeriatrics, 25(10): 1679-1685, 2013.

[18] Mahieux-Laurent, F., Fabre, C., Galbrun, E., Dubrulle, A. & Moroni, C., "Validation d'une batterie brève d'évaluation des praxies gestuelles pour consultation mémoire. Evaluation chez 419 témoins, 127 patients atteints de troubles cognitifs légers et 320 patients atteints d'une démence", Revue neurologique, 65: 560-567, 2009.

[19] McNeill, D., "Language and Gesture", Cambridge University Press: 313-328, 2000.

[20] OMS., "La démence", Aide-mémoire, 362, 2015. Online: http://www.who.int/mediacentre/factsheets/fs362/fr/, accessed on 22 April 2015.

[21] Parakh, R., Roy, E., Koo, E. & Black, S., "Pantomime and imitation in limb gestures in relation to the severity of Alzheimer's disease", Brain and Cognition, 55: 272-274, 2010.

[22] Pinto, S. & Ghio, A., "Troubles du contrôle moteur de la parole : contribution de l'étude des dysarthries et dysphonies à la compréhension de la parole normale", Revue française de inguistique appliquée, 8 : 45-57, 2008.

[23] Rousseau, T., "Maladie d'Alzheimer et troubles de la communication", Elsevier-Masson, Issy-les-Moulineaux, 2011.

[24] Sataloff, R., T., Caputo Rosen, D., Hawks, M., Spiegel, J., R., "The three ages of voice: The aging adult voice", Journal of voice, 2(11): 156-160, 1997.

[25] Taglialatela, J. P., Russell, J. L., Schaeffer, J. A. & Hopkins, W. D., "Chimpanzee Vocal Signaling Points to a Multimodal Origin of Human Language", PLoS One, 2011.

[26] Viader, F., Lambert, J., De La Sayette, V., Eustache, F., Morin, P., Morin, L & Lechavalier, B., "Aphasie. Encyclopédie Médicale Chirurgicale", Neurologie, Editions Scientifiques et Médicales Elsevier SAS, Paris, 17: 32, 2002.

Towards a Theory of Compositionality in Displays of Extreme Emotion

Federica Cavicchio, Wendy Sandler

University of Haifa, Sign Language Research Lab
federica.cavicchio@gmail.com, wsandler@research.haifa.ac.il

Abstract

Compositionality – the combination and recombination of meaningful units to create more complex structure, – is a defining property of human language. Here we seek the foundations of this property in a more basic form of communication: the expression of emotion. We collected 300 pictures of athletes, moments after winning or losing a competition. We annotated face and body displays in detail, and checked prototypical displays in winning and in losing contexts. We identified features of face and body reliably used in each situation, and some used in both, paving the way for a theory of compositionality in the expression of emotions.

Index Terms: emotion theory, compositionality, multimodal communication

1. Introduction

Language is a compositional system in which the meaning of a complex structure is determined by the meanings of its constituent components and the way they combine.. This property characterizes all human language, whether spoken [1] or signed [2]. Here we seek to determine whether nonverbal communication has compositional properties as well. Specifically, we hypothesize that compositionality transcends language and is rooted in the most "primitive" of the human communication systems: the expression of emotions. To this end, we ask whether facial expressions and body postures are combined and recombined to convey different emotional meanings in extreme displays of emotions. Specifically, we consider two approaches, each of which makes different predictions for our data. The **compositional** approach predicts that individual components can be reliably associated with particular interpretations and may recombine, lending their interpretations to different arrays. The **holistic** approach makes the opposite prediction, that multi-component configurations are interpreted as gestalts. Here we take a first step toward distinguishing the two by identifying prototypical face and body elements present in victory and defeat situations, each of which often triggers an array of intense emotions.

Since Darwin's seminal work [3], many models of emotion have attempted to explain the concept of emotion and how the body “contributes a content that is part and parcel of the workings of the mind” [4]. Broadly speaking, there are currently two main approaches to the description of emotion: the Basic/Prototypical Emotions approach ([5], [6], [7], [8]) which we call here the holistic approach, and the Dimensional/Appraisal approach ([9], [10], [11], [12], [13]). As we will show, the dimensional approach is conceptually closer to our notion of compositionality, though the motivations and methodologies differ.

In the holistic view, emotions are “affect programs” and facial expressions are residual actions of more complex behavioral responses combining vocal, postural, gestural and skeletal muscle movements. For example, a basic emotion such as fear is a hardwired response to a threatening stimulus that activates a certain brain area (or brain circuit) associated with a "fight or flight" response, which in turn activates particular facial expressions and body postures. Facial expressions of emotion may also be modified or inhibited by cultural display rules. All the other emotive states beyond the basic set are considered to be "blends" of basic emotions. Facial expressions are usually coded using the Facial Action Coding System (FACS, [14]), which annotates each observable facial movement as an Action Unit (AU), so that all displays perceived as facial expressions can be coded in terms of their constituent AUs. In the holistic view, although the facial expressions of basic emotions are comprised of a number of action units, they are considered to be gestalts.

On the other hand, dimensional models of emotions, such as 2D circular models of valence and arousal [9], do not view basic emotions as biologically hardwired gestalts, but rather as phenomena that emerge from combinations of behavioral responses. For example, in the expression of fear, a complex facial expression involving a number of action units, the specific characteristic, widening of the eyes, (AU 5), is hypothesized to have evolved from the attempt to widen the visual field in response to threatening stimuli ([15], [16]).

Another group of emotion models that adopts the dimensional approach are appraisal models. Appraisal theories of emotions propose a model according to which the final emotive status (and the consequent facial expression) is a product of a series of appraisals checks on the part of the experiencer ([17], [18]). Appraisal models go beyond the classic valence and arousal distinction to propose that several dimensions are at play when we appraise an emotion-inducing stimulus, and that these are reflected in different facial movements. These dimensions are: relevance of a stimulus, intrinsic pleasure, implications in terms of goal conduciveness, coping potential and norm compatibility. These five dimensions are appraisal domains that can be decomposed by appraisal check. For example, relevance can be decomposed into two appraisal checks, novelty and pleasantness. These move along the continua sudden/familiar (for novelty) and pleasant/unpleasant (for pleasantness). Appraisal theories do not endorse the idea of a small number of basic emotions, but rather propose that there is a large number of different emotions which may combine with one another ([17], [18], [19]).

To test this hypothesis, Scherer et al [20] analyzed the facial expressions of four positive emotions in the GEMEP corpus using FACS. In the GEMEP (GEneva Multimodal Emotion Portrayal, [21], [22]) corpus, 10 actors expressed 18 emotions, uttering the same meaningless speech strings in different emotional contexts. For this study, the authors selected a subset of the emotions portrayed in the corpus: interest, joy, pride, and pleasure. Results of the FACS coding showed that the frequency and patterning of the AUs could not be explained using holistic emotional categories such as these. The facial expressions did not show significant differences between joy and pride, for example. Instead, contrasting emotions for appraisal checks was a more accurate predictor of different facial displays. In particular, the appraisal dimension of novelty in interest and joy was reflected in the degree of eye opening (Action Unit 5 of FACS), whereas cheek raise (AU6) was characteristic of intrinsically pleasant emotions (such as joy and pleasure), and eyelid tightening (AU7), of goal conduciveness (as in pride).

Though Darwin's observations included the whole body, body posture in the expression of emotions has not received the same attention as facial expression. In fact, it has long been assumed that, whereas a number of facial muscle configurations are reliable indicators of specific emotions, body movements or postures provide information of intensity only ([23], [24], [25]). However, recent studies show that variations in body movement and posture convey specific information about emotional states ([26], [27], [28], [29]), and that a change in body context ([30], [31]) or in the external context in which the body and face are inserted ([32], [33]) changes the way in which the emotion is perceived and categorized. As noted, only a limited number of studies have measured the physical cues that express emotion in the body ([34], [35], [36], [37], [38]). The main reason for this dearth of research is the lack of an established coding system for the body that would be comparable to the face and voice measurement techniques (e.g., [39]) that have facilitated systematic research on emotion expression in those modalities. Another problem is that the few systems that have been developed to investigate body expressions (e.g. [37], [40]) have usually relied on displays of actors rather than on spontaneous emotional displays. For example, Dael et al.[41] explored a subset of the GEMEP corpus using 49 behavioral categories belonging to 12 emotions, both basic and subtle, representing the two poles of the valence and the arousal continua. They found that hot anger, amusement and pleasure were characterized by distinct patterns of body behaviors, such as forward body movement for hot anger, self-touching and neutral head position for amusement, and head tilted up for pleasure. In contrast, many emotions considered basic, such as joy, panic and fear, were not reliably represented by any specific body pattern. What emerged instead were two bi-dimensional patterns grouped around the arousal and valence dimensions, which were not sufficient to explain all the body displays. Distinct clusters of behaviors also emerged for emotions having the same potency (on a strong vs weak continuum) and attentional activity (interesting vs not interesting). Those results are consistent with previous findings on facial expressions of emotions [42]. Results showed that an emotion could be encoded by a variety of behavior patterns, suggesting that emotion dimensions such as valence, arousal, power and attention - and not classic affect programs like fear, happiness, etc. - drive the bodily expression of emotions. It is interesting to note that Dael et al. [41] also found that some displays were shared by different emotions: panic fear and elated joy share symmetry of arm actions and knee movements; sadness and relief had the same "arm along the body" posture; and interest and irritation share asymmetrical one-arm action and trunk leaning forward movements. These results suggest to us that the same body behaviors with different combinations of face and head movements may convey different emotional meanings in a compositional fashion, a hypothesis we wish to test.

In the present study we try to overcome the limitation of using actors to pose stimuli by investigating the facial expressions and body postures of athletes' pictures taken moments after they won or lost a high-stakes competition, in order to capture expressions that were extreme and spontaneous. We assume that emotional displays that are both extreme and spontaneous are less likely to be filtered by social or cultural conventions and inhibitions than other expressions of emotion. Following Aviezer et al. [30], we collected 300 pictures of athletes shot seconds after their victory or defeat. These two contexts ensure both spontaneity and emotions of opposite valence in high arousal contexts. We annotated the facial expressions using FACS, and the body features using a similarly motivated coding scheme that we developed and validated, which codes 25 different components of body positions. We found that specific sets of facial and body features were highly correlated with winning and losing contexts, respectively, whereas other features were mildly correlated with each context. Finally, a small set of facial and body features were shared by the two contexts, and we hypothesize that they share particular dimensions of emotion contributing to the interpretation of these displays. Our data show that particular face and body actions combine in the expression of emotions, paving the way for the development of a compositional model encompassing the whole human form. We aim to incorporate insights from the dimension approach by explicitly evaluating the interaction of face and body features in ongoing perception experiments.

2. Method

2.1 Data Collection

Following Aviezer et al. [30], we searched Google Images for strings of text such as "reaction to win" and "reaction to lose", but, unlike Aviezer et al [30], who restricted his research to 180 pictures from tennis matches, we collected 300 pictures from badminton, boxing, fencing, judo, rugby, tennis, table tennis, football, volleyball, and track and field, most of them from the 2012 London Olympics. Of these 300 pictures of athletes taken seconds after winning or losing a competition, 136 images pictured defeat, and 164 victory. For the defeat category, 50 pictures portrayed women and 86 men, and for win, 70 images portrayed women and 94 men. Athletes' country of origin varied, including both Western and Eastern countries. To ensure extreme, spontaneous displays, we sought pictures of athletes in high stakes competitions moments after their victory or loss was determined (and not when medals are awarded for example). To verify that the pictures were taken a few seconds after the event, we Google searched for the corresponding videos of the sport events and confirmed that the pictures were taken in a time span no longer than 10 seconds after the win or the loss. In this study, pictures were preferred to videos because the quality of videos taken from the Internet was often too poor for accurate coding of facial expressions.

2.2 Data Coding

To code facial expressions, neck tightening and head positions, a certified coder used FACS. To code the body features we developed our own coding scheme, the Body Arrangement Coding System (BACS), which focuses on the position of different parts of the body with respect to the main articulators and joints. Our system also facilitates coding of interaction among articulators. For example, we coded the type of interaction between hands and head/face/body (when applicable), using labels such as hand in front of the face, covering mouth, covering eyes, on top of the head, on the back of the head, on the knees, on the chest etc. Each body articulator was coded separately: head, neck, shoulders, arm position along the X, Y and Z space axes; chest, torso, leg split, knees, palm direction, hand shape. Right and left articulators were coded separately to capture asymmetries (e.g. right arm vs left arm, right shoulder vs left shoulder etc.). To assess coding scheme reliability, 4 coders independently annotated 40 pictures taken from the corpus. The 12 categories yielded an intercoder agreement with kappa scores between 0.73 and 0.95, which are considered good for multimodal annotation of emotions [43].

2.3 Data Analysis

A total of 305 features distributed over 29 categories were used to code facial expressions, head positions, hands to head, neck and body posture. To reduce the data dimensions, we performed a Multiple Correspondence Analysis (MCA), a particular type of Correspondence Analysis suited to multiple categorical variables. The MCA model collapsed and simplified the data by reducing the number of parameters in our dataset and finding the ones that were significant for the descriptions of win and loss in terms of face and body features. We ran two separate statistical models: one included all units of the face and head: facial expression, head position, and neck tightening, as well as hands to head/face The other included all the body features beneath the neck. As we were interested in the facial expressions and body postures in Win and Loss contexts, we tagged each picture according to Win or Loss context of occurrence, and included Win and Loss in the statistical analysis, to see whether there was a high correlation between these contexts and the face and body features coded. We tagged pictures for Gender as well, as a potentially correlated factor. MCA models were run using FactoMineR package implemented in R 3.0.3 [44].

A first MCA was run on the whole set of pictures (N=300) for the face and head: facial Action Units (divided according to upper face, lower face and nasal area Action Units), Head Position AUs and Neck AUs, and position of the Hands on Face/Head. The first component of the MCA accounted for 15.7% of the total variance of the data, and the second component for 10.5%. Correlations are observable according to the proximity of features/tags that occur together. Surprisingly, Gender was correlated neither with the first nor the second component, whereas Win and Loss were highly correlated with the first component.. As shown in Table 1, particular groupings of facial AUs of different parts of the face -- the lower face, the nasal area, and the upper face -- were highly correlated with the first component and described most of the data variability (**R**^{2} >0.5). Neck AUs and head position AUs were fairly well correlated with the first component (**R**^{2}~0.5). Hand to Face/Head was highly correlated with the first component (**R**^{2} >0.5). In the table, coded features appear above the line, and tagged features of Win/Loss and Gender appear below the line.

Specific features typically clustered with win, and others with loss, with a few overlapping between the two contexts. Winning athletes typically produced a more complex set of facial expressions than losing athletes, exemplified in Fig 1. In particular, for upper face, AUs 4 (brow lowerer), 6 (cheek raiser) and 7 (lid tightener) were frequently found in combination with other AUs. For lower face, AUs 25 (lips part) and 27 (mouth stretch) were found in many of the combinations. In contrast, loss was typically characterized by neutral or "not visible" facial features (see Fig. 1). However, some features correlated with both win and loss. We found that closed eyes (AU43) occurred with both victorious and defeated athletes, but in defeated athletes it occurred without other upper face AUs, while in winning athletes, it occurred in combination with AUs6 and 7 (cheek raise and lower lid tightening). Lip parting (AU 25) was also found in winning and losing athletes, but each context contributed different additional features of mouth opening.

Table 1. Correlation coefficients and p values between the face, neck, head and hands to head variables and the first component of the MCA.

	R^2	p.value
Lower Face_AUs	0.9	>0.001
NasalArea_AUs	0.8	>0.001
UpperFace_AUs	0.8	>0.001
LeftHandtoFace/Head	0.8	>0.001
RightHandtoFace/Head	0.8	>0.001
Neck_AUs	0.5	>0.001
HeadPosition_AUs	0.4	>0.001
Win_Loss	0.7	>0.001
Gender	0.01	0.6

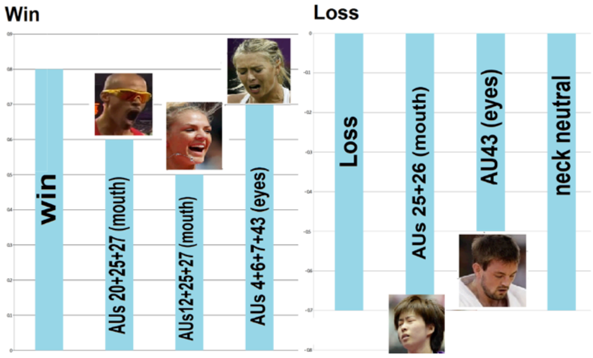


Figure 1. Estimate values of the Face and Neck Action Units for the first component. AUs with positive estimates belong to the winning context. A selection of the AUs that yield an Estimate >0.5 are reported.

Regarding head position, winning athletes had their heads up (AU53) in combination with other head positions such as head forward (AU57) or turned left (AU51, see Fig. 2).

Interestingly, head up (AU53) is found in defeated athletes too, but alone, not in combination with other head features. Losing athletes often had head down (AU54) sometimes in combination with head forward (AU57). Regarding hands to face/head, winning athletes tend to put their hands away from the face, or to place their hands on the mouth or on top of the head, whereas defeated athletes tend to cover the whole face with their hands or place one or both hands on the upper face and eyes area, or (less often) on the back of the head. When only one hand touches the forehead, winning athletes tend to place their right hand on the forehead, whereas athletes that just lost tend to cover their forehead with their left hand.

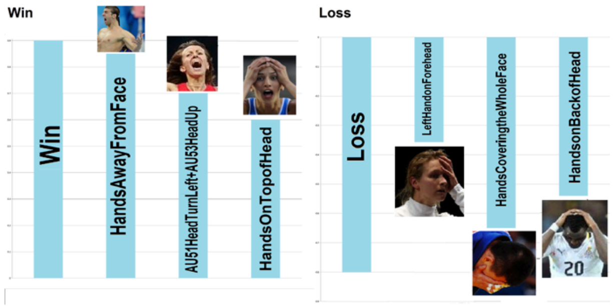


Figure 2. Estimate values of the Head Movement Action Units and Hands to Body/Face for the first component. A selection of the features that yield an Estimate >0.5 are reported.

For the body features, we have coded 80 pictures so far. A second MCA was run on the results of this coding. The first component explained 16.7% of the variability and the second component explained 8.4% of the total variability. Table 2 reports the R^2 and p. values of the body features that were found significant.

Table 2. Correlation coefficients and p values between the body features and the first component of the MCA.

	R^2	p.value
ArmRight&Left_Z	0.6	>0.001
ArmRight&Left_XY	0.5	>0.001
ArmForearmR&L	0.45	>0.001
ShoulderR&L	0.3	>0.01
PalmR&L	0.45	>0.001
PalmDirectionR&L	0.2	>0.001
HandTouchBodyR&L	0.15	=0.01
Chest	0.4	>0.001
Torso	0.4	>0.001
LegR&L	0.2	>0.01
TouchingGround	0.4	>0.01
Win_Loss	0.6	>0.001
Gender	0.01	0.3

Win/loss is fairly well correlated with the first component. Again, Gender was not correlated significantly with either the first or the second component of the model. We found that the arm position was fairly well correlated with the first component, as were the shoulders, chest and torso positions and the palm configuration. The position of lower parts of the body was less correlated with the first component, but the athletes' proximity to the ground was well correlated (standing, sitting, touching the ground with the hand(s), forehead, etc.)

In Fig. 3 we report the body features along the win and loss axis. Broadly speaking, winners' bodies are open and extended while those of losers are closed and diminished in size. Winning athletes are typically standing, and stretch their arms up over their heads, shoulders raised, palms clenched and directed away from the body. Defeated athletes typically hold their arms down and bent more than 90 degrees at the elbow, often to cover their face with their hands. Shoulders forward, chest closed and torso and legs bent; palms touching in the praying position or stretched (fingers are stretched with respect to the palm and separated from each other) and directed towards the body. We are now in the process of coding the remaining 220 pictures to test our initial findings for robustness.

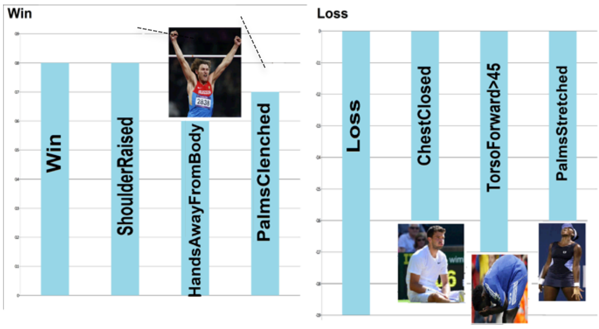


Figure 3. Estimate values for body. Features with positive estimates belong to the winning context. A selection of the body features that yield an Estimate >0.5 are reported.

3. Discussion and Conclusions

In the previous section we reported the face and body features that were highly correlated with winning and losing contexts. A small set of such features was shared between the two contexts. In particular, eye closure, mouth opening, and head forward were found in both win and loss sets of pictures. Head up is another component shared between the two emotion contexts, as was touching the upper part of the head, though on different parts of the head, with different hands, and in combination with different units in each context. While Aviezer et al's [30] study uses very similar pictures and contexts, it only reports judgments of positive or negative/winning or losing and did not analyze the face and body displays themselves. Our results may help to explain why participants in that study were not able to judge the outcome of a tennis match by looking only at the athlete's facial expression: features shared by winning and defeated athletes may have confounded their judgements. It is possible that precisely those features that are shared are more salient than those that we found to reliably distinguish the two displays, a suggestion that we will follow up in ongoing research.

On the other hand, Aviezer et al. [30] found that participants were capable of correctly discerning a winning from a defeated tennis player from the body posture alone. In our study no components of the body that were highly correlated with either winning or losing were shared between the two contexts, explaining the participants' success. In short, facial displays can be ambiguous while body displays are not (or are less so). Our preliminary interpretation is that the correspondence between positions of the large, salient articulators of the body and the emotions that prompt them is both more clearly perceivable and less complex and therefore less ambiguous than that between articulations of the face and their corresponding emotions. As we have said, there have been few studies of body displays, and those that have been conducted were in different contexts. The body displays we found in our 80 pictures are quite different from the ones found by Dael et al. [41], where, for example, head up was a distinctive characteristic of pleasure, while in our contexts we found that head up was a feature shared between win (presumably pleasurable) and loss. It is too early to say whether such differences are due to the different coding schemes, the use of posed vs. spontaneous displays, differences in extremeness/intensity of emotion, or differences in the head and face units with which they combine.

As regards the emotion models, our results are in contrast with the basic emotion (holistic) theory, which holds that whole configurations of facial action units characterize each basic emotion. Although some units overlap between different emotions in the holistic model (e.g., brow lowerer and upper lid raise in both prototypical anger and fear), their contribution is not compositional; i.e., neither the individual units nor groups of units on different parts of the face are analyzed as making independent contributions of meaning on the holistic approach.

Our results are partially compatible with the dimensional model of emotions. For example, as high stakes winning and losing are potentially both high arousal events with opposite valence, one could hypothesize that the shared components such as those mentioned above might be linked to the degree of arousal and not to the nature of that arousal, i.e., not to valence. Our working hypothesis is that individual units, or minimal combinations of units of the upper face, the lower face, and the upper and lower body, will distinguish interpretations of corporeal expression; i.e., the displays are compositional.

Comparison of findings in the contexts we are examining with those of other studies is expected to elucidate what these units and combinations are, and how they contribute to interpretation. Interesting contrasts in this direction emerge when comparing facial features associated with contexts of opposite valence such as elated happiness and sadness/despair in Scherer and Ellgring's study using actors [18] with those in our study of spontaneous reactions to victory and defeat. For example, AU4, brow lowerer, is common in sadness and despair in [18], but it is common in winners (and not losers) in our study. Brow lowering in winners is problematic for the dimensional/appraisal approach, because this AU is predicted to be present in appraisals of unpleasantness, relevant discrepancy, or lack of coping control, none of which is compatible with victory. The presence of brow lowering in spontaneous victory displays in our study, as well as in the unpleasantness contexts of the laboratory study suggests that this feature, whatever its 'meaning', is not part of a holistic display, thus lending support to our compositionality hypothesis.

In sum, our initial results show that a compositional approach to understanding corporeal displays of emotion is crucial for investigating emotion. Importantly, we are now conducting experiments to determine how participants categorize the emotions conveyed by different combinations of features in the same naturally occurring displays of emotion. To further test how the facial and body features re-combine and whether they convey meanings alone or in combination with other features, we are working to create new stimuli in which body and facial expressions highly correlated to win will be combined with lower correlated ones or with facial and body expressions of loss, to try to isolate and test the contributions of individual features and feature groupings. We expect these studies to lead to the creation of further complex stimuli to use in interpretation experiments. By comparing the results of these different lines of research, we aim to derive testable hypotheses about compositionality in the expression of emotion.

4. Acknowledgements

This research is supported by ERC Grant Agreement 340140 for the Grammar of the Body project.

5. References

[1] Jackendoff, R., "What is the human language faculty?:Two views", Language 87(3): 586-624, 2011.

[2] Sandler, W., Lillo-Martin, D., "Sign language and linguistic universals". Cambridge University Press; Cambridge, UK: 2006.

[3] Darwin, C., "The expression of the emotions in man and animals." Chicago: University of Chicago Press. 1965. (Original work published 1872)

[4] Damasio A.R., "The somatic marker hypothesis and the possible functions of the prefrontal cortex". Transactions of the Royal Society, 351 (1346): 1413–1420, 1996

[5] Tomkins, S., "Affect Imagery Consciousness: Volume I, The Positive Affects". London: 1962.

[6] Izard, C. E.,"Innate and universal facial expressions: Evidence from developmental and cross-cultural research". Psychological Bulletin, 115, 288-299, 1994.

[7] Ekman, P., "Universals and cultural differences in facial expressions of emotion". In J. Cole (Ed.), Nebraska Symposium on Motivation, 1971 (Vol. 19, pp. 207–283). Lincoln: University of Nebraska Press, 1972.

[8] LeDoux J.E., Cicchetti P., Xagoraris A., Romanski L.M., "The lateral amygdaloid nucleus: sensory interface of the amygdala in fear conditioning". Journal of Neuroscience, 10(4): 1062-1069, 1990.

[9] Russell, J. A., "Culture and the categorization of emotions". Psychological Bulletin, 110, 426–450, 1991.

[10] Barrett, L. F. "Solving the emotion paradox: Categorization and the experience of emotion". Personality and Social Psychology Review, 10, 20-46, 2006.

[11] Barrett, L. F., Wilson-Mendenhall, C. D., & Barsalou, L. W. "The conceptual act theory: a road map". Chapter in L. F. Barrett and J. A. Russell (Eds.), The psychological construction of emotion (p. 83-110). New York: Guilford, 2015.

[12] Ortony, A., Turner, T. J., "What's basic about basic emotions?" Psychological Review, *97*, 315-331, 1990.

[13] Scherer, K. R., "On the nature and function of emotion: A component process approach". In K. R. Scherer & P. Ekman (Eds.), Approaches to emotion (pp. 293–317). Hillsdale, NJ: Erlbaum. 1984.

[14] Ekman, P., & Friesen, W. V.,The Facial Action Coding System: A technique for the measurement of facial movement. Palo Alto, CA: Consulting Psychologists Press, 1978.

[15] Gould, S.J., Lewontin, R.C. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme" Proceedings Royal. Society London 205 581–598. 1979.

[16] Barrett, L. F. "Emotions as natural kinds?" Perspectives on Psychological Science, 1, 28-58. 2006.

[17] Lazarus, R. S., Folkman, S. Stress,appraisal, and coping. New York: Springer, 1984.

[18] Scherer, K. R.; Ellgring, H. "Multimodal expression of emotion: Affect programs or componential appraisal patterns?" Emotion,Volume 7(1), 158-171, 2007.

[19] Scherer, K. R., Mortillaro, M., Mehu, M., "Understanding the mechanisms underlying the production of facial expression of emotion: A componential perspective". Emotion Review, 5(1), 47-53 (2013).

[20] Mortillaro, M., Mehu, M., Scherer, K. R., "Subtly different positive emotions can be distinguished by their facial expressions". Social Psychological & Personality Science, 2(3), 262-271, 2011.

[21] Scherer, K. R., & Bänziger, T., "On the use of actor portrayals in research on emotional expression". In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), Blueprint for affective computing: A sourcebook, 166-176. Oxford, England: Oxford university Press, 2010.

[22] Bänziger, T., Mortillaro, M., Scherer, K. R., "Introducing the Geneva Multimodal Expression Corpus for Experimental

Research on Emotion Perception". Emotion, 12(5), 1161-1179, 2012.

[23] Ekman, P., "Differential Communication of Affect by Head and Body Cues". Journal of Personality and Social Psychology, 2(5),726-735, 1965.

[24] Ekman, P., Friesen, W. V., "Head and Body Cues in the Judgement of Emotion: A Reformulation". Perceptual and Motor Skills, 24, 711-724. (1967).

[25] Harrigan, J. A. "Proxemics, Kinesics, and Gaze. The New Handbook of Methods in Nonverbal Behavior" Research: 137–198, 2005.

[26] Atkinson, A. P., Winand H. Dittrich, A. J. Gemmell, A., Young, A.W. "Emotion Perception from Dynamic and Static Body Expressions in Point-light and Full-light Displays." Perception-London 33: 717–746, 2004.

[27] Coulson, M. "Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence." Journal of Nonverbal Behavior 28 (2): 117–139. 2004.

[28] Tracy, J. L., Robins, R. W. "Show your pride: Evidence for a discrete emotion expression". Psychological Science, 15, 194–197.

[29] Peelen, M.V., Downing P.E., "The Neural basis of visual body perception". Nature Neuroscience Review, 636-648, 2007.

[30] Aviezer, H., Trope, Y., Todorov. A., "Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions". Science, 338 (6111): 1225, 2012.

[31] Aviezer, H., Hassin, R.R., Ryan, J., Grady, C., Susskind, J., Anderson, A.,"Angry, Disgusted or Afraid?" Psychological Science, 19(7), 724-732.

[32] de Gelder, B., Meeren, H. K. M., Righart, R., Van den Stock, J. , van de Riet, W. A. C., Tamietto, M., "Beyond the face: Exploring rapid influences of context on face processing". Progress in Brain Research, 155, 37-48. 2006.

[33] de Gelder, B., Van den Stock, J., "Real faces, real emotions: perceiving facial expressions in naturalistic contexts of voices, bodies and scenes". In A.J. Calder, G. Rhodes, J.V. Haxby & M.H. Johnson (Eds.), The handbook of face perception. Oxford: Oxford University Press. 2012.

[34] Boone, R. T., Cunningham, J. G., "Children's expression of emotional meaning in music through expressive body movement". Journal of Nonverbal Behavior, 25, 21– 41. 2001.

[35] Gross, M. M., Crane, E. A., Fredrickson, B. L., "Methodology for assessing bodily expression of emotion". Journal of Nonverbal Behavior, 34, 223–248. 2010.

[36] Sawada, M., Suda, K., Ishii, M., "Expression of emotions in dance: Relation between arm movement characteristics and emotion". Perceptual and Motor Skills, 97, 697–708. 2003.

[37] Wallbott, H. G., "Bodily expression of emotion". European Journal of Social Psychology, 28(6), 879–896. 1998.

[38] Scherer, K. R., Wallbott, H. G. "Analysis of nonverbal behavior". In T. A. van Dijk (Ed.), Handbook of discourse analysis (pp. 199–230). London: Academic Press. 1985.

[39] Scherer, K. R., "Vocal affect expression: A review and a model for future research". Psychological Bulletin, 99, 143-165. 1986.

[40] Dael, N., Mortillaro, M., Scherer, K. R., "The Body Action and Posture Coding System (BAP): Development and Reliability". Journal of Nonverbal Behavior, 36(2), 97–121. 2012.

[41] Dael, N., Mortillaro, M., Scherer, K., "Emotion expression in body action and posture". Emotion, 12(5), 1085-1101. 2012.

[42] Osgood, C. E., "Dimensionality of the semantic space for communication via facial expressions". Scandinavian Journal of Psychology, 7, 1–30. 1966.

[43] Cavicchio, F., Poesio, M., "Multimodal Corpora Annotation: Validation Methods to Assess Coding Scheme Reliability". In M. Kipp, J.-C. Martin, P. Paggio, and D. Heylen (Eds.),Multimodal Corpora, LNAI 5509, Berlin: Springer-Verlag, 109-121. 2009.

[44] Le, S., Josse, J., Husson, F., "FactoMineR: An R Package for Multivariate Analysis", Journal of statistical software, 25(1), 2008.

Disposition and noise interference as factors of zero-lag interpersonal coordination

Carlos Cornejo[1], Esteban Hurtado[1,2], Himmbler Olivares[1]

[1] Pontificia Universidad Católica de Chile
[2] Universidad Diego Portales, Santiago, Chile.
cca@uc.cl

Abstract

Pursuing the goal to study interpersonal coordination from a more ecological point of view we conducted a study on interpersonal coordination using a MoCap system. A total of 20 female and 16 male undergraduates (ages 18 to 28) were randomly matched in 18 couples for having a conversation. Each couple was randomly assigned to one of three conditions: empathic, non-empathic and noise. We found three main results. First, in all conditions correlation is maximum at a delay near zero. Second, the magnitude of the peak correlation near zero delay is higher for empathic condition (r=0.2059), followed by non-empathic (r=0.1892), with noise condition displaying the lower value (r=0.1779). Third, noise curve distinctively displays local peaks at around -1.5 and 1.5 second delays. This suggests that in this condition delayed bodily reactions to gestures are more present than in the other two conditions.

Index Terms: interpersonal coordination, zero-lag coordination, imitation, empathy

1. Introduction

It is well known that phenomena of spontaneous bodily coordination happen at different observation levels (Hurley & Chuter, 2005). There is robust evidence that interpersonal coordination promotes positive emotions, such as rapport and liking, among interactants (Batson, 2009). Coordination plays an important role in creating and maintaining joint actions. Kirschner and Tomasello (2009) postulated a specifically human motivation to coordinate with social others, which might be characterized as the human "*desire to move in synchrony*" (Kirschner & Tomasello, 2009, p. 32). However, we still do not have strong evidence of the association between interpersonal coordination phenomena and empathic disposition from natural or ecological settings. Pursuing the goal to study interpersonal coordination from a more ecological point of view -that is, to study the whole person in a real interacting situation (Schmidt, Nie, Franco, & Richardson, 2014; Musa, Carré, & Cornejo, 2015), we conducted a study on interpersonal coordination using a MoCap system. Our main hypothesis that interpersonal coordination plays an essential role in maintaining the affective mood of an ongoing conversations, so that it should be more evident in empathic rather than in non-empathic encounters. Additionally, if interpersonal coordination is helping to follow a conversational rhythm, we hypothesized that the amount of coordination should increase under conditions that impair the verbal communication -such as, a conversation occurring along with background noise.

2. Method

2.1. Participants

A total of 20 female and 16 male university students (ages 18 to 28) were randomly matched in 18 couples. Each couple was randomly assigned to one of three conditions: empathic, non-empathic and noise, which differ as explained below in the procedure section. Special care was put in making sure that the two participants in a couple did not know each other.

2.2. Materials

Two backless chairs were used so that participants could sit face to face with their backs remaining visible to measurement equipment. Questions included ice-breakers and casual conversation topics. Questions were self-paced by the participants through two sets of 11 cards -one for each interactant- with the questions printed on them. Ninth question requested telling first person experiences during Chilean earthquake of 2010 and was designed to produce a stronger affective engagement. Motion of interacting dyads during this question was analyzed for this study.

Motion during conversations was recorded using a motion capture system consisting of 18 OptiTrack V100:R2 cameras manufactured by NaturalPoint (Corvallis, Oregon, USA). Arena software provided by the same manufacturer allowed us to reconstruct 3D motion afterwards and export it for further analysis with custom software. In order for this system to work, each subject had to wear 15 little spherical reflective markers. One experimental condition required the use of computer speakers (stereo 20W Edifier brand with 4 inch mid/woofer), and a computer to reproduce a noise composed from the superposition of several speech sources, making it sound unintelligible like in a loud cocktail party.

2.3. Procedure

Each participant was given a brief description of the experiment and signed an informed consent document. After that, 15 reflective markers were attached to each interactant by a same sex assistant by means of elastic bands. Marker localizations were hands, elbows, feet, knees, plus three markers in a fixed arrangement attached to the back of the head, and four markers on the back (see Figure 1). This worked well with diverse clothes without requiring the use of a special suit and allowed a short setup time, all of which contributed to keep interactions reasonably natural.

Only after that participants sat together while a member of the research team gave instructions without leaving space for talking before the experiment started. In brief, each question had to be answered by the two participants before moving on to the next one, alternating who answered each question first. There were no restrictions on how they could move, and they didn't have to keep track of time. However, it was requested that turns proceeded in a strict way if possible, and that the conversation was kept on topic with an estimated duration of 20 to 40 minutes for the whole session as a reference. Motion data was captured with the motion capture system during the conversation.

Previous description corresponds to an *empathic* condition, under the assumption that empathy is the default expected disposition in this situation. A *non-empathic* condition included a manipulation that consisted in an additional instruction. Participants were told that question cards could include some text below commanding to give a fake answer. They were told that some decks did not have this command, so it was a matter of luck if they got a “lier” deck. In actuality, no deck had that command. But this consign introduced the possibility of lying into the conversation without participants truly lying (because it was part of the game and was never commanded anyway), or researchers lying to them (since actual game was within explicitly stated possibilities). We expected this to make participants reluctant to share intimate stories in an authentically empathic disposition. In fact, participants reported believing that the other may have lied.

Finally, a *noise* condition was similar to empathic condition with the only difference that noise was produced through speakers so to make it harder to hear each other and see whether and how gestures compensated this.

2.4. Data analysis

After using manufacturer-supplied software for capturing motion and exporting data to a standard format (C3D), custom software was used to visually label the body part corresponding to each marker and checking for potential problems. We designed an analysis procedure with the following goals in mind. First, it had to avoid subjective segmentation or categorization of gestures. Second, it had to detect similarities between motion events of the two participants even if they occurred with a difference in time of a few seconds. Third, it had to detect those similarities even if they occurred between different parts of the body, or involved different directions in space.

The first goal was met by making an automatic analysis of the continuous motion signals without segmentation, except for the fact that only motion during conversation about the ninth question was considered. The second goal lead us to compute cross-correlation curves. These show an immediate Pearson correlation coefficient that informs about similar events occurring at the same time, but also show the same coefficient for each possible delay in time between potentially similar events in a range that goes up to a few seconds. Delay times are shown in the horizontal axis of our cross-correlation plot, and can be negative or positive, because similar events can occur with one or the other participant producing the first event.

The third goal was met by taking the 45 motion variables of each subject (15 markers, each with an x, y, z position), and performing a principal component analysis (PCA) based on the correlation matrix, similar to a factor analysis. This process linearly extracts maximum variance axes. The result was then rotated with the varimax algorithm. Each resulting dimension was cross-correlated between the two participants, and absolute value of all resulting curves were averaged together, which can be shown to be equivalent to the cross correlation of PCA transformed vectors, using vector dot product instead of the usual deviation product of the Pearson correlation for scalar series. In other words, this correlation measure will tend to be bigger when the principal axes of variance deviate from average to the same direction in both subjects. And since this is after PCA, this *same* direction is referenced to the particular directions and body parts in which each individual shows more motion, so they don't need to be the same directions in the original 3D space.

2.5. Results

We found three main results. First, in all conditions correlation is maximum at a delay near zero. At equipment's temporal resolution of 100 frames per second, the average cross-correlation curves displayed on Figure 2 peak at exactly 0.00 seconds for noise and non-empathic conditions, and at -0.01 for empathic. The curve shape and similarity between the three conditions indicates that this is highly unlikely to have occurred by chance.

Second, the magnitude of the peak correlation near zero delay is higher for empathic condition (r=0.2059), followed by non-empathic (r=0.1892), with noise condition displaying the lower value (r=0.1779). Additional work is needed in order to find the statistical significance of this pattern. Preliminary Montecarlo resampling suggested statistical significance of the difference between peak correlations for empathic and noise conditions.

Third, there is a relevant qualitative shape difference between noise curve and the other two. It displays local peaks at around -1.5 and 1.5 second delays, and several other peaks at delays of bigger magnitude. This suggests that in this condition delayed bodily reactions to gestures are more present than in the other two conditions.

3. Discussion

One of our results strongly suggest that natural conversations display an immediate coordination between participants with a delay much lower than the smallest possible human reaction time: coordination lag seems to be no more than 10 milliseconds in our study, in contrast to 100+ milliseconds reaction time of well trained athletes. This means that our finding, if replicated, cannot be explained as a reaction. In principle, *tracking* the motion of another person with such a tight timing would require knowing the future beforehand. But only if this phenomenon is actually seen as tracking, which would be a sort of precognitive imitation. We think there is no need to view results that way. Actually, many physical phenomena start out of phase, but soon display a coupling that produces an immediately coordinated pattern. The well known phenomenon of sympathetic resonance is a good example, in which an object responds to the vibrations of a nearby object. There is no reason for excluding something as complex as human interaction from the possibility of such patterns, and our study strongly suggests that this is indeed the case regarding bodily coordination. This adds to recent evidence of tightly timed coordination (Paxton & Dale, 2013;

Schmidt, Morr, Fitzpatrick, & Richardson, 2012; Schmidt, Fitzpatrick, Caron, & Mergeche, 2011) found with techniques that involve varying degrees of subjectivity while labeling video sequences of human interaction.

As expected, non-empathic interaction displays a lower amount of immediate coordination than empathic when using motion correlation as a measure. This result, however, requires more statistical work, and ideally replication in order to discard that it can be attributed to chance. On the other hand, we found more support for the hypothesis that the noise condition involves less immediate coordination than the empathic one. This is compensated by the former showing more delayed coordination, at different time lags, remarkably at 1.5 seconds. This is a consistent result if we consider that the noise condition generates a handicap for verbal communication, so that bodily resources should be focused on compensating for that by emphasizing gesture in order to explain, and also in order to acknowledge, which are events that need to occur with a delay. Nevertheless, while lower in magnitude, immediate correlation does not disappear in the noise condition. In fact it is quite high when viewed in the context of the whole cross-correlation curve.

Our finding of clear immediate Pearson correlations of magnitudes around 0.2 is remarkable if we consider that immediate coordination information should be buried below all other complex motion patterns and relationships that occur in human interaction. This suggests that the method of correlating PCA transformed motion data is a useful tool for bodily coordination research.

4. Acknowledgements

This article was supported by the Chilean National Fund for Scientific and Technological Development, FONDECYT (grant number 1141136).

5. Bibliography

Batson, C. (2009). These things called empathy: Eight related but distinct phenomena. In J. Decety & W. Ickes (Eds.), *The social neuroscience of Empathy* (pp. 3–15). Cambridge, MA: MIT Press.

Bernieri, F. J., Reznick, J. S., & Rosenthal, R. (1988). Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother–infant interactions. *Journal of Personality and Social - Psychology, 54*(2), 243–253.

Hurley, S. & Chater, N. (Eds.) (2005). *Perspectives on imitation: From neuroscience to social science, Vol. 1.* Cambridge, MA: MIT Press.

Kirschner, S. & Tomasello, M. (2009). Joint drumming: Social context facilitates synchronization in preschool children. *Journal of Experimental Child Psychology, 102*(3), 299–314.

Musa, R., Carré, D., & Cornejo, C. (2015). Bodily synchronization and ecological validity: a relevant concern for nonlinear dynamical systems theory. *Frontiers in Human Neuroscience, 9*, 64. doi: 10.3389/fnhum.2015.00064

Paxton, A., & Dale, R. (2013). Frame-differencing methods for measuring bodily synchrony in conversation. *Behavior Research Methods*, *45*(2), 329-343. doi: 10.3758/s13428-012-0249-2

Schmidt, R. C., Fitzpatrick, P., Caron, R., & Mergeche, J. (2011). Understanding social motor coordination. *Human Movement Science, 30*(5), 834-845. doi: 10.1016/j.humov.2010.05.014

Schmidt, R. C., Morr, S., Fitzpatrick, P., & Richardson, M. J. (2012). Measuring the dynamics of interactional synchrony. *Journal of Nonverbal Behavior*, *36*(4), 263-279. doi: 10.1007/s10919-012-0138-5

Schmidt, R. C., Nie, L., Franco, A., and Richardson, M. J. (2014). Bodily synchronization underlying joke telling. *Frontiers in Human Neuroscience, 8,* 633. doi:10.3389/fnhum.2014.00633

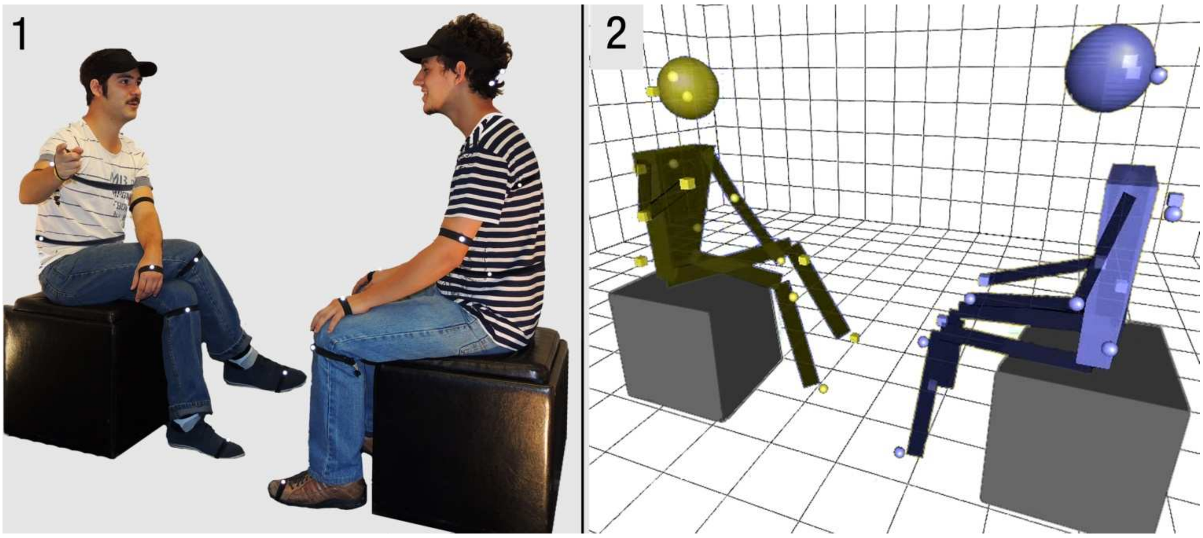


Figure 1. Participants with little reflective markers attached to their bodies by means of elastic bands, and 3D reconstruction.

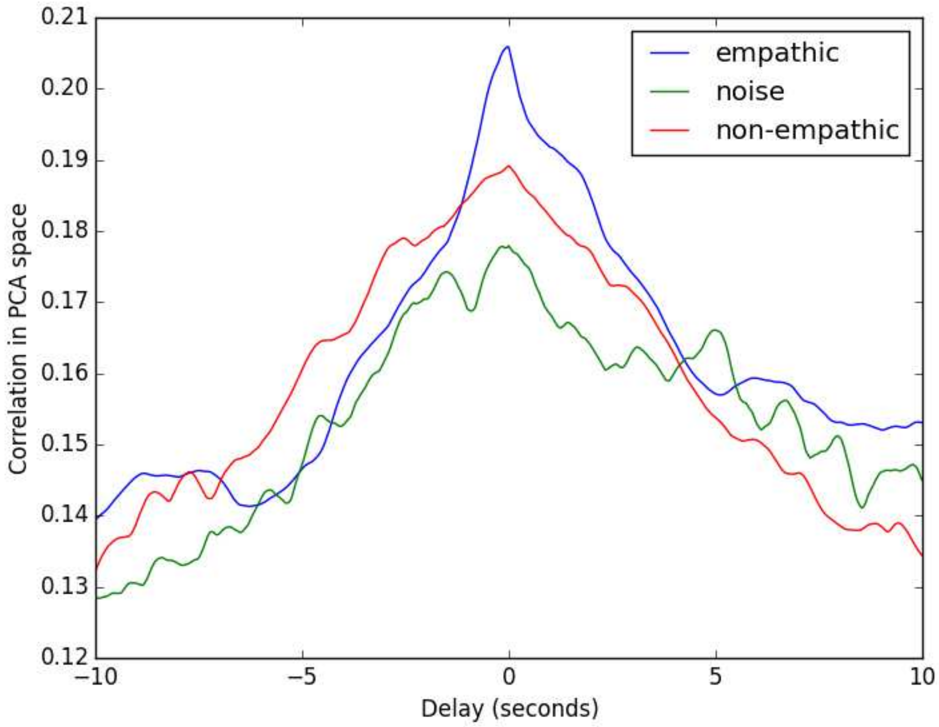


Figure 2. Average cross-correlation curve between the PCA transformed motion of the two participants, computed for each condition. All three curves display a zero-lag (i.e., immediate) motion coordination between participants, but with different magnitudes in each case.

Prosody and gesture in dialogue: Cross-modal interactions

Agnieszka Czoska, Katarzyna Klessa, Maciej Karpiński, Ewa Jarmołowicz-Nowikow

Institute of Linguistics
Adam Mickiewicz University in Poznań, Poland

agaczoska@gmail.com, klessa@amu.edu.pl, maciejk@amu.edu.pl, ewa@jarmolowicz.art.pl

Abstract

In this paper, some measures of cross-modal interactions are proposed and implemented in the analysis of a multimodal corpus of task-oriented dialogues. The corpus includes multi-level annotations of speakers' verbal and gestural behaviour, e.g., hand gestures, gaze direction, utterance content or intonational phrasing. A moving time-window approach is adopted to analyse changes in the communicative behaviour of dialogue participants over time. The study is focused on how gestures and speech of the Instruction Giver influence the speech of the Instruction Follower in the course of dialogue.

Index Terms: cross-modal interaction, rate of events, gesture, gaze direction, entrainment

1. Communicative entrainment in dialogue

Interaction between dialogue participants means not only building an overt structure of dialogue turn exchanges such as questions and responses or statements and acknowledgements. Conversational parties exert mutual influence on each other that starts on the shallow, behavioural level and goes up to the level of mental representations. Various aspects of behaviour of dialogue partners are mutually adjusted but these externally observable changes influence and are influenced by deeper processes of mental representation alignment. Early ideas of mutual accommodation in communication were formulated by Giles [e.g., 1, 2]. They certainly inspired more recent influential works on interactive alignment by Pickering, Garrod, and others [3, 4]. Many of these interactive processes seem to be subtle, susceptible to factors normally occurring in conversation, and relatively difficult to track in empirical studies. Nevertheless, their importance is rarely questioned as they have been shown to predict success or failure of communication [e.g., 5, 6, 7]. They may also influence other cognitive processes like speech perception and phonological processing [8].

A relatively large number of studies have been devoted to lexical [e.g., 9, 10] and prosodic alignment [e.g., 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. Although the involvement of body motion and gesture in these processes has gained a significant amount of attention only recently, certain influential ideas can be found already in the early works of Kendon [24].

The notions of "alignment", "convergence","entrainment", "co-ordination" or "synchrony" are defined from different viewpoints and with different purposes in mind [e.g., 25, 26] but the phenomena they refer to remain strongly interrelated. In their proposal of Interactional Phonology, Wagner et al. [8] postulate that alignment may percolate across various levels of representation. In our research, in turn, we hypothesize that the mutual influences between dialogue parties are not necessarily held within a single sensory modality. For example, gesture rate in one speaker may influence speech rate in the other or vice-versa. Such processes of entrainment seem to be more difficult to explain but may be of equal importance and strength as the intra-modal ones.

In the present contribution, cross-modal interactions are brought to focus with the use of DiaGest2 multimodal corpus as speech material, and the SRMA plugin as an implementation of moving time-windows approach (Section 2). The results of multiple regression used for analysis of interaction are discussed in Section 3 and concluded in Section 4 of the paper.

2. Cross-modal interactions in task oriented dialogues

2.1. DiaGest2 multimodal corpus

The study is based on DiaGest2 multimodal corpus of task-oriented dialogues using "origami" task [27]. The task itself is reconstructing a paper figure visible only to one of the participants (instruction giver, IG) and invisible to the other (instruction follower, IF). The latter is provided with all the necessary materials and then instructed by the IG how to build the figure. The corpus consists of ten sessions recorded in the mutual visibility condition and ten in the limited mutual visibility condition. Only the former are analysed in the present study. Participants were female and male students of philology. There was an equal number of females and males in the role of IG while IFs were always females.

DiaGest2 corpus includes transcriptions and annotations of a variety of phenomena both on the level of speech and gestures. Speech was segmented into intonational macro- and micro-phrases, transcribed orthographically and phonetically for both IGs and IFs. For IG, gestures were annotated with a varying level of detail, depending on the gesture category. For all IG speakers, however, gaze shifts, the boundaries of gesture phrase as well as changes of hand movement direction within gesture phrases were tagged [27].

As the original DiaGest2 project was focused on the multimodal realisation of selected categories of dialogue acts that are much more frequently produced by IGs than IFs, some behaviour (e.g., gesticulation) was annotated only for IGs.

2.2. Assumptions and methods

A previous analysis of interactions based on the annotations of DiaGest2 corpus was performed using simple univariate correlation measures between speech rates and nPVI [28] measurements in dialogue utterances [22]. This choice was made for the sake of comparisons between the results obtained from DiaGest2 and from another corpus including only telephone conversations with no eye contact nor gesture annotation [29]. Because of that it was decided to use only the cues available for both groups of speakers, i.e. those included in the speech signal. Since DiaGest2 multimodal corpus includes also dialogues recorded with eye contact between interlocutors, it provides appropriate material for analyses of both speech and gestures. In the present study, we thus extend the scope of interest towards potential cross-modal interactions based on both verbal and non-verbal cues.

In order to observe the phenomena related to the rate and (co-)occurrence of speech and gesture events we used moving time-windows approach according to which a selected parameter is measured and averaged within a fixed-size time window moving along the time axis. A similar method was employed by Kousidis et al. [13, 15] who additionally used a weighted mean, where the interval durations were the weights. We applied the SRMA (Segment Rate Moving Average) plugin developed for Annotation Pro [22, 31]. With this plugin, the size of the moving window as well as the overlap between subsequent time-windows can be adjusted by the user. Different values may be required depending on, e.g., the overall rate or number of segments and silence intervals within an annotation tier which may, in turn, differ significantly depending on the level of annotation detail or communicative channel and modality.

Since speech and gesture tend to synchronise in one speaker [31, 32, 33, 34, 35, 8], using only simple regression and correlation for multimodal data may result in numerous artefacts (e. g. finding spurious correlations variables A and B which in fact is dependant on another variable, C, with which B is intercorrelated). In order to analyse the present data, multiple regression is used which is expected to provide better control over all the independent variables, accounting also for their cross-correlations [36]. With multiple regression it is possible to include all the independent variables (e.g. the number of syllables, gestures and movements preformed by one participant) in one equation and measure their influence on the dependent variable. Thus, for each time-window several measurements of one participant's behaviour were used to predict the single measure of communicative behaviour (mainly speech rate) of the second speaker in the same window.

In the abovementioned earlier analysis of convergence in DiaGest2 corpus [22], measurements for separate dialogues were in focus, while in the present study a grand average approach is used. Accordingly, all the dialogues are treated as a single set of data and the results of multiple regression may be interpreted more generally.

The main goal of the present study was to investigate the potential interactions between communicative behaviour of dialogue participants, and the possible association between IG's gestures and gaze shifts with IF's gaze shifts and speech. The main hypotheses were:

1. IG speech rate (number of syllables or micro-phrases per time-window) will correlate with IF speech rate;

2. IG gesture rate (number of gestures or movements per time-window) will correlate with IF gaze shifts rate (number of focus changes).

2.3. Data extraction

Five different time-window size settings were used with the SRMA plugin (Section 2.2), i.e. 20, 30, 40, 50 and 60 seconds, with two overlap settings (33% of overlap for window sizes of 20-40, and 75% for sizes of 50-60). Increased overlap size allowed for collecting more data points. The analysis of the narrowest time-windows (20 and 30 seconds) was interpreted as referring to the local entrainment, while the widest ones (50 and 60 s) – to the global one.

For each of the time-window sizes, the values based on the number of the following events were calculated:

- information giver (IG) left and right hand gestures (gesture phrases) (G-PhraseLH and G-PhraseRH);
- IG left and right hand movements performed during the phrases (movRH and movLH);
- IG and information follower (IF) gaze shifts (GazeIF and GazeIG; gaze was annotated in terms of object viewed with 3 distinct categories: OnPartner, OnFigure and OnElse, and thus gaze shifts may be interpreted as focus changes);
- IG and IF intonational micro-phrases and syllables (MiIP_IG, MiIP_IF, SylIG and SylIF).

To explain which independent variables (features of communicational behaviour of IG) affect dependent variables (features of communicational behaviour of IF: GazeIF, SyllIF and MiIP-IF), multiple regression analyses were performed. All the variables describing the behaviour of IG were taken into account: speech (SyllIG and MiIP-IG), gestures (G-PhraseLH, G-PhraseRH, movRH and movLH) and gaze shifts (GazeIG). The effect sizes reported below are measured with R^2 for the whole multiple regression results and with semipartial *R* when separate independent variables are analysed. Semipartial *R* was chosen because it shows correlation coefficient between a given independent variable and the dependent variable that remains while influence of all the other independent variables is controlled; in other words it may be seen as an extraction of the single independent variable from the multiple regression equation.

The analyses were performed initially for all the data as one dataset. However, DiaGest2 dialogues can be divided into two distinct groups comprising: female–female (FF) and mixed, female–male (MF) dialogues. There are examples in the literature indicating that the gender of participants may influence the nature of entrainment [37]. In order to test that hypothesis, multiple regression analyses for FF and MF dialogues separately were also conducted.

3. Interaction analysis results

The statistical analyses described further in this section were conducted with IBM SPSS Statistics for Windows. Measures of the ten variables (3 for IF and 7 for IG) were extracted for each time-window. The cases with more than six empty data-points, i.e. no annotation within a given time-window, were excluded from further analyses.

3.1.1. Multiple regression analysis results (grand average)

Detailed results of multiple regression analysis are presented in Table 1 while Figures 1 – 3 illustrate selected global and local tendencies.

IF's gaze shift frequency (see Figure 1) appears to be affected by IG's speech in the local time-scale (windows of 20-50 seconds size) and by IG's gaze shifts in more global scale (the correlation emerges in a 40 s time-window). Contrary to our expectations, semi-partial correlation with the gestures of IG was found in only one time-window (50 s), which may be caused by its high correlation with IG's speech rate. The shift from negative to positive correlation between GazeIF and GazeIG is probably an outcome of the change of data resolution (from about 15 data points per dialogue to about 10), because of which the individual time-windows showing negative association were averaged with adjacent time points showing positive correlation.

The syllable rate in IG influences IF's speech rate (the number of syllables per time-window, see Figure 2). The shift from negative to positive correlation between 20 and 30 s time-windows may be caused by the change of data resolution: from about 30 time-points per dialogue to about 20. A weak negative correlation was also observed between the number of syllables produced by IF and the number of IG's left hand gestures. All the aforementioned correlations are local. No single independent variable crucial for between-speakers coordination emerges across the global time-windows, although the communicational behaviour of IG taken as whole correlates significantly with IF's speech rate measured with both measures (syllables and micro-phrases). For the speech of IF in general, the effect size increases slightly with the increase in the size of the time-window (see Table 1).

The number of IF's intonational micro-phrases (see Figure 3) in a given time-window appears to be affected mainly by IG's speech measured as the number of syllables.

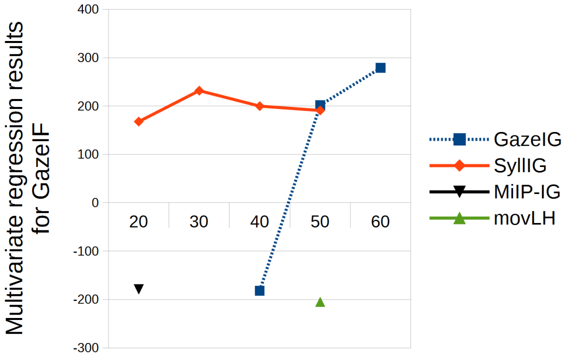


Figure 1: *Independent variables correlated with GazeIF; the points represent values of correlation coefficient (semipartial R) found significant in each time window.*

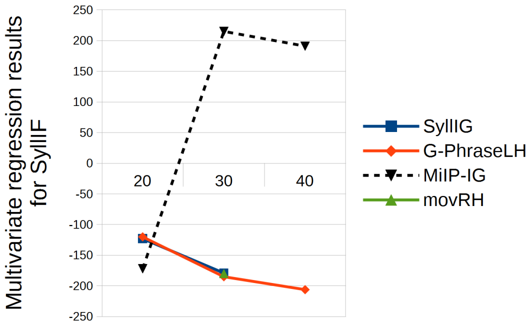


Figure 2: *Independent variables correlated with SyllIF; the points represent values of correlation coefficient (semipartial R) found significant in each time window.*

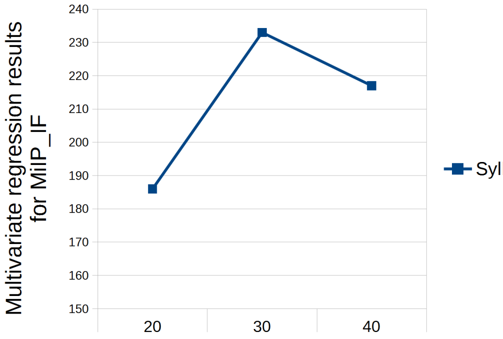


Figure 3: *Independent variables correlated with MiIP_IF; the points represent values of correlation coefficient (semipartial R) found significant in each time window.*

Table 1. Results of multivariate regression analysis: general results (first row for each dependent variable, R^2 as effect size) and independent variables (second row; semipartial R (RSem) as effect size) that emerge as significantly influencing each dependent variable in different time-window sizes when all the other variables are controlled for. All the reported R measures are significant with p<0.005.

	20 s time-window	30 s time-window	40 s time-window	50 s time-window	60 s time-window
GazeIF	$R^2 = 0.179$	$R^2 = 0.269$	$R^2 = 0.292$	$R^2 = 0.318$	$R^2 = 0.414$

	SyllIG: RSem=0.168;	**SyllIG:** RSem=0.232;
		MiIP-IG: RSem=-0.179;
			SyllIG: RSem=0.2;
			GazeIG: RSem=-0.182;
				SyllIG: RSem=0.191;
				GazeIG: RSem=-0.202;
				movLH: RSem=-0.205;
					GazeIG: RSem=-0.279;

SyllIF	$R^2 = 0.07$	$R^2 = 0.102$	$R^2 = 0.144$	$R^2 = 0.15$	$R^2 = 0.162$

	MiIP-IG: RSem=-0.172;
	SyllIG: RSem=-0.123;
	G-PhraseLH: RSem=-0.12;
		MiIP-IG: RSem=0.215;
		SyllIG: RSem=-0.179;
		movRH: RSem=-0.18;
		G-PhraseLH: RSem=-0.185;
			MiIP-IG: RSem=0.191;
			G-PhraseLH: RSem=-0.206;
				no significant semipartial correlations	no significant semipartial correlations

MiIP-IF	$R^2 = 0.062$	$R^2 = 0.114$	$R^2 = 0.196$	$R^2 = 0.217$	$R^2 = 0.212$
	MiIP-IG: RSem=0.186;	**MiIP-IG:** RSem=0.233;	**MiIP-IG:** RSem=0.217;	no significant semipartial correlations	no significant semipartial correlations

3.1.2. Multiple regression analysis results by groups

Detailed results of multiple regression analysis by groups are presented in Table 2 and (to illustrate global and local tendencies) in Figures 4-7.

Evidence for significant associations between IG's and IF's behaviours in both groups was found only for GazeIF. There is a stable (with a global tendency in FF group) medium correlation between IF gaze shifts and IG communicational behaviour. However, the structure of co-ordination is different in the two groups: the rate of syllables per time-window in IG is crucial in the FF group (see Figure 4), whereas in the MF group none of the independent variables remains in significant correlation with GazeIF through adjacent time-windows, only GazeIG seems to be important locally (in time-windows of 20 and 40 seconds size; see figure 5).

There is also evidence for weak and unstable local association between GazeIF and IG gestures. The lack of global association between gaze shift rates in the MF dialogues may be the effect of the lack of coordination, but may also be caused by the fact that in one of the MF dialogues a reverse association (a negative one) between dialogue participants' gaze shift rates was observed.

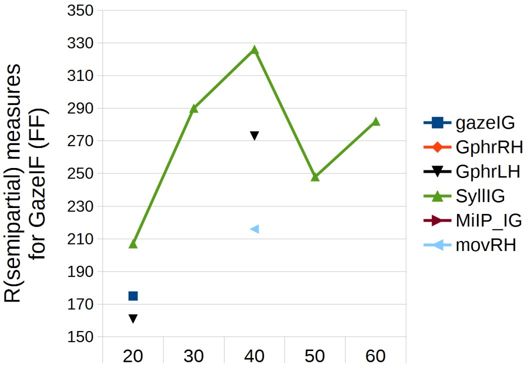


Figure 4: *Independent variables correlated with GazeIF (FF dialogues); the points represent values of correlation coefficient (semipartial R) found significant in each time window.*

*Table 2. Results of multivariate regression analysis in FF (female-female dialogues) and MF (male-female dialogues) groups. General results for each dependent variable in each group are presented in the first rows and independent variables that emerge as significantly influencing dependent variables in different time window sizes are presented in second rows (with semipartial R (Rsem) as effect size). All the reported R measures are significant with p<0.005 or p<0.006 (indicated with *).*

	20 s time-window	30 s time-window	40 s time-window	50 s time-window	60 s time-window
GazeIF	FF: $R^2 = 0.303$; MF: $R^2 = 0.288$;	FF: $R^2 = 0.335$; MF: $R^2 = 0.371$;	FF: $R^2 = 0.39$; MF: $R^2 = 0.31$;	FF: $R^2 = 0.386$; MF: $R^2 = 0.277$;	FF: $R^2 = 0.498$; MF: $R^2 = 0.442$;
	FF: GazeIG (RSem=0.175) GphrLH (RSem=0.161) SyllIG (RSem=0.207)	FF: SyllIG (RSem=0.29)	FF: GphrLH (RSem=0.273) SyllIG (RSem=0.326)	FF: SyllIG* (RSem=0.248)	FF: SyllIG (RSem=0.282) movRH (RSem=0.216)
	MF: GazeIG (RSem=0.359) GphrLH (RSem=0.191) GphrRH (RSem= -0.159)	MF: no significant semipartal correlations	MF: GazeIG (RSem=0.271)	MF: no significant semipartal correlations	MF: no significant semipartal correlations
SyllIF	FF: $R^2 = 0.151$; MF: no significant correlation	FF: $R^2 = 0.188$; MF: no significant correlation	FF: $R^2 = 0.282$; MF: no significant correlation	FF: $R^2 = 0.328$; MF: no significant correlation	FF: $R^2 = 0.503$; MF: no significant correlation
	FF: MiIP_IF (RSem=0.233) GphrLH (RSem= -0.238)	FF: MiIP_IF (RSem=0.321)	FF: MiIP_IF (RSem=0.273)	FF: MiIP_IF (RSem=0.313)	FF: movLH (RSem= -0.279)
MiIP-IF	FF: $R^2 = 0.184$; MF: no significant correlation	FF: $R^2 = 0.273$; MF: $R^2 = 0.287$;	FF: $R^2 = 0.318$; MF: no significant correlation	FF: $R^2 = 0.43$; MF: no significant correlation	FF: $R^2 = 0.507$; MF: no significant correlation
	FF: MiIP_IG (RSem=0.29)	FF: MiIP_IG (RSem=0.39)	FF: MiIP_IG (RSem=0.446) GazeIG (RSem= -0.289)	FF: MiIP_IG (RSem=0.353)	FF: movLH (RSem=-0.29) movRH* (RSem=0.259) GazeIG (RSem= -0.312)
		MF: GphrRH (RSem=0.487) GphrLH (RSem= -0.487)			

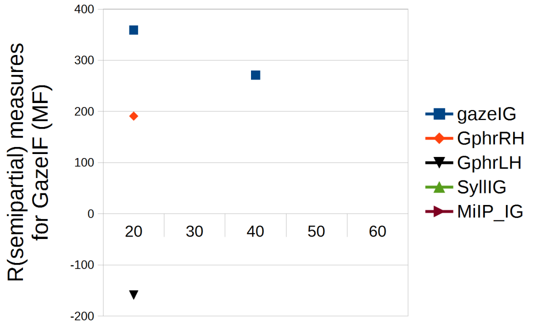


Figure 5: *Independent variables correlated with GazeIF (MF dialogues); the points represent values of correlation coefficient (semipartial R) found significant in each time window.*

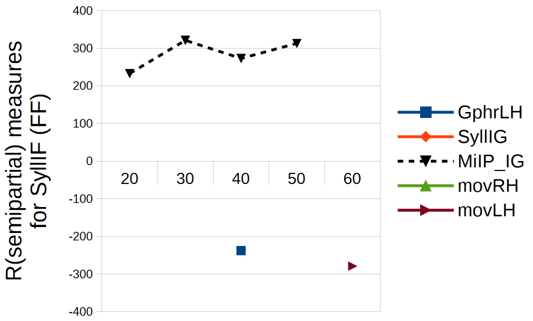


Figure 6: *Independent variables correlated with SyllIF (FF dialogues) ; the points represent values of correlation coefficient (semipartial R) found significant in each time window.*

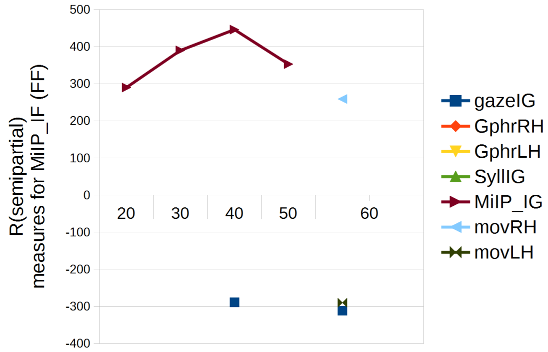


Figure 7: *Independent variables correlated with MiIP_IF (FF dialogues); the points represent values of correlation coefficient (semipartial R) found significant in each time window.*

Evidence for entrainment between IG's and IF's speech measures was found only in the FF group. As when all the dialogues were analysed together (see Figures 2 and 3), MiIP_IG seems to be the most important independent variable. The lack of significant association between IG's other behaviour and IF's speech may be caused by social variables (gender and the degree of familiarity), but may also be caused by the fact, that in one of the MF dialogues, the average syllable and micro-phrase rate is lower in IG, than IF.

4. Conclusion and further work

Verbal and paraverbal behaviour tend to be highly correlated within one speaker in terms of frequency or rhythm. Consequently, the results of simple correlations between those measures and behaviour of another speaker may be spurious or epiphenomenal. That was the reason why multiple regression analysis was applied in this study.

In the short task-oriented dialogues requiring instruction and error monitoring certain evidence of multimodal between-speaker coordination was found, especially between IF gaze shifts and IG speech, and (although only a local one) between IF speech and IG gestures. Since it was IG who played the leading role of dialogue manager, usually speaking more and first, it might be expected that it was rather IF adjusting her behaviour to IG. Though the strongest correlations were found between IF and IG speech rate measures, we also observed quite stable correlations between IF gaze shifts and IG behaviour as well as an indication of a weak correlation between IG left-hand gestures and IF speech rate.

Contrary to our expectations, IF gaze shifts correlated rather with IG speech than gestures. This (together with the fact that IF speech correlated with IG speech as well) may mean that speech is the main indicator of rhythm to which dialogue participants may entrain. However, gestures also emerged as important for entrainment, especially when IF syllable rate was measured. What is more, in FF group there was some indication of local correlation between IG gestures and IF gaze shifts.

When analysed separately, FF and MF dialogues seem to be very different in terms of multimodal entrainment. What is most interesting, is that more significant and more stable correlations were found in the dialogues in the FF (uni-gender) group than in the mixed one. The difference is most visible in speech co-ordination, since significant regression measures for verbal behaviour were found only in FF group. Co-ordination seems to be rather global in both cases, since the effect size tends to rise with the increase of the time-window size. The difference may be motivated by the gender of the participants, with female IF in mixed groups less eager to adjust their behaviour to IG's.

While discussing the differences between the FF and MF dialogues in DiaGest2 corpus, a potential distracting factor should be mentioned related to the fact that the pairs of interlocutors might have differed in terms of the degree to which they knew each other at the time of the recording session. During the recording the data was not gathered methodically, but it may be the case that within FF dialogues there were more pairs who were more familiar to each other or were friends. Friends might be expected to exhibit different communication strategies, including different degree of multimodal entrainment than distant colleagues or strangers (cf. also the considerations in [38] as regards interactional convergence in storytelling by friends and good colleagues recorded for [39]). Further investigation in this direction might also provide a broader perspective also for the studies of gender-dependent relationships.

Results of the present analyses may also suggest that there is relatively little multimodal entrainment based on gesticulation between dialogue partners. However, this observation should be treated as a tentative one due to fact that so far we looked only for linear correlations while the character of such association may be different and can vary through interaction. Moreover, one should take into account the characteristics of task-oriented dialogues we explored where participants often were loosing eye contact and sometimes paused verbal interaction in order to work on the manual part of the task.

5. References

[1] Giles, H., Taylor, D. M., and Bourhis, R. Y. “Towards a theory of interpersonal accommodation through language: Some Canadian data,” Language in Society 2, 1973, pp. 177-192.

[2] Giles, H., and Smith, P. "Accommodation theory: Optimal levels of convergence," In H. Giles and R. N. St. Clair (Eds.), Language and Social Psychology. Baltimore: University Park Press, 1979, pp. 45-65.

[3] Pickering, M. J. and Garrod, S. "Toward a mechanistic psychology of dialogue," Behavioral and Brain Sciences, 27, 2004, pp. 169–226.

[4] Garrod, S. & Doherty, G. 1994. Conversation, co-ordination and convention: and empirical investigation of how groups establish linguistic conventions, Cognition, 53, pp. 181–215.

[5] Porzel, R., Scheffler, A. and Malaka, R. "How entrainment increases dialogical efficiency," Proceedings of Workshop on Effective Multimodal Dialogue Interfaces, Sydney 2006.

[6] Reitter, D. and Moore, J. D. Predicting success in dialogue," Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 808–815.

[7] Ramseyer, F. and Tschaecher, W. "Nonverbal synchrony or random coincidence? How to tell the difference," In A. Esposito (Ed.) COST 2102 International Training School, Heidelberg: Springer Verlag, 2009, pp. 182–196.

[8] Wagner, P., Malisz, Z., Inden, B., Wachsmuth, I. "Interaction phonology – a temporal co-ordination component enabling representational alignment within a model of communication," In: Wachsmuth, I., de Ruiter, J., Jaecks, P., Kopp, S. (Eds.) Alignment in Communication: Towards a New Theory of Communication. Amsterdam: Benjamins, 2013.

[9] Garrod, S., and Anderson, A. "Saying what you mean in dialogue: A study in conceptual and semantic co-ordination," Cognition, 27(2), 1987, pp. 181-218.

[10] Brennan, S. E. and Clark, H. H. "Conceptual pacts and lexical choice in conversation," Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 1996, pp. 1482–93.

[11] Street, R. L. "Speech convergence and evaluation in fact-finding interviews". Human Communication Research, 11(2), 1994, pp. 139–169.

[12] Street, R. L., Brady, R. M., and Putman, W. B. "The influence of speech rate stereotypes and rate similarity or listeners' evaluations of speakers," Journal of Language and Social Psychology, 2(1), 1983, pp. 37-56.

[13] Kousidis, S., Dorran, D., Wang, Y., B. Vaughan, Cullen, C., Campbell, D., McDonnell, C. and E. Coyle, "Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues," Proceedings of Interspeech 2008, pp. 1692–1695.

[14] Edlund, J., Heldner, M. and J. Hirschberg 2009. "Pause and gap length in face-to-face interaction," Proceedings of Interspeech 2009, pp. 2779–2782.

[15] Kousidis, S. "A Study of Accomodation of Prosodic and Temporal Features in Spoken Dialogues in View of Speech Technology Applications," Doctoral Thesis. Dublin Institute of Technology, 2010.

[16] Levitan, R., Hirschberg, J. 2011. "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," Proceedings of Interspeech 2011, Florence, Italy, August 2011, pp. 3081–3084.

[17] Looze, de, C. and Rauzy, S. 2011. "Measuring speakers' similarity in speech by means of prosodic cues: methods and potential," Proceedings of Interspeech 2011, Florence, Italy, August 2011, pp. 1393–1396.

[18] Vaughan, B. 2011. "Prosodic synchrony in co-operative task-based dialogues: A measure of agreement and disagreement," Proceedings of Interspeech 2011, Florence, Italy, August 2011, pp. 1865–1867.

[19] Truong, K. P. and Heylen, D. 2012. "Measuring prosodic alignment in cooperative task-based conversations," Proceedings of Interspeech 2012, 9-13 September 2012, Portland, OR, USA.

[20] Włodarczak, M., Simko, J. and Wagner, P. 2012. "Syllable boundary effect: Temporal entrainment in overlapped speech," Proceedings of Speech Prosody 2012, Shanghai, China, May 2012.

[21] Wagner, P. Malisz, Z. Kopp, S. "Gesture and speech in interaction: an overview." Speech Communication, 57, 2014, pp. 209-232.

[22] Karpiński, M., Klessa, K., Czoska, A. "Local and global convergence in the temporal domain in Polish task-oriented dialogue," Proceedings of the 7th Speech Prosody Conference, 20-23 May 2014, Dublin, Ireland. ISSN: 2333-2042.

[23] Gravano, A. Benus, S., Levitan, R., Hirschberg, J. "Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement," Spoken Language Technology Workshop (SLT), IEEE, 2014, pp. 578-583.

[24] Kendon, A. "Movement coordination in social interaction". Acta Psychologica, 32:1-25, 1970.

[25] Levitan, R., Hirschberg, J. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions, Proceedings of Interspeech 2011, Florence, Italy, August 2011, pp. 3081–3084.

[26] De Looze, C., Scherer, S., Vaughan, B., & Campbell, N., "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction." Speech Communication, 58, 11-34, 2014.

[27] Jarmołowicz-Nowikow, E. and Karpiński, M. "Communicative intentions behind pointing gestures in task-oriented dialogues," In P. Wagner, Z. Malisz, C. Kirchhof (Eds.) Proceedings of GESPIN 2011: Gesture and Speech in Interaction Conference, 2011.

[28] Low, E. L., Grabe, E., Nolan F. "Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English," Language and Speech 43(4), 2001, 377-401.

[29] Klessa, K., Wagner, A., Oleśkowicz-Popiel, M., Karpiński, M. "Paralingua – a new speech corpus for the studies of paralinguistic features," In Vargas-Sierra, Ch. (Ed.) Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions. Procedia – Social and Behavioral Science. Vol. 95, 2013. pp. 48-58.

[30] Klessa, K. "Annotation Pro," [Software tool]. Version 2.2.1.5. Retrieved from: http://annotationpro.org/ on 2015-04-02.

[31] McClave, E. "Pitch and manual gestures," Journal of Psycholinguistic Research, 27(1), 1998 , pp. 69-89.

[32] McNeill, D. Hand and Mind: What Gestures Reveal about Thought.Chicago: University of Chicago Press, 1995.

[33] Cummins, F. Port, R. F. "Rhythmic commonalities between hand gestures and speech. In Proceedings of the eighteenth meeting of the Cognitive" Science Society. 1996. pp. 415-419.

[34] Cummins, F. "Rhythm as entrainment: The case of synchronous speech." Journal of Phonetics, 37(1), 2009, pp. 16-28.

[35] Wachsmuth, I. "Communicative rhythm in gesture and speech. "In Gesture-based communication in human-computer interaction., 1999. pp. 277-289.

[36] Hayes, A. F. Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford Press, 2013.

[37] Levitan, R., Gravano, A., Willson, L., Benus, S., Hirschberg, J., and Nenkova, A. "Acoustic-prosodic entrainment and social behavior," Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies. Association for Computational Linguistics, 2012.

[38] Guardiola, M., Bertrand, R.. "Interactional convergence in conversational storytelling: when reported speech is a cue of alignment and/or affiliation," Frontiers in Psychology 4, 2013.

[39] Bertrand R., Blache P., Espesser R., Ferré G., Meunier C., Priego-Valverde B., et al. "Le CID—Corpus of Interactional Data—Annotation et Exploitation Multimodale de Parole Conversationnelle," Traitement Automatique des Langues 49, 2008, pp. 105–134.

Which gesture types make a difference?

Interpretation of semantic content communicated by PWA via different gesture types

Carola de Beer[1]*, Marcella Carragher*[2]*, Karin van Nispen*[3]*, Jan P. de Ruiter*[1]*, Katharina Hogrefe*[4] *& Miranda L. Rose*[2, 5]

[1] Faculty of Linguistics and Literature Science, Bielefeld University, Bielefeld, Germany
[2] Department of Community and Clinical Allied Health, La Trobe University, Melbourne, Australia
[3] Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, The Netherlands
[4] Clinical Neuropsychology Research Group (EKN), Clinic of Neuropsychology, Bogenhausen Hospital, Munich, Germany
[5] Centre for Clinical Research Excellence in Aphasia Rehabilitation (CCRE), Australia

carola.de_beer@uni-bielefeld.de, M.Carragher@latrobe.edu.au, K.vanNispen@uvt.nl, jan.deruiter@uni-bielefeld.de, katharina.hogrefe@ekn-muenchen.de, M.Rose@latrobe.edu.au

Abstract

People with aphasia (PWA) spontaneously use various gesture types. Such gestures can potentially express semantic content that complements speech.

We investigated whether production of different gesture types adds crucial semantic content to the spoken output produced by PWA. In a perception experiment using multiple choice questions, naïve judges reported their information uptake from messages communicated by PWA in a speech-only vs. gesture+speech condition. The results show that the choice of response-options differed between conditions for all tested gesture types. We conclude that gestures in PWA disambiguate the interpretation of communicated messages and therefore markedly influence the expression of semantic content.

Index Terms: gesture, aphasia, spontaneous communication, semantic content

1. Introduction

The relationship between gesture and speech is assumed to vary between different gesture types. Kendon [1] distinguishes between gesticulation, pantomimes, emblems and sign language. These gesture types show different characteristics in terms of their relationship to speech, their degree of conventionalization and their linguistic properties. Gesticulations are not conventionalized, only appear with speech and have no linguistic properties. In contrast, emblems and pantomimes are conventionalized to a certain degree and hold some linguistic properties. Therefore, the latter two gesture types hold the potential to be understood without accompanying speech, whilst the interpretation of gesticulations is closely related to the accompanying speech.

The role of gestures in the expression of semantic content has been investigated in a number of studies. One line of enquiry relates to whether the content expressed via gesture is redundant to the accompanying speech or complementary. Some researchers argue that iconic gestures do not play an important role in the communication of relevant information [e.g. 2]. This assumption is based on the finding that participants' interpretation of semantic content was not improved with the accessibility of visual information compared to only audio information. In contrast, Bangerter [3] as well as Melinger and Levelt [4] report that spatial information is completely omitted from spoken output in the presence of deictic or iconic gestures in target-identification tasks. Furthermore in narratives, [5] parts of the informational content expressed via gesture was not inferable from the content of the spoken output.

The coordination and link between gesture and speech can be conceptualised by the planning and production processes underlying each. Non-parallel expression of content in gesture and speech can be accounted for by models of gesture production that assume a shared origin of gesture and speech and tightly coordinated but separate production processes of the two channels, for example the *Sketch Model* [6, 7]. Parts of a speaker's communicative intention can be conveyed via gesture and do not necessarily have to be specified in speech as well. This is especially evident in people with impaired spoken output, as is the case in PWA [8, 7]. However there is evidence against this compensative or *trade-off relationship* of gesture and speech in non-impaired speakers [9]. Regarding people with aphasia some researchers were able to demonstrate a spontaneous and compensative use of gestures that is especially true for those individuals presenting with severe aphasia [10, 11]. But this potential compensative role of gesture for PWA has been debated, with evidence against an effective compensative use of gestures [12]. Furthermore, there is evidence that both gesture and speech are vulnerable to simultaneous break down in PWA [13]. These findings clearly call into question the view that gesture plays a compensatory role in the case of aphasia.

Whilst acknowledging the lack of consensus regarding the role of gesture in communication, it is widely accepted that PWA make use of various gesture types in spontaneous communication [e.g. 14, 15]. Amongst many other gesture types, Sekine and colleagues [15] identified emblems, pantomimes and referential gestures as frequently used by PWA in spontaneous communication. Whilst we know that PWA with different aphasic types and severities make spontaneous use of a variety of gestures in communication, previous studies have not investigated the content expressed via gesture. Furthermore, it cannot be inferred from previously reported evidence what information listeners were able to comprehend when gesture, speech or both channels were accessible.

Hogrefe et al. [11] investigated the comprehensibility of cartoon-narratives produced by PWA based on the responses of naïve judges. The PWA recalled the cartoon narratives in two conditions: 1) they were asked to retell the cartoons they watched without any specific instructions (speech+gesture condition) and 2) they were explicitly asked to retell the cartoons only by the use of gestures (gesture only condition). Judges' information uptake from the first condition was compared between gesture and speech. The reactions to the audio stimuli were more accurate for 8 (out of 16) PWA. For 2 of the 16 PWA, judges' reactions to the gesture stimuli were more accurate. Judges' reactions to the gesture stimuli from the first condition (speech+gesture) were also compared to the gesture stimuli from the second condition (gesture only). The judges' responses were more accurate for 8 PWA in the second condition. In summary, speech was more informative than speech+gestures in most PWA. However, for some PWA their speech-replacing gestures (gesture only) were more informative than their speech-accompanying gestures (speech+gesture).

In an additional analysis, Hogrefe et al. [16] evaluated the information content that six judges identified from the speech vs. gesture (speech+gesture condition) stimuli used by PWA. The judges were presented with choices from a list of predefined content-related propositions and asked to identify which propositions they were able to recognize from the stimuli. For 5 of the 16 PWA, more propositions were correctly detected from the gestures. Similarly, for 5 of the 16 PWA, more propositions were correctly detected from the speech by the judges. A subsequent analysis per proposition was carried out to investigate if there were a) any cases in which no information was understood from either of the communication channels, b) propositions were recognized from both modalities (redundant), c) propositions were solely recognized from gesture, and d) propositions were solely recognized from speech. The redundant score did not significantly differ from the gesture-only score for the whole group. For individuals presenting with severe aphasia, more propositions were shown to be conveyed solely by gesture. These results suggest that individuals with severe aphasia produce gestures to compensate for their reduced verbal output. However, whilst Hogrefe et al. [16] considered the effects of all gestures used in the narrative, they did not distinguish between different gesture types and their respective influence on the judges' perception.

Rose and colleagues [17] tested the comprehensibility of pantomimes produced by PWA. The data were extracted from spontaneous conversations and presented in a) audio+video b) audio only and c) video only. Seventy-four student participants answered open-ended questions (OQ) and multiple-choice questions (MCQ). The combined audio+video stimuli led to the most accurate responses to both the OQ and MCQ.

In a follow-up study by De Beer et al. [18], the impact of gestures on the communicative effectiveness in PWA was investigated. The accuracy of information uptake from messages communicated by PWA was studied for three different gesture types; referential gestures, emblems and pantomimes. Clips from conversation samples of PWA were presented in a gesture+speech condition or a speech-only condition. Participants answered OQ and MCQ and their responses were scored. Participants' responses were more accurate in the gesture+speech condition for all tested gesture types for both OQ and MCQ. The choice of the MCQ options was compared between conditions: analysis indicated that participants’ responses differed significantly between the two conditions. In other words, the participants’ perception of information content differed between the gesture+speech and speech-only conditions. However, the choice of response options was not tested for each of the specific gesture types. Hence it is not possible to infer from the data if all three different gesture types (pantomimes, emblems and referential gestures) express information that differs from verbal speech to a different extent.

The present study represents a follow-up analysis of participants’ choice of response-options from the multiple-choice questionnaire for pantomimes (as defined by Kendon [1]), emblems (as defined by Kendon [1]) and referential gestures (reflecting what Kendon [1] named gesticulations and subsuming McNeill's [19] deictic and iconic gestures). We compared participants' responses between two different presentation conditions 1) gesture+speech (G+S) and 2) speech-only (S-O). The analysis aimed to further differentiate various gesture types and their respective effects on listeners’ uptake of messages produced by PWA.

2. Method

A subsequent analysis was conducted using data collected in a perception experiment. In the original study, we tested participants' reactions to 30 stimulus clips taken from spontaneous conversation samples of PWA [18].

2.1. Participants

10 participants with aphasia were chosen from the AphasiaBank Database (http://www.talkbank.org/ Aphasia Bank). They presented with primarily productive deficits and varying degrees of severity of aphasia (for details on the participants, see De Beer et al. [18]).

60 student participants were recruited as naïve judges for the study. The participants were blinded to the aims of the study.

2.2. Material

a) Video and Audio Stimuli

The clips for the experiment were chosen from conversational samples of the AphasiaBank Database. These clips are recordings of PWA reporting their stroke story and also an important event of their lives. For each PWA, one clip per gesture type was chosen (i.e., pantomimes, emblems and referential gestures). An exception to this was Subject 2, who did not produce any pantomimes in the samples. To ensure an equal number of clips per gesture type, two clips with pantomime gestures were chosen from the conversation sample of Subject 4. This yielded a total of 30 clips containing the gestures of interest. For each of the 30 clips, an audio and a video version were created. The chosen clips were of varying lengths (2 to 10 seconds) due to differing complexities of the communicated messages. Gesture classification was conducted by the first author. The classification for the 30 gestures was checked by a second blinded rater who was familiar with the categorisation system used. Agreement between the two raters was reached for 83.3% of all cases. Cohen’s kappa for inter-rater reliability was acceptable at .75.

b) Multiple Choice Questions

MCQ were constructed to identify the information that the judges understood from the clips. The four multiple choice options included:

1) gesture+speech (G+S) message, i.e., the target message based on the information from the video and the audio versions of the clips;

2) G+S distractor which was semantically related to the G+S message;

3) speech-only (S-O) message, i.e., a message solely based on the information from the audio versions of the clips;

4) S-O distractor which was semantically and phonetically related to the S-O message.

The transcript of one of the stimulus clips (clip 20) is presented below. Table 1 displays the four constructed response options for clip 20.

The four response options were generated by two of the authors. For the construction of the S-O messages, one rater listened to the audio versions of the clips without knowing the video versions.

Example for one stimulus clip: Transcript of the target gesture and the accompanying speech for Clip 20.

S: and one le uh left

H: *left hand in front of the body, palm turned upwards (preparation)*

[/1.5/]

H: *pantomime: left hand and arm on chest height, hand is oriented downwards, circular movement above the table, imitates sprinkling something on top of a round object (target gesture)*

S: [and decorate] cakes an'

S	spoken output
H	hand movements (in italics)
/	silent pause (duration in seconds reported in brackets)
[]	stroke of gesture

Table 1. Overview of the messages and distractors for Clip 20

1) G+S message	I was decorating cakes left-handed
2) G+S distractor	I was baking cakes
3) S-O-message	When they left I was decorating cakes
4) S-O-distractor	I was decorating the house and baking a cake after they left

2.3. Procedure

Participants were randomly assigned to one of the two experimental groups. In group 1 (n=30) clips 1 - 15 represented the audio or S-O version and clips 16 - 30 represented the video or G+S version. For group 2 (n=30) the presentation modes were reversed. In the experimental sessions all participants started with the S-O condition to avoid any unwanted effects of order of condition. Each clip was presented twice before participants were asked to report what they understood from the clips by answering to one OQ per clip and the subsequent MCQ (for more information about the OQ see De Beer et al. [18]).

Participants recorded their responses in a response booklet in written form. For the MCQ, participants were asked to choose the option they felt best matched the message the PWA in the respective clip was trying to communicate. Gestures were not mentioned in the instructions or any of the written forms. The number of choices of each option was counted per clip and per condition.

Analysis

Clip number 4 was removed from the analysis because of poor sound quality. The gesture type presented in clip 4 was an emblem. Thus for the category of emblems only 9 clips were included in the final data analysis.

Two-tailed Wilcoxon Sign Ranked Test for related samples was used for the statistical analysis.

3. Results

a) Referential Gestures

For the category of referential gestures, the G+S message was chosen significantly more often (*Z* = -2.549, *p* = .011) in the G+S condition (*mean* = 21.6, *SD* = 6.931) compared to the S-O condition (*mean* = 10.6, *SD* = 8.249). The G+S distractor was chosen more often in the G+S condition (*mean* = 3.6, *SD* = 4.671) compared to the S-O condition (*mean* = 2.8, *SD* = 4.686), but this difference did not reach statistical significance (*Z* = -.06, *p* = .952). The S-O message was chosen more often in the S-O condition (*mean* = 8.90 , *SD* = 5.744) than in the G+S condition (*mean* = 2.2, *SD* = 3.155). This difference was significant (*Z* = -2.553, *p* = .011). Also the S-O distractor was picked significantly more often (Z = -2.492, p = .013) in the S-O condition (*mean* = 7.7, *SD* = 7.273) compared to the G+S condition (*mean* = 2.6, *SD* = 2.989). See Figure 1.

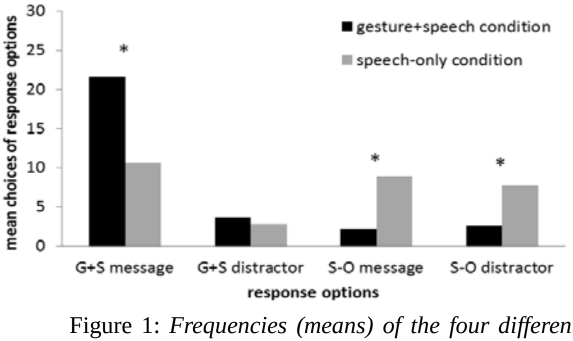


Figure 1: *Frequencies (means) of the four different choices of response options for referential gestures compared between the gesture + speech condition (black) and the speech-only condition (grey). Significant differences are indicated by asterisks.*

b) Emblems

For the category of emblems, participants chose the G+S message more often in the G+S condition (*mean* = 16.56, *SD* = 10.43). This difference to the S-O condition (*mean* = 9.11, *SD* = 7.132) reached statistical significance (Z = -2.556, p = .011). The difference for the G+S distractor between the G+S condition (*mean* = 3.22, *SD* = 5.426) and the S-O condition (*mean* = 3.56, *SD* = 5.615) was not significant (*Z* = -.632, *p* = .527). Participants' choices of the S-O message differed significantly between conditions (*Z* = -2,075, *p* = .038) and it was more often chosen in the S-O condition (*mean* = 11.22, *SD* = 7.513) compared to the G+S condition (*mean* = 6.33, *SD* = 8.602). Participants chose the S-O distractor significantly more often (*Z* = -2.2, *p* = .028) in the S-O condition (*mean* = 6, *SD* = 7.826) compared to the G+S condition (*mean* = 4, *SD* = 7.632). See Figure 2.

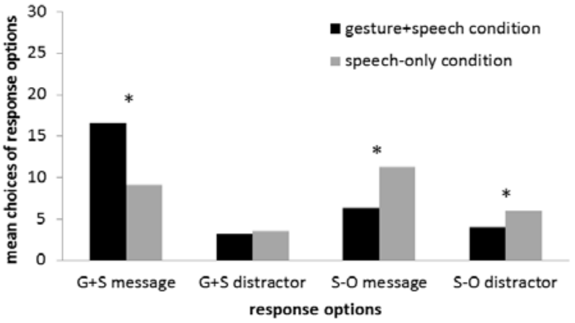


Figure 2: *Frequencies (means) of the four different choices of response options for emblems compared between the gesture + speech condition (black) and the speech-only condition (grey). Significant differences are indicated by asterisks.*

c) Pantomimes

For the category of pantomime gestures, the G+S message was chosen more often in the G+S condition (*mean* = 20, *SD* = 8) compared to the S-O condition (*mean* = 11.7, *SD* = 9.638). This difference was statistically significant (*Z* = -2.67, *p* = .008). No significant difference (*Z* = -.768, *p* = .443) was found for the choice of the G+S distractor between the G+S condition (*mean* = 2.2, *SD* = 3.736) and the S-O condition (*mean* = 3.5, *SD* = 5.642). The S-O message was chosen more often in the S-O condition (*mean* = 10.6, *SD* = 8.884) compared to the G+S condition (*mean* = 6.5, *SD* = 5.421). This difference did not reach statistical significance (*Z* = -1.899, *p* = .058). Participants' choices of the S-O distractor differed significantly between conditions (*Z* = -2.536, *p* = .011). It was chosen more often in the S-O condition (*mean* = 4.3, *SD* = 3.622) compared to the G+S condition (*mean* = 1.2, *SD* = 1.135). See Figure 3.

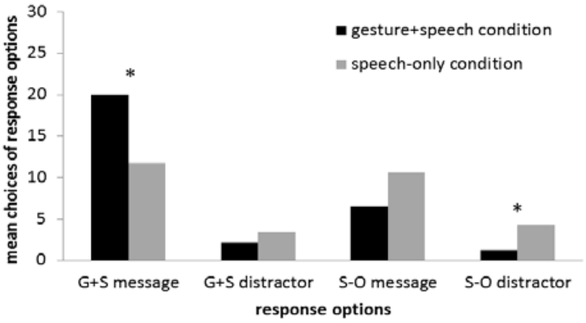


Figure 3: *Frequencies (means) of the four different choices of response options for pantomimes compared between the gesture + speech condition (black) and the speech-only condition (grey). Significant differences are indicated by asterisks.*

4. Discussion

In summary, the participants' choices of response options in the MCQ differed between conditions for all three gesture types. The G+S message and the S-O message were chosen more often in their respective conditions. These effects were significant, apart from the number of choices of the S-O message for pantomime gestures. For the G+S distractors no remarkable effects of condition were found for either of the three gesture types. The S-O distractor was chosen significantly more often in the S-O condition for all three gesture types.

The number of choices of the response options indicates overall that the participants did pay attention to the type of gesture that the PWA produced in the clips and that the information expressed via all gestures was used for the interpretation of the messages. This supports earlier findings by De Beer et al. [18].

In the G+S condition, participants demonstrated a clear preference for the G+S message (the target message); this was true for all three gesture types. However, in the S-O condition, participants did not choose the S-O message with a similar frequency. Participants' choices of the response options were less stable in the S-O condition; here, the target message was chosen with a similar frequency as the S-O message for all three gesture types. A remarkable number of participants in the S-O condition still chose the target message which is not surprising, because for many clips most of the semantic content was expressed in speech. The presentation of the MCQ options might have influenced participants' interpretation of the messages. Particularly in the S-O condition, when participants did not have access to the complete informational content (i.e., information conveyed via gesture), the presentation of the target message might have led to reinterpretation of the audio-stimuli. Combining these assumptions together with the effects of condition, it can be inferred that the accessibility of the information from the gesture channel decreased the ambiguity of the communicated messages in the stimuli. Therefore in the G+S condition when participants had access to the information from both modalities, they were able to identify the target message with higher accuracy.

Strikingly the G+S distractors were rarely chosen in both conditions across gesture types. There were no clear effects of condition found for this distractor. This finding may be due to the construction of the distractors, because the G+S distractor was only semantically related to the G+S message and not always phonetically related to the information presented in verbal speech. Hence the G+S distractors may not have been sufficiently closely related to the target messages.

The effects of condition were shown for all three gesture types. This indicates that all tested gesture types did influence the participants' information uptake. By their nature, pantomimes and emblems hold the potential to convey content that complements or even replaces spoken output. Referential gestures are assumed to be more tightly related to spoken output and only completely interpretable in the context of the accompanying speech. Surprisingly, within this study, the effect of gesture on participants' interpretation of semantic content was not limited to pantomimes and emblems; participants showed similar effects for all three gesture types on information uptake, though one would expect stronger effects of gestures that can replace speech in the case of impaired production of speech. For at least some PWA, gesture might necessarily be used to replace speech in the event of severely compromised spoken output. It is crucial to mention that some content is still expressed in speech by PWA in most cases. One-word utterances as well as sentences interrupted by unsuccessful word retrieval still serve as a source for semantic content for listeners. Gestures produced in spontaneous conversation can be interpreted in the context of even very reduced speech production. Within this study, all tested gesture types played a significant role in the expression of semantic content. This semantic content can complement spoken output, but it is still interpreted in the context of spoken production. The findings of the current study support our earlier conclusions [18] and serve to further our understanding of the impact of different gesture types on the expression of semantic content in PWA. Therefore, we were able to contribute to the evidence suggesting a compensative use of gestures in PWA, i.e. argue against the assumption that gesture and speech break down in parallel in PWA.

We acknowledge that the choice of stimulus clips might have influenced the results of the study. This would be true if only sequences were chosen in which gestures were used in a speech-replacing way. However we included clips of sequences in which gestures were complementary but also redundant to the spoken output. Thus the stimuli were chosen to reflect varying degrees of complement or redundancy. Future studies might wish to consider constructing the target messages and distractors on the basis of independent judges' interpretation of the audio and video stimuli to improve validity. We also acknowledge the use of short messages in a perception study has been criticised by Beattie & Shovelton [5], who argued that the information expressed via gestures is often inferable from the wider context of a narrative. In the present study we used parts out of spontaneous conversation samples. Whilst it is plausible that contextual information influenced judges' perception of messages, we took care not to choose any clips that could only be interpreted with context knowledge of the whole conversation. Finally, the work of Hogrefe et al. [16], who investigated the information uptake from narrations produced by PWA, also suggests that in some individuals with aphasia gestures are more informative than speech.

5. Conclusion

All three gesture types under investigation (pantomimes, emblems and referential gestures) influence the interpretation of the messages communicated by PWA. Gestures produced by PWA are used by listeners to disambiguate messages from spoken output. Gestures do not necessarily have to be used in a speech-replacing way by PWA to play a role in the expression of semantic content. Therefore, communication in PWA has to be viewed as a multi-modal process. Gesture types which differ in the degrees of conventionalisation and relation to speech have been demonstrated to hold the potential of expressing semantic content. This was true even for gestures that are closely related to spoken output (referential gestures). Our results clearly suggest a compensative use of different gesture types and broaden the knowledge about their role for communication for PWA.

6. Acknowledgements

The first author of this study was funded by a short term-PhD-scholarship of the DAAD (German Academic Exchange Service). Further acknowledgements go to Kaziku Sekine and Annett Jorschick for supporting the statistical analysis, to Abby Foster and Lucy Knox for their support in the preparatory phase of the experiment and to the lecturers of the School of Allied Health at La Trobe University who helped with participant recruitment.

7. References

[1] Kendon, A., "Gesture: Visible Action as Utterance", Cambridge: University Press, 2004.

[2] Krauss, R., Dushay, R.A., Chen, Y., & Rauscher, F., "The communicative value of conversational hand gestures", J Exp Soc Psychol, 31(6):533–552, 1995.

[3] Bangerter, A., "Using pointing and describing to achieve joint focus of attention in dialogue", Psychol Sci, 15(6):415–419, 2004.

[4] Melinger, A., & Levelt, W. J. M. (2004), "Gesture and the communicative intention of the speaker", Gesture, 4(2):119–141, 2004.

[5] Beattie, G., & Shovelton, H., "An exploration of the other side of semantic communication: How the spontaneous movements

of the human hand add crucial meaning to narrative", Semiotica, 184(1-4):33–51, 2011.

[6] De Ruiter, J. P., "The production of gesture and speech", In D. McNeill [Ed], Language and Gesture, 284–311, Cambridge: Cambridge University Press, 2000.

[7] De Ruiter, J. P., & De Beer, C., "A critical evaluation of models of gesture and speech production for understanding gesture in aphasia", Aphasiology, 27(9):1015–1030, 2013.

[8] De Ruiter, J. P., "Can gesticulation help aphasic people speak, or rather, communicate?", Int J Speech Lang Pathol, 8(2):124–127, 2006.

[9] De Ruiter, J. P., Bangerter, A., & Dings, P., "The interplay between gesture and speech in the production of referring expression: Investigating the Tradeoff Hypothesis", Top Cogn Sci, 4(2):232–248, 2012.

[10] Goodwin, C., "Gesture, aphasia and interaction", in D. McNeill [Ed], Language and Gesture, 84–98, Cambridge: Cambridge University Press, 2000.

[11] Hogrefe, K., Ziegler, W., Wiesmayer, S., Weidinger, N., & Goldenberg, G., "The actual and potential use of gestures for communication in aphasia", Aphasiology, 27(9):1070–1089, 2013.

[12] Cicone, M., Wapner, W., Foldi, N., Zurif, E., & Gardner, H., "The relationship between language and gesture in aphasic communication." Brain Lang, 8(3):324–349, 1979.

[13] Duffy, R. J., & Duffy, J. R., "Three studies of deficits in pantomimic expression and pantomimic recognition in aphasia." J Speech Lang Hear Res, 24(1):70–84, 1981.

[14] Carlomagno, S., & Cristilli, C., "Semantic attributes of iconic gestures in fluent and non-fluent aphasic adults", Brain Lang, 99(1-2):102–103, 2006.

[15] Sekine, K., Rose, M. L., Foster, A. M., Attard, M. C., & Lanyon, L. E., "Gesture production patterns in aphasic discourse: In-depth description and preliminary predictions", Aphasiology, 27(9):1031–1049, 2013.

[16] Hogrefe, K., Ziegler, W., Weidinger, N., & Goldenberg, G., "Gestural expression in narrations of aphasic speakers: redundant or complementary to the spoken expression?", Proceedings of the Tilburg Gesture Research Meeting (TIGER), Netherlands, 2013.

[17] Rose, M.L., Mok, Z., Katthagen, S., & Sekine, K., "The communicative effectiveness of pantomime gesture in people with aphasia", Aphasiology, in prep.

[18] De Beer, C., Carragher, M., Van Nispen, K., De Ruiter, J.P, Hogrefe, K., & Rose, M. L., "How much information do people with aphasia convey via gesture?", Am J of Speech Lang Pathol, under revision.

[19] McNeill, D., "Hand and Mind", Chicago: University Press, 1992.

Stance-taking functions of multimodal constructed dialogue during spoken interaction

Camille Debras[1]

[1] Department of English Studies, University Paris Ouest Nanterre La Défense, Nanterre, France
cdebras@u-paris10.fr

Abstract

Based on qualitative analyses of spontaneous interactions between native speakers of British English, this paper argues that speakers' use of multimodal enactment during constructed dialogue can be motivated by stance-taking processes. Speakers use multimodal enactment (i.e. change in voice pitch, pantomime) when *dis(s-)tancing* themselves from a stance attributed to an absent subject. When *endorsing* an absent subject's stance, they don't use multimodal enactment, thereby iconically representing the outside stance as their own. Theoretically, this study re-evaluates Du Bois's (2007) Stance Triangle as a Stance Tetrad: speakers simultaneously position themselves with respect to an object and *both* present *and absent* subjects.

Index Terms: multimodality, enactment, stance-taking, constructed dialogue, interaction

1. Introduction

As remarked by Tannen [1], reported speech is an inaccurate term to describe direct discourse attributed to another source than the speaker here and now. There is no point, she explains (among others), in assessing the truthfulness of the representation of speech by direct discourse: the original discourse can usually not be accessed and the direct discourse is nothing but a production of the speaker here and now. When using direct speech, the speaker is not so much representing somebody's speech as presenting discourse in the form of "constructed dialogue" [1]. Constructed dialogue can have a much larger range of pragmatic functions than just referring to speech, just as direct speech can be used to characterize non-speaking entities, like objects, through fictive interaction [2].

Direct speech, which dissociates the speaker's voice from his responsibility, has been studied from a rich variety of approaches. Goffman's [3] sociolinguistic description of institutional speech distinguishes *author* (the source of the speech), *animator* (the person who is voicing the speech) and *principal* (the entity that is responsible for the speech). A large body of research in French enunciative linguistics has accounted for such polyphony with a distinction between *locutor* (the speaking voice), and *enunciator* (the origin of the speech), which can be distinct from the speaker himself ([4], [5], [6] among others). A locutor's utterance can hence contain multiple enunciators. This distinction has a lot in common with Martin & White's appraisal theory (anchored in systemic-functional linguistics) [7]. In Martin & White's Bakhtinian approach, a speaker's utterance always exists against a backdrop of other possible utterances on the same theme. Since "whenever speakers (or writers) say anything, they encode their point of view towards it" ([8]: 197), any utterance makes a speaker agrees or disagrees with the explicit or potential perspectives of present interlocutors and/or absent parties. From this perspective, direct speech is only a case where the dialogical nature of all discourse is made explicit.

In the course of spoken interaction, constructed dialogue can be supplemented by non-verbal components, such as a change in voice pitch and/or coordinated body movements: constructed dialogue can turn into *multimodal enactment* ([9], [10], [11]). Indeed, if gestures are often used to represent objects, one of the most familiar things to represent with a talking body is another talking body ([12]: 16), in contexts where so-called quotations actually function as multimodal demonstrations [13].

Enactment is a well-documented phenomenon in Sign Languages, under the name of *role shift* ([14] on ASL), *personal transfer* ([15] on LSF) or *constructed action* ([16] on Auslan), but has received less attention in spoken languages (apart from [10], [17] and [18], where McNeill shows how adults and children use enactments differently to express observer and character viewpoint). This paper aims to show that multimodal enactments during constructed dialogue in the course of interaction do not only fulfill representational functions but also stance-taking ones. More specifically, a speaker's use of voice change and bodily enactment can be used as a resource to take a stance simultaneously with respect to present subjects (interlocutors) and absent ones.

Stance has been studied from various approaches in corpus linguistics, and broadly corresponds to "a display of a socially recognized point of view or attitude" ([19]). When speakers take stances, they simultaneously position themselves with respect to a discourse object and an interlocutor: "stance is a public act by a social actor, achieved dialogically through overt communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects, with respect to any salient dimension of the sociocultural field" ([20]: 163). Studying stance is hence fundamentally concerned with how propositional content is always intermingled with the expression of intersubjective, interpersonal relations. More particularly, this paper argues that agreeing or disagreeing with *absent* subjects is another basic dimension of stance-taking.

Based on qualitative analyses of spontaneous interactions between native speakers of British English, I analyze how speakers' use of multimodal enactment during constructed dialogue is motivated by stance-taking processes as follows. (i) Speakers use multimodal enactment (i.e. change in voice pitch and pantomime) when they *distance* themselves from a stance attributed to an absent third party by constructed dialogue. (ii) Speakers don't use multimodal enactment (i.e. there is continuity in their gesturing style and tone of voice) when they *endorse* a stance attributed to an absent third party by constructed dialogue.

2. Corpus and Method

2.1. Corpus

The corpus under scrutiny is a collection of videotaped semi-guided discussions between pairs of friends (2 hours and 20 minutes in total), recorded in Spring 2011. All 16 speakers

(7 male, 9 female) are university students (aged 18-30) who are native speakers of British English. During approximately 15 minutes, the participants pick and discuss questions bearing on environmental issues, a classical topic in applied ethics ([21]) that invite them to take stances, evaluate, and position themselves with respects to norms and knowledge. All participants signed informed consents before participating in the data collection, and are anonymously identified by trigram code names. Speakers sat in the familiar setting of a college supervision room and were free to skip a question if they wished. Recording pairs of friends made the conversation spontaneous and familiar and sitting on chairs did not prevent them from moving and gesturing freely from the waist up. Although using multiple cameras allows for collecting visual information ([22]) these naturalistic conversations were filmed with just one camera, which is less intrusive.

2.2. Method

The chosen approach is founded in multimodal interaction analysis ([23], [24], [25]). For each occurrence of constructed dialogue in the corpus, voice pitch is analyzed in PRAAT ([26]) and the following features are coded in ELAN ([27]):

- Affiliation/ disaffiliation ([28]) with the absent subject: based on the analysis of the sequential context, does the speaker agree or disagree with the absent subject to whom the speech content is attributed?

- Affiliation/ disaffiliation with the interlocutor: based on the analysis of the sequential context, does the speaker agree or disagree with his interlocutor on the topic?

- Vocal features: is there continuity or a marked change in the speaker's voice pitch range?

- Gestures: is there continuity or a marked change in the quality of the speaker's gestures (e.g. in speed or amplitude)?

- Gaze and posture: is there a shift sideways indicating the creation of an imaginary story space and multimodally expressing mixed viewpoints ([29])?

The results presented here focus on the qualitative study of three sequences, which exemplify processes at work in the data as a whole. The analyzed passages are transcribed in intonation units ([30]), in which punctuation reflects intonation, not syntax. Specific gestures are shown after each transcript in screen captures. For the sake of clarity, turns at talk are numbered and instances of constructed dialogue are transcribed in bold.

3. Results

3.1. Endorsing an absent subject's stance

In a previous study, we showed how absent subjects are sometimes quoted as experts to serve as warrants for the speaker's discourse ([31]). But endorsing an absent subject's stance can have other forms and functions. In Excerpt 1, ANT tells his interlocutor ELI a side anecdote from his adolescence, in which a homeless man who smelt bad used to regularly visit his local library. Each time he left, the librarian sprayed the library with an air freshener, perplexing some library users.

Excerpt 1

1 ANT: and er she'd literally just like as soon as he left she kind of like,
right I'm gonna go round with the febreze now,
and everybody else was like,
why why is she like going round with the febreze?
and she was like,
well you know the smelly guy's been round again.
2 ELI: (laughs) oh god,
3 ANT: (laughs) it was quite common knowledge.

In the passage, ANT presents both the library users' incomprehension and the librarian's justification of her action in the form of constructed dialogue. His attribution of direct discourse to the library users (*why is she going round with the febreze?*) and to the librarian (*well you know the smelly guy's been round again*) is probably a reformulation rather than a quote: in the silent environment of a library, they would more likely have expressed their incomprehension by silent visual displays (e.g. raised eyebrows) rather than voiced utterances. Presenting their reaction in the form of direct discourse is hence rather a strategy to stage the anecdote and create humor. There is continuity in ANT's gesturing style and voice pitch as he shifts to the two points of view. In none of the three occurrences of constructed dialogue (ANT, turn 1) does ANT resort to a shift of posture and/or gaze to create a visual story space. Rather, he keeps the same body orientation towards ELI and his gaze fixed on her throughout constructed dialogue, even though he embodies the mentioned absent subjects to some extent. For instance, on *right I'm gonna go round with the febreze now*, he keeps the same gaze and body posture orientations as when he was speaking in his own name, but combines them with an enactment of the librarian spraying febreze in the library (Fig. 1). Enacting an absent subject's actions without explicitly marking the difference between self and other, between real space and story space ([29]), is a way for the speaker to iconically express his endorsement of the absent subject's attitude.

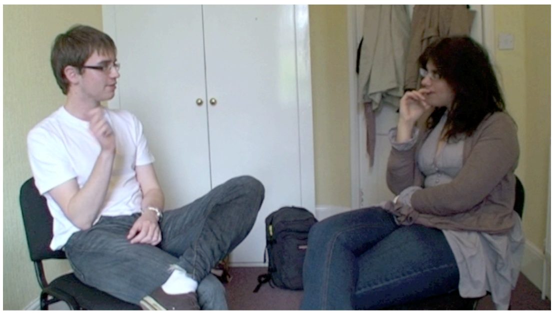


Figure 1: ANT (left)'s enactment of right I'm gonna go round with the febreze now.

Likewise, ANT uses the upper range of his own voice pitch to voice the library users' reaction as questioning an observable state of affairs ([32]), the way he would do to ask a question himself. He also maintains his usual pitch range during both the quotative utterance ([33]) *she was like* and the direct discourse itself *well you know the smelly guy's been round again*, which iconically suggests that he puts himself in the absent subject's shoes when voicing her stance. The visual modality anticipates on the verbal content: the quotative utterance is synchronized with a small palm-up shrug (lifted shoulders, palm-up flip of the left hand, in Fig. 2) expressing shared knowledge ([34]), which is later taken up verbally by the discourse marker *you know* in the utterance attributed to the librarian. The small amplitude of this shrug is in line with the speaker's gesturing style when he speaks in his own name.

Continuity in voice pitch range and gesturing style to represent points of view that originally did not involve speech in the form of constructed dialogue allows the speaker to achieve several effects. Two different points of view (the library users' and the librarian's) on the same event are presented on an equal footing. By lending his own voice and gesturing style to both of them, he endorses each viewpoint in turn. More precisely, his whole talking body is mobilized to lend a voice to each of them, thereby suggesting that he could well have reacted the same way in their place. ELI aligns with ANT, empathetically laughing at the incongruous situation he has just described, and their shared laughter (turns 2 and 3) indicates their aligned stances (i.e. shared perspectives) on the story.

Figure 2: ANT's palm-up shrug on and she was like

3.2. Dis(-s)tancing oneself from an absent subject's stance

In the data, distance with respect to an absent subject's stance presented as constructed dialogue is largely expressed by non-verbal resources. Excerpt 2 is taken from the conversation between SIM and DAN. DAN presents the opinion of geographers about climate change in the form of constructed dialogue. He heard them speak at a debate organized by a geographical society, and was surprised by their position on combating climate change: they argued that a country should develop either renewable energies or nuclear power, while DAN thinks that both should be developed together. SIM joins him in questioning the absent subject's stance and DAN eventually confirms that he rejects it too.

Excerpt 2

1 DAN: like people I went to this debate,
they were a geographical society,
and they were saying,
oh it's it's either one or the other you know,
we can't direct our attention to both.
but I definitely think we can,
2 SIM: really,
why not?
3 DAN: well this is what I didn't understand,
none of them gave a good argument.

In this passage, the main function of direct discourse cannot be truthfully quoting an absent subject: the source of the direct discourse is explicitly identified as a group of people (*they were a geographic society*) during a debate. DAN uses multimodal constructed dialogue to sum up a collective stance on a given topic, and makes extensive use of multimodal resources to enact it. A first striking aspect is the use of a change in voice pitch range in synchrony with the direct discourse attributed to the geographers. On *and they were saying, oh it's it's either one or the other you know we can't direct our attention to both*, DAN uses a markedly low voice pitch (around 100Hz, see Fig. 3) that reaches far lower than his default pitch range. His own voice pitch (around 200 Hz) reappears when he starts speaking in his own name again, on *but I definitely think we can*. This contrast in voice pitch iconically marks the introduction of an outside enunciator, whose voice is perceptually different from his own. He uses his own voice as a medium to present an absent subject's stance while simultaneously reminding his interlocutor that this outside voice is distinct from his. The difference in voice pitch iconically represents the speaker's disaffiliation with the absent subject's stance. The transition from self to other is also marked on the verbal level: the direct discourse opens with the utterance-initial discourse marker *oh*, which usually indicates a change of state for the speaker ([35]). *Oh* is highlighted by a low initial pitch, marking a shift from the speaker's viewpoint to the absent subject's viewpoint.

The visual modality reinforces the speaker's distance: DAN accompanies the constructed dialogue utterance with pantomime including exaggerated head movements and facial displays ([36]), and a shift in posture and eye gaze (Fig. 5). These visual changes are timed with the vocal distanciation and all begin on the quotative utterance (*and they were saying*). In that respect, the non-verbal components slightly anticipate the verbal one. In contrast with his previous physical attitude (Fig. 4), the use of visual markers borders on caricature (Fig. 5), informing the interlocutor that the stance presented by the speaker has nothing to do with what he believes here and now.

Using a markedly lower voice pitch (Fig. 3) adds to the caricature, as it mimics the voice of a phlegmatic old professor. Furthermore, DAN's simultaneous shifts in gaze and trunk posture (Fig. 4) suggest that the rejected stance is positioned in another, abstract dialogue space different from the real dialogue space ([29]) of his conversation with SIM (Fig. 5). In this specific context, creating a virtual dialogue space does not only serve a narrative purpose. Locating the constructed dialogue outside the here and now is another way for the speaker to iconically represent disaffiliation with the absent subject's stance. In all, verbal strategies (direct discourse, *oh*) as well as vocal (marked change in voice pitch range) and visual ones (exaggerated pantomime, gaze sideways) are carefully timed and combined in the sequential unfolding of actions to multimodally construct the rejection of an absent party's stance.

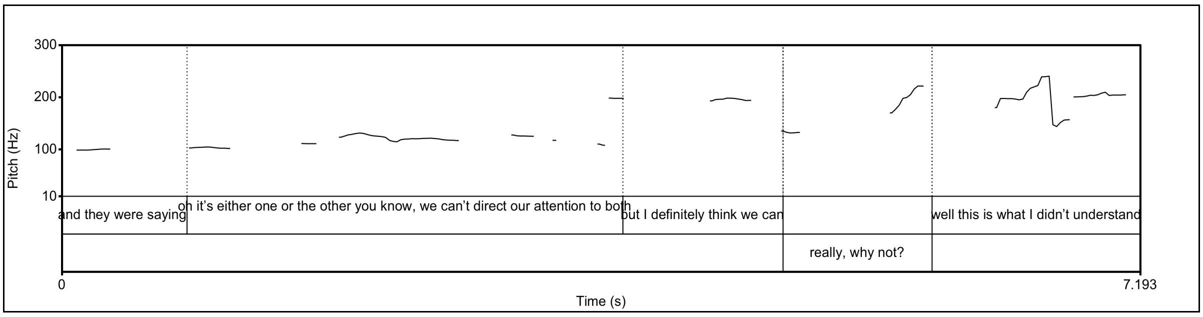


Figure 3: DAN's change in voice pitch between constructed dialogue (and they were saying) *and his own voice* (but I..)

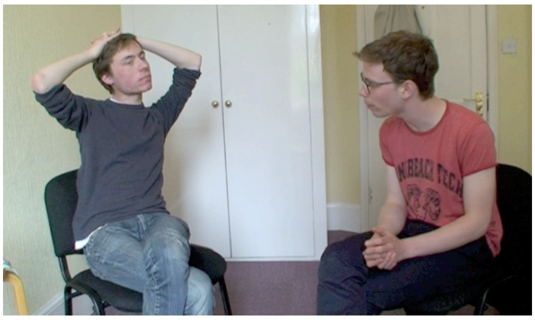


Figure 4: DAN (right) gazing at SIM (left) before constructed dialogue

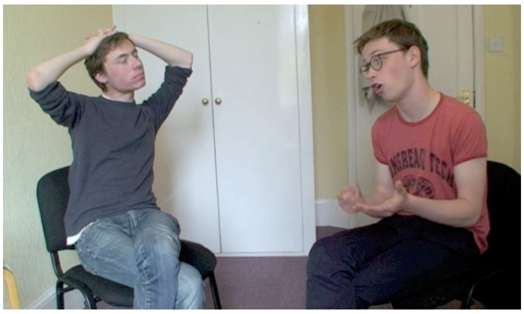


Figure 5: DAN (right) multimodal enactment on and they were saying*: trunk back, gaze away, exaggerated facial expression*

There is a meta-pragmatic ([11]) quality in the speaker's use of multimodal constructed dialogue. By using a full range of verbal, vocal and visual resources, he reminds his interlocutor that *this is an enactment*, i.e. that his words, voice and body are only temporarily used to display another subject's stance and in no way represent his personal beliefs. His interlocutor SIM immediately aligns with him by questioning the absent subject's stance described with a rising intonation on *why not*, and DAN sides with him in verbally questioning the absent subject's stance by presenting it as incomprehensible (*this is what I didn't understand*). DAN's multimodal enactment has allowed him to put the absent subject at a distance, while simultaneously fostering agreement with his interlocutor.

3.3. An in-between, more complex case

In Excerpt 3, AMY has just picked up the question *how can we solve climate change* as part of the semi-guided conversation protocol, and asks it to her interlocutor JOE. As an answer, JOE develops the following stance: nuclear power is a relevant solution to combat climate change (e.g. it replaces polluting coal stations) and it is safe technology since accidents like Fukushima remain rare. AMY's stance in response to his is two-fold. She starts with a concession that is compatible with JOE's stance, thereby partially aligning with him (anti-nuclear activists can oversimplify matters), but eventually disagrees with him (one huge nuclear accident is already one too many).

<u>Excerpt 3</u>

1 AMY: ok er how can we solve climate change?
2 JOE: er pff lots of nuclear power. (small laugh)
3 AMY: mmh, (small laugh)
4 JOE: I know that's a bit controversial at the moment,
but I th… I think it's still a valid point.
(argues in favor of nuclear power for 21 seconds)
5 AMY: I think like a lot of em a lot of anti nuclear sentiment is really not informed at all,
and rather kind of like,
nuclear stuff's poisonous and that's bad,
6 JOE: yeah,
7 AMY: em,
8 JOE: yeah I I think it' a real shame with the with the thing in Japan,
9 AMY: mmh,
10 JOE: er from the point of view of nuclear power as well,
11 JOE: cause it's sort of the,
12 JOE: actually what happened in Japan was this really big exception,
13 AMY: mmh,
(*JOE argues in favor of nuclear power for 8 seconds*)
14 JOE: and then sort of well actually if we if we're just careful,
then then nuclear power is fine.
15 AMY: I guess like the the problem is,
a lot of people understandably will say like,
even if it happens once it's once too often,
but,

Expressing disagreement is a sensitive phenomenon that involves face work ([3]), and agreement is usually preferred to (i.e. is more frequent than) disagreement in interaction ([35], [37]). Owing to politeness mechanisms ([38]), speakers tend to attenuate the potential threat posed to their interlocutor's face thanks to diverse strategies. As exemplified by Excerpt 3, agreement prefacing disagreement, in the form of concession, is one way of downplaying disagreement. AMY starts by adopting a stance that is compatible with JOE's as she criticizes the oversimplified criticisms of anti-nuclear activists. To do so, she uses constructed dialogue introduced by a quotative utterance (*I think like a lot of em a lot of anti nuclear sentiment is (…) kind of like*) to reject the absent subjects' stance just as she provides them with a voice. Her critical distance with respect to them is marked in the verbal modality (*really not informed at all*) as well as vocally. On the direct discourse attributed to the absent subjects, *nuclear stuff's poisonous and that's bad*, her voice pitch markedly shifts to a very high range which is not common at all in her usual way of speaking (Fig. 6).

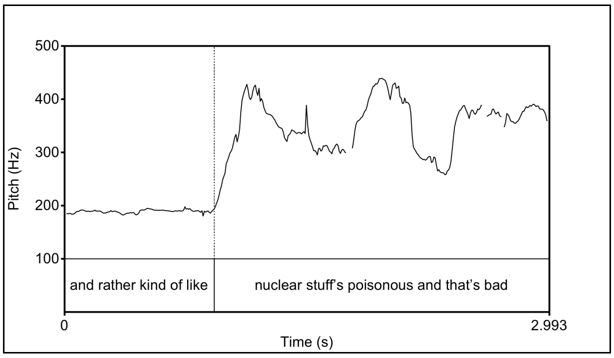


Figure 6. AMY's shift in voice pitch range on nuclear stuff's poisonous and that's bad *(turn 5)*

Yet when using this high-pitched voice, iconic of the scatterbrain attitude she is criticizing, she does not gesture at all. This comes out as slightly incongruous. Indeed, prosody and gesture usually work hand in hand ([39]), with heightened intensity in the vocal modality being simultaneously expressed in some way in the visual modality, and vice versa. The larger stance-taking processes at work here are a plausible explanation for this partial (vocal not visual) enactment of a criticized absent subject's stance. This critical instance of constructed dialogue is not the core of her stance, but only a concession and preface to her real (i.e. anti-nuclear) stance. She is not using the full range of multimodal resources to caricature the absent subjects to the full, because that is not her main point and she partially agrees with them.

To formulate her disagreement with JOE (turn 15), AMY uses oblique strategies that allow her to express a divergent opinion while preserving her interlocutor's face. Her dissent is expressed by a turn-initial *I guess*. As Kärkkäinen ([40]) remarks, when *I guess* is used in second position in a sequence (i.e. in responsive actions to some other actions), it usually indicates some "degree of disagreement and disaffiliation between the participants", as "the current speaker wishes to modify, withdraw, and redefine his or her original stance at this point" ([40]: 197). Disagreement with the interlocutor is also marked on the vocal level: AMY uses a distinct pitch reset on *I guess*: this break in intonation is iconic of a break away from her interlocutor's stance ([41]).

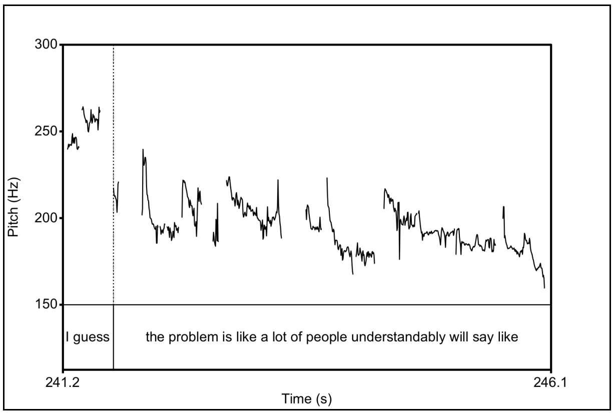


Figure 6: AMY's initial pitch reset on I guess *(turn 15)*

Then AMY presents her discordant stance in an indirect way, resorting to constructed dialogue as an intermediate to express her opinion. She attributes a stance (*even if it happens once it's once too often but*) to the underdetermined, generic absent subject *people* and positions herself as endorsing this outside point of view by way of the stance adverb *understandably* in the quotative utterance (*a lot of people understandably will say like*). Her endorsement of this utterance is vocally and visually indicated by the continuity in her vocal pitch range and personal gesturing style. This supposedly outside voice cannot be traced to anyone in particular: more likely, it is hers in disguise. In this example, direct discourse works as a hedging technique to avoid disagreeing with the interlocutor too bluntly. AMY's cautiousness in taking an adversative stance is confirmed by a final shoulder shrug, an epistemic emblem expressing uncertainty and disengagement ([34]) just before her final *but*.

4. Discussion

This qualitative study has evidenced that speakers do not use the multimodal potential of constructed dialogue only to represent interactions that have taken place or to narrate past events. Constructed dialogue can often not be traced to a speaker's original utterance at another time and place. It is also a pragmatic strategy that allows the speaker here and now to present a person or a group's stance in a more vivid, embodied way. Constructed dialogue allows speakers to articulate two levels of intersubjectivity: they position themselves with respect to both present subject (interlocutors) and absent ones (brought in by constructed dialogue). More specifically, positioning themselves with respect to absent subjects is one way of positioning themselves with respect to present ones. Many combinations are possible: the speaker enacts the absent subjects' stance to take on their perspective, and the interlocutor aligns, empathetically sharing the experience put on display by the speaker (Excerpt 1). In other cases, the speaker can mobilize his own talking body as a medium to ridicule an absent subject, thereby inviting his interlocutor to side with him on the topic at stake (Excerpt 2). The enactment of an absent subject's stance to put it at a distance can also be partial (verbal and vocal only, not visual) when the speaker caricatures this absent subject's stance to side with the interlocutor only temporarily and partially in a movement of concession, just before disagreeing with him (Excerpt 3). Constructed dialogue can be used as a hedge to downplay disaffiliation with the interlocutor, so as to ensure the politeness of the exchange: the speaker lessens her endorsement of the disagreeing stance by attributing it to an absent subject and agreeing with it (Excerpt 3).

There is a continuum in the multimodal intensity of constructed dialogue: not all instances of constructed dialogue include enactments of the absent subject's body or voice. Non-verbal resources, and most strikingly voice pitch, seem iconic of the speaker's stance with respect to the absent subject. When speakers make a distinction between they own voice and the other voice through a marked change of pitch, they distance themselves from this other voice/stance by marking it as different. Conversely, continuity in one's voice pitch when presenting another voice can indicate the speaker's endorsement of that voice/stance. Likewise, continuity in one's gesturing style (e.g. similar speed and amplitude) can mark the speaker's endorsement of the absent subject's stance, while suddenly using more ample, faster gestures can express distance through pantomimic caricature. In all, constructed dialogue takes on different stancetaking functions in context, depending on the kind of multimodal resources that are mobilized.

5. Conclusion

On a theoretical level, this qualitative study invites to a re-evaluation of Du Bois's ([18]) model of stance as a triangle between two subjects (the speaker and the interlocutor) and a discourse object. Constructed dialogue makes explicit not only the backdrop of possible perspectives ([7]) on a given topic, but also the other, absent subjects who take on these stances. The Stance Triangle could be redefined as a Stance Tetrad, where speakers position themselves with respect not only to an object and a present subject but also to absent subjects. This in turn invites a redefinition of the interaction context. As the speaker positions himself with respect to absent subjects as well, the interaction context becomes indexical of the larger social context ([42]).

This qualitative study opens up further research perspectives. A larger corpus and quantitative methods could permit to operationalize "self" and "other" voice pitch and gesturing style according to a set of specific features. Variations in the gestures' speed and amplitude could be measured by motion capture equipment and a PRAAT script could be designed to measure a speaker's average pitch and standard deviation, so as to test whether continuity or change in voice pitch and/or gesture quality function as predictors of endorsement or distance vis-à-vis an absent subject's stance presented as constructed dialogue.

6. References

[1] Tannen, D., "Introducing Constructed Dialogue in Greek and American Conversational and Literary Narratives", in F. Coulmas [Ed], *Direct and Indirect Speech*, 311-322, Mouton de Gruyter, 1986.

[2] Pascual, E., *Fictive interaction. The conversation frame in thought, language and discourse*, John Benjamins, 2014.

[3] Goffman, E., *Interaction Ritual: Essays on Face-to-face Behavior*, Anchor Books, 1967.

[4] Authier-Revuz, J., Hétérogénéités(s) énonciative(s), *Langages*, 73: 98-111.

[5] Ducrot, O., Quelques raisons de distinguer "locuteurs" et "énonciateurs", *Polyphonie – linguistique* et *littéraire* III: 19-41, 2001.

[6] Celle, A., Tense, Modality and Commitment in Modes of Mixed Enunciation, *Belgian Journal of Linguistics* 22: 15-36, 2008.

[7] Martin, J. and White, P., *The Language of Evaluation: Appraisal in English*, Palgrave Macmillan, 2005.

[8] Stubbs, M., A matter of prolonged fieldwork: notes towards a modal grammar of English, *Applied Linguistics* 7: 1–25, 1986.

[9] Kendon, A., *Gesture: Visible Action as Utterance*, Cambridge University Press, 2004.

[10] Sidnell, J., Coordinating gesture, talk and gaze in reenactments, *Research on Language and Social Interaction*, 39(4): 377-409, 2006.

[11] Streeck, J., *Gesturecraft: The manu-facture of meaning*, John Benjamins, 2009.

[12] Sweetser, E., Introduction: viewpoint and perspective in language and gesture, from the ground down, in B. Dancygier and E. Sweetser [Eds], *Viewpoint in Language: A Multimodal Perspective*, Cambridge University Press, 1-24, 2012.

[13] Clark, H. and Gerrig, R., Quotations as demonstrations, *Language*, 66(4): 764-805, 1990.

[14] Janzen, T., Two ways of conceptualizing space: motivating the use of static and rotated vantage point space in ASL discourse, in B. Dancygier and E. Sweetser [Eds], *Viewpoint in Language: A Multimodal Perspective*, Cambridge University Press, 156-175, 2012.

[15] Cuxac, C., *La Langue des signes française. Les voies de l'iconicité*, Ophrys, 2000.

[16] Ferrara, L. and Johnston, T., Elaborating Who's What: A Study of Constructed Action and Clause Structure in Auslan (Australian Sign Language), *Australian Journal of Linguistics*, 34(2): 193-215, 2014.

[17] Stec, K., Meaningful shifts: A review of viewpoint markers in co-speech gesture and sign language, *Gesture*, 12(3): 327-360, 2012.

[18] McNeill, D., *Hand and Mind*, University of Chicago Press, 1992.

[19] Ochs, E., Constructing Social Identity: A Language Socialization Perspective, *Research on Language and Social Interaction*, 26(3): 287-306, 1993.

[20] Du Bois, J., The Stance Triangle, in R. Englebretson [Ed], *Stancetaking in Discourse: Subjectivity, evaluation, interaction*, John Benjamins, 139-182, 2007.

[21] Marzano, M., *L'éthique appliquée : De la théorie à la pratique*, Presses Universitaires de France, 2008.

[22] Mondada, L., Video Recording as the Preservation of Fundamental Features for Analysis, in H. Knoblauch, J. Raab, H. Soeffner, B. Schnettler [Eds], *Video Analysis: Methodology and Methods*, Peter Lang, 51-68, 2006.

[23] Norris, S., Analyzing *Multimodal Interaction: a methodological framework*, Routledge, 2004.

[24] Stivers T. and Sidnell, J., Introduction: Multimodal Interaction, *Semiotica*, 156(1/4): 1-20, 2005.

[25] Mondada, L. Participants' online analysis and multimodal practices: projecting the end of the turn and the closing of the sequence, *Discourse Studies,* 8(1): 117-129, 2006.

[26] Boersma, P. and Weenink, D., Praat: doing phonetics by computer [Computer program]. Version 5.4.08. Online: http://www.praat.org/

[27] Sloetjes, H. and Wittenburg, P., Annotation by category – ELAN and ISO DCR, in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008. Online: http://tla.mpi.nl/tools/tla-tools/elan/

[28] Stivers, T., Stance, Alignment, and Affiliation During Storytelling: When Nodding Is a Token of Affiliation, *Research on Language and Social Interaction*, 41(1): 31-57, 2008.

[29] Stec, K., and Sweetser, E., Managing Multiple Viewpoints, presentation at ICLC Theme Session on Mixed Points of View in Narrative, 2013.

[30] Chafe, W. *Discourse, Consciousness and Time: the Flow and Displacement of Conscious Experience in Speaking and Writing*, University of Chicago Press, 1994.

[31] Debras, C., Formes et fonctions du discours d'expert dans des discussions sur l'environnement entre étudiants britanniques : une étude multimodale de la prise de position en interaction, *ASp*, 65: 45-58, 2014.

[32] Heritage, J., The limits of questioning: negative interrogatives and hostile question content, *Journal of Pragmatics,* 34: 1427-1446, 2002.

[33] Fuchs, Y., *Les quotatifs en interaction en anglais contemporain*, Presses Sorbonne Nouvelle, 2013.

[34] Debras C., *L'expression multimodal du positionnement interactionnel (multimodal stance-taking)*, Unpublished thesis manuscript, Sorbonne Nouvelle, France, 2013.

[35] Heritage, J., A change-of-state token and aspects of its sequential placement, in J. Atkinson and J. Heritage [Eds], *Structures of Social Action: Studies in Conversation Analysis*, Cambridge University Press and Éditions de la Maison des Sciences de l'Homme, 99-345, 1984.

[36] Kraut, R. and Johnston, R., Social and Emotional Messages of Smiling, *Journal of Personality and Social Psychology,* 37(9): 1539-1553, 1979.

[37] Pomerantz, Anita, Agreeing and disagreeing with assessments: some features of preferred/ dispreferred turn shapes, in J. Atkinson and J. Heritage [Eds], *Structures of Social Action: Studies in Conversation Analysis*, Cambridge University Press and Éditions de la Maison des Sciences de l'Homme, 57-101, 1984.

[38] Brown, P. and Levinson, S., *Politeness: Some Universals in Language*, Cambridge University Press, 1987.

[39] Bolinger, D., Intonation and Gesture, *American Speech*, 58(2): 156-174, 1983.

[40] Kärkkäinen, E., The role of *I guess* in conversational stancetaking, in R. Englebretson [Ed], *Stancetaking in Discourse: Subjectivity, evaluation, interaction*, John Benjamins, 183-219, 2007.

[41] Morel, M. and Danon-Boileau, L., *Grammaire de l'intonation : L'exemple du français oral*, Ophrys, 1998.

[42] Goodwin, C. and Duranti, A., Rethinking Context: an introduction, in A. Duranti and C. Goodwin [Eds], *Rethinking Context*, Cambridge University Press, 1-42, 1992.

"This area of rain will stick South in the far North"

Pointing and Deixis in Weather Reports

Gaëlle Ferré & Quentin Brisson

Faculty of Languages, Laboratory of Linguistics (LLING), University of Nantes, France
Gaelle.Ferre@univ-nantes.fr, Quentin.Brisson@etu.univ-nantes.fr

Abstract

Pointing in face-to-face interactions has been largely studied in the literature, as well as route and spatial descriptions in map task experiments. Weather reports share features with both when it comes to the expression of spatial information. Although monologic by nature, weather reports, as will be shown in this preliminary study, are highly complex descriptions with constant switches from one viewpoint to another as well as frequent shifts in deictic center. The changes in viewpoints and deictic center may be congruent in speech and gesture, but they may also occur in one modality at a time, thus creating mismatches between the gestural and verbal information. Pointing in weather reports is also conditioned by the type of medium used, and the mixed frames of reference involved by the activity itself.

Index Terms: pointing, weather reports, viewpoint, frames of reference, deictic shifts

1. Introduction

Pointing has been the object of quite a large body of research as shown by a multidisciplinary publication such as [1]. In a variety of areas, from child development to psychology, linguistics and anthropology, research has shown the intimate link between the oral and gestural production in face-to-face interactions, especially as far as deixis is concerned, as noted in [2]. So, for instance, [2] and [3] noted that pointing is culturally determined and that, depending on whether the information encoded in speech is speaker-oriented or not in a given culture, then the information encoded in gesture will be of a completely different nature. Personal, temporal, and spatial deixis is central to discourse in weather reports. Yet, strangely enough, this type of description has only been studied in the field of automatic gesture recognition to the best of our knowledge (as in [4] and [5] to name just a few studies). The reason for this is possibly that weather reports are likely to abound in pointing gestures and that pointing is easier to recognize in terms of form that other more complex gesture types. We are all familiar with weather reports, and at first look, the mapping of gesture-speech units in this type of presentation may seem extremely straightforward, with nothing more than a person pointing to different locations on a map while describing the weather expected in those locations for the next day or couple of days. Now consider the sentence given as a quote from our corpus in the title of this paper: "this area of rain will stick South in the far North", said as the speaker is making a sweeping gesture over an area of the map to his left. What could be the deictic center of such an utterance? If we posit the speaker as deictic center of the utterance, it works well for the beginning of the sentence and for the gesture, but it doesn't work so well for the end of the sentence: "South of what?", "the far North of what?" This is not an isolated case in weather reports. Based on the analysis by [6], who describes a deictic split between the verbal and gestural deictic *origo* in some route directions, we will show how weather reports involve a constant shift of viewpoint and deictic center in the description of weather conditions. Although time plays a major role in weather reports, this paper will focus on person and space deixis, which are complex enough notions in the limited space of this paper.

2. Theoretical background

In [7], Lyons defines *deixis* as referring to "the function of personal and demonstrative pronouns, of tense and of a variety of other grammatical and lexical features which relate utterances to the spatiotemporal co-ordinates of the act of utterance" (p. 636). Levinson ([8]) adds to this definition by describing deixis as pertaining to the personal, temporal, spatial, discursive and social domains.

2.1. Orientational frames of reference

As stated by Levinson in [2], there are several ways in which utterances can be related to the spatiotemporal context of utterance. These ways are what he calls the "orientational frames of reference". A discourse entity may be conceived of in an *absolute* frame of reference, with fixed bearings like cardinal directions. It may also be conceived of in a *relative* frame of reference, so that its location is defined with respect to another location (which can be the speaker, but also the viewer or another object in the field). This means that in order to define the location of the discourse entity, you have to identify the element that serves as a reference point and its own bearings. At last, a discourse entity may be referred to in an *intrinsic* frame of reference, i.e. in relation to a cultural conception of an object. For instance, unless under special circumstances, when we say that we're "sitting in front of a computer", we usually mean that we're facing the screen, not the plugs on the other side.

As shown by [2] and [3], there is great linguistic variation as to which frame of reference is preferred in different cultures. In English, there is a strong preference for the relative frame of reference, although speakers may also use the other two frames. Shifts in frames may occur over a discourse unit but also within a single utterance, and this is true of verbal information as well as information imparted in other modalities, like gesture or body orientation.

2.2. Deictic center

The reference point in frames was originally called *origo* by Bühler ([9]), although others call it the *deictic center*, a term that will be adopted in this paper. Since the frame of reference may change over the course of an utterance, as mentioned in the preceding paragraph, the deictic center may also change, and this applies to both speech and gesture. [2] mentions for instance that “ complex switches of this reference point or origo may occur in a single sequence of gestures” (p. 250). However, he seems to consider examples in which the reference point in speech and co-occurring gesture is the same. In [6], Fricke describes an example of route description and shows that there is a mismatch between the deictic center in speech and co-occurring gesture: over the course of the utterance, the origo of the gesture is the speaker’s body while the origo in speech content corresponds to the addressee imagined as wanderer. This is a crucial point for the study of weather reports as we will see that the nature of the description involves not only shifts in frames, but also of course shifts in deictic center and that the shifts may also generate mismatches between gesture and speech in terms of reference point.

3. Data

3.1. Corpus

The corpus on which this preliminary study is based consists in a small collection of 10 weather reports posted on the internet between 2011 and 2015. All the reports are in English. 5 of them are presented by female speakers, and the other 5 by male speakers, although no difference was noted in speech or gesture for male and female speakers. The reports we selected were from different TV channels located in various English speaking countries in order to avoid falling into the analysis of presenters’ speech and description habits. Table 1 below presents the channels, the location of their head offices, and the number of reports we downloaded from their websites.

Table 1. Channel, location of head offices, and number of reports downloaded for the present study.

Channel	Location of head offices	Nb of reports
National Weather Channel	Atlantla, US	2
Local CBS	New York City, US	1
BBC weather	London, UK	4
eNCA	Johannesburg, South Africa	1
Local CBS	Los Angeles, US	1
CNN	Atlanta, US	1

The total duration of the recordings was 15.41 minutes (8.04 min for male speakers and 7 min for female speakers), with report duration varying from 55 seconds to 2.35 minutes (average report length: 1:30 minutes).

3.2. Annotations

Pointing gestures and co-occurring speech in each report were coded using Elan ([10]) for easier retrieval and comparison of examples (181 occurrences). As far as gestures are concerned, we counted the preparation and retraction phases ([11]) as parts of pointing if there were any. We noted hand shape (open palm, extended index finger, or index and middle finger, or thumb, or little finger). We also noted if the gesture was a fixed pointing or described a path (that includes the *contour following gestures* or *area-sweep gestures* noted in [12]). At last, we noted gesture direction (away from or towards speaker) and arm extension (small, medium or full extension).

On two more tracks, we noted the target type of the pointing gesture (point or area), and whether the map was static or dynamic (scrolling map or animated icon on the map) during gesture production.

4. Formal description of weather reports

4.1. Overall structure

Although everyone is quite familiar with weather reports, a short formal description of their major features will provide a starting point for the analysis of deixis. Basically, a weather man or woman standing beside a map of a region or nation is describing what the weather is or is going to be like in different locations on the map. In addition to location points, icons and drawings representing weather types or fronts but also text may be superimposed on the map.

After greetings and an optional short introduction, the speaker launches into the weather report generally starting with the present state of affairs (which constitutes the spatiotemporal reference point) and proceeds with a description of the weather conditions expected for a later time (which can be the same day or the following day). This is immediately followed by a description of the expected temperatures. At last, the forecast is extended to the next couple of days or even the coming week before closing the presentation with farewells to the audience. A weather report is a monologue throughout, although it may be launched with a short question-answer sequence if it is included in the larger context of a news report.

4.2. Types of maps

The traditional weather report uses a background projection screen behind the speaker who monitors his gestures with the help of a control screen (placed either in front or to the side) in order to avoid turning his back on the audience during the description. With this setup, speakers point at locations without looking at the map. Some reports use an oblique screen with respect to the speaker and a different camera angle which presents the screen in the larger frame of the recording studio. In this case, no control screen is used and speakers gaze more freely at the map. The difference between the two types of screen is shown in Figure 1.

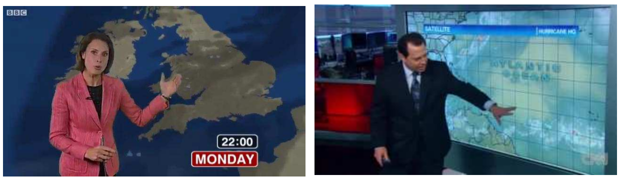


Figure 1: Traditional background projection screen (left) [17] and oblique screen (right) [18] used by different channels for weather reports.

In some reports, the scale of the map remains constant throughout the report with possible dynamic objects superimposed on the map, but in others, the scale of the map changes as the report progresses which has an obvious impact on gesture size and direction as shown in Figure 2. In this report, although the speaker basically points to a same area on the map of the UK in the two frames, the two gestures are radically different due the shift in scale. The fact that the medium may change is specific to weather reports and does not occur when people point at maps in route descriptions where the map is held constant.

This has an impact on gesture-speech mapping in terms of deixis since speakers may refer in speech to locations “further away” from a reference point previously established while gesturing closer to their own bodies.

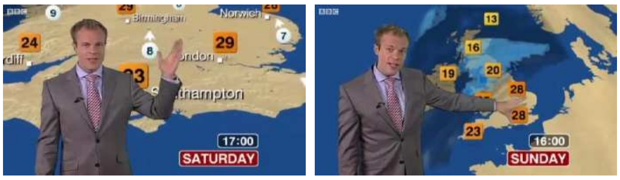


Figure 2: Pointing at the South-East area of the UK on a smaller scale map (left) and larger scale map (right) [19].

Another aspect of the maps used in weather reports is that at times, the scale may be held constant while the speaker is pointing at two different locations on the map, but the map may be scrolling down which may generate yet another type of gesture-speech mismatch as in example (1) below (the two gestures of interest are marked in the text with square brackets and are illustrated in Figure 3 below):

(1) (a) [Maybe some patch of rain in Cambria] by the end of today, (b) [but down towards Manchester], Birmingham, temperatures widely up into the high 20s.

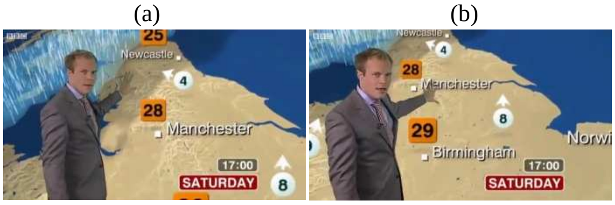


Figure 3: Two pointing gestures in example (1): effect of scrolling map ([19]).

As he mentions Cambria, the weather man points at a location on the map as shown in (a) and he then partly retracts his gesture (arm still extended but somewhat lower, fist closed). While saying "down in Manchester", he opens up his hand and points slightly upwards again towards the location on the map as illustrated in (b). In between the two gestures, the map has scrolled down, which creates the gesture-speech mismatch typical of weather reports when a speaker points upwards while uttering "down". When in "real life", a gesture would follow a moving target; this is not what happens with a scrolling map in the context of a weather report.

4.3. Frames of reference

The survey maps in weather reports use an absolute frame of reference with externally fixed coordinates. By convention, the North has been placed at the top of the map since the discovery of the magnetic North and the invention of the compass. However, this absolute frame of reference is inserted in the relative frame of reference of the speaker, since the North point of the map is not oriented to the North in the recording studio, and since the speaker is standing on one or the other side of the map (which may change as well in the course of the report). The two frames of reference are mixed in speech in example (2) below with a first mention of the absolute frame of reference ("the North West", map internal) followed by the relative frame of the speaker ("down across", map external) in the next sentence:

(2) This is the result of this weather front. It's pushing in <u>from the North West</u>. It's gonna sink <u>down across</u> the country overnight.

5. Viewpoint in weather reports

The viewpoint adopted in weather reports can be quite straightforward. Example (3) below shows a perfect example of a weather woman addressing the audience. The viewpoint adopted is that of the speaker both in speech and gesture:

(3) Let's come back to that cluster of clouds I showed you earlier. (a) [There it is on our forecast charge], (b) [an intense area of rain].

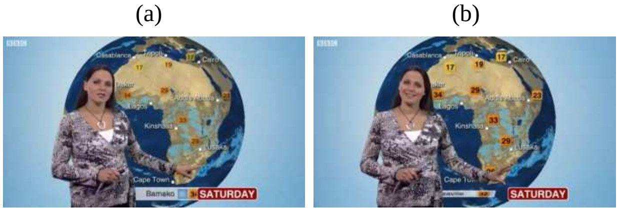


Figure 4: Two pointing gestures in example (3): speaker viewpoint ([20]).

In this utterance, the pronoun "I" clearly makes reference to the speaker, and "earlier" makes reference to a time prior to the time of speaking so that the frame of reference is clearly established as the "here and now" of the speaker. The two pointing gestures that accompany speech are static gestures, one being a point with the index finger while she says "there it is", thus locating a precise location on the large scale map, the other a point with extended open palm in the same direction, matching the word "area" in speech. Both gestures locate a point or an area on the map with respect to the speaker's body so that there is a perfect match in between the "here and now" in speech and the "here and now" in gesture, with a viewpoint that is completely external to the map. Such a perfect match is however not always met in weather reports as will be shown below, and mismatches may arise due to switches in viewpoint, in speech, in gesture or both, or to shifts in deictic center.

5.1. Switches in viewpoint

Consider the following example from a BBC national report in the UK:

(4) That cloud then stays with us in the South. To the north of it, here we've got some clearer spells.

The question is: Who is "us/we"? Or in other words, what is the viewpoint adopted in this stretch of talk? Traditional grammars describe the pronoun "we" as including the speaker and any number of other people. This is perfectly appropriate for the mention of "us" in the first sentence. Since the recording studio is located in London, it makes sense to consider that "us" refers to the speaker and the part of the audience who lives in the South. The mention of "we" in the second sentence cannot however refer to the same people. For one thing, another part of the audience is included in the reference, and besides, the speaker cannot be "in the South" and "to the north of it" at the same time, which means that "we" can't include the speaker this time.

However, in weather reports, speakers are supposed to be objective and not take their hometowns as systematic deictic centers in the description. In order to remain as objective as possible, they generally assume the role of *narrator*, whose deictic center is as neutral as possible and is placed on the map in the middle of the space they describe (a nation, a country, a region...).

So in example (4) above, one can consider that in the second sentence “we” refers to the part of the audience living north of the cloud and the narrator of the report. The viewpoint in speech is now map-internal. Yet, the neutral deictic center (supposedly located in the middle of the UK in this national weather report) has shifted north as a new focus of attention ([13]) has been opened and the narrator’s deictic center has been “transposed” ([14]) to the North as indicated by the use of the proximal deictic pronoun “here”.

Now consider example (5) from a BBC report that presents the weather in some parts of Africa (some major towns on the continent are mentioned but not in every country though):

(5) In Maputo we’re in for a bit of a deluge over the next couple of days.

The head office for this report is located in London and the report is not addressed to people living in Africa, but actually to people in the UK who might want to travel to the African countries mentioned in the report, or who have relatives there. The speaker of the report is therefore not in Maputo at the time of speaking and neither is the audience. “We” in this utterance includes the space of the narrator again, as well as some imagined people living there but not the audience of the report and the deictic center has once again shifted from some neutral deictic center to “Maputo” which becomes the temporary “present” location for the description.

Example (6) below is an extract from a CBS report about a blizzard over New York City. While uttering this sentence, the speaker makes 4 pointing gestures: 2 static gestures toward fixed locations on the map (first with index finger, then with tip of all fingers a little bit below the first gesture), followed by a dynamic sweeping gesture upward with the open palm of his hand, and at last a double-handed open palm gesture towards the map. The 4 gestures marked in the text with square brackets are illustrated below the example in Figure 5.

(6) $_{(a)}$ [Right now], $_{(b)}$ [we’re dealing with them and we’re out in the East end of Long Island] $_{(c)}$ [all the way up through the coast] of Maine and $_{(d)}$ [this is your picture at around 3 o’clock] in the afternoon.

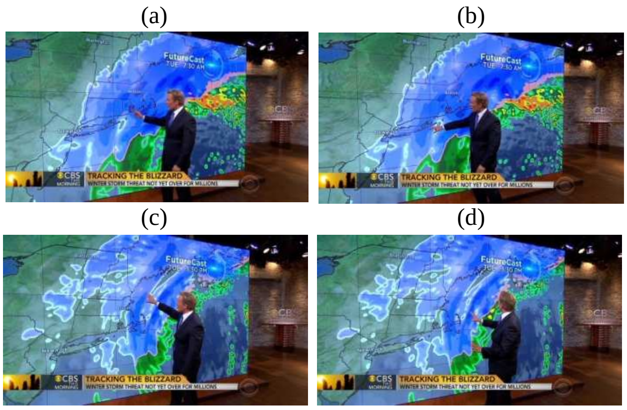

(a) (b) (c) (d)

Figure 5: Four pointing gestures illustrating viewpoint switches over the course of an utterance ([21]).

In (a), the gestural viewpoint adopted is that of the speaker, pointing to a particular location on the map (New York City) that establishes a first reference point, but the deictic center is then shifted south and north of New York in (b) and (c) which reveals not the viewpoint of the speaker but that of a narrator moving along the coast as he follows the expected progression of the represented snow front. The deictic center has shifted from New York City to the East end of Long Island which enables the narrator to express a distal information in speech with “all the way up”. In (d), the viewpoint is again that of the speaker standing in front of the map (and actually looking at the top right corner of the map where the expected time is noted in script) and addressing the viewers so we are back to a speaker/audience relationship, instead of the previous narrator/narrated object point of view.

As we can therefore see, switches in viewpoint may occur over the course of a verbal utterance, but may also be visible in the gestures made by the weather person. So far, gestures and speech have been congruent in the expression of viewpoint. We will now look at examples in which the viewpoint expressed in speech is not the same as the viewpoint expressed in gestures.

5.2. Viewpoint mismatches in speech and gesture

Example (7) below was uttered in the same CBS weather report as the extract just mentioned, whose head office is located in New York. The weather reporter is talking of a blizzard over New York City:

(7) While we were sleeping [it just pushed a little bit more off to the East].

The verbal viewpoint adopted in this example is multiple since the deictic center shown by the use of “we” is New York City, and this is the space of the speaker, of at least part of the audience, as well as the deictic center of the narrator so the deictic center adopted lies at the intersection of the three spaces as represented in Figure 6 below:

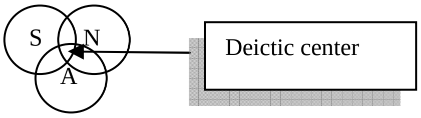


Figure 6: Deictic center at the intersection of speaker (S), narrator (N) and audience (A) space in example (7).

Let’s now look at the pointing gesture made by the speaker when uttering example (7). Index finger extended the speaker points to New York on the map and draws a short a line towards the right, as shown in the left-hand side picture of Figure 7:

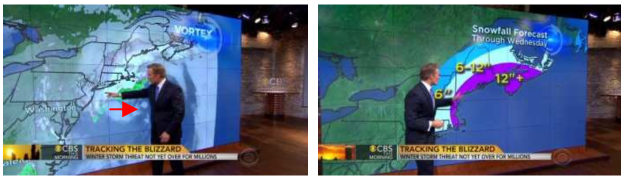


Figure 7: Left: pointing gesture made by the weatherman in example (7); right: position of weatherman later in the video ([21]).

With this gesture, the weatherman illustrates the route taken by the blizzard overnight away from the city. If we consider the viewpoint of the audience and narrator, then the gesture moves away from the deictic center (a point on the map representing New York City), and this is consistent with the semantics of “off” in speech. But at this time of the report, the speaker is standing on the right side of the map, which involves that while saying “off to the East”, the direction of his gesture is towards himself, thus creating a mismatch in between verbal content and gesture. The mismatch could have been avoided with a shift of the speaker to the other side of the map as he does later in the video.

Another mismatch is observed in example (8)

(8) Temperatures a little bit warmer, still running low, somewhere about 70 to 75 [and here comes the next front in] which will arrive late Sunday.

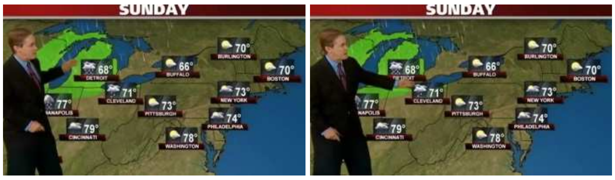


Figure 8: Dynamic pointing gesture produced during example (8) [22].

As he is describing the temperatures in example (8), the weatherman is standing beside the map, facing the audience and gesturing but not pointing at the map. The deictic center is neutral, i.e. located towards the middle of the region of the US he is describing. This deictic center is still reflected in speech when he utters “here comes the next front in”, with the use of the verb “come in” which is appropriate from a viewpoint located in the middle of the map. The gesture however is made from the speaker’s viewpoint, since the weatherman makes a sweeping movement of the end towards the middle of the map. The viewpoint in gesture is then external to the map as it depicts the route the rain front will adopt.

In example (9) below, the reporter makes a continuous pointing gesture towards the map as she utters:

(9) The temperatures are rising as we head for the South.

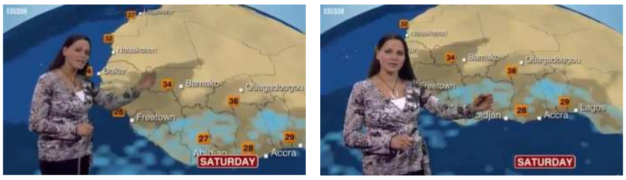


Figure 9: Pointing gesture made in example (9). Arm position during preparation phase (left) and at the end of preparation (right) [20].

This is the same weather woman as for example (5) and we’ve already said that while describing the weather in parts of Africa, she is not in Africa herself and thus does not share the physical space with the object of her description nor does the audience. So when she says “we head for the South”, she adopts the narrator’s point of view, as well as the viewpoint of the viewers, not a speaker’s point of view. Her gesture, however, is made from the speaker’s viewpoint, rather than from the narrator’s: she makes a single static gesture to her left, but during the preparation phase of her gesture, the map begins to scroll down so that her pointing gesture encompasses an area of the map, but does not in itself encode movement southward.

5.3. Shifts in deictic center

As shown in section 5.1, different gestures produced in a sequence may indicate switches in viewpoint, especially switches from speaker to narrator point of view and vice versa. They may also reveal shifts in deictic center. Consider example (10) below, illustrated in Figure 10:

(10) (a) [Now, through this evening], (b) and (c) [that rain will gradually sink its way slowly southward] (d) [in the far North of England].

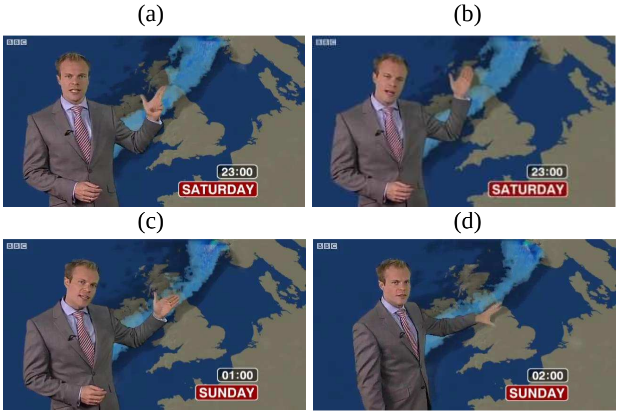


Figure 10: Sequence of gestures produced during example (10) [19].

Just before this stretch of talk, the weatherman was describing rainy conditions over Northern Ireland, with an area of rain extending northward (shown in blue on the map). The pointing gesture with index and middle finger extended in (a) enables the narrator to establish a new deictic center located in the middle of Scotland, which serves as a new reference point for the description of the path that will be adopted by the water front over the coming hours. In speech, this shift of deictic center is shown by the use of discourse marker “now” which does not refer to the present time of speaking but opens up a new focus of attention in a new discourse unit. The direction of the water front is itself depicted in the small sweeping hand movement downward illustrated in (b) and (c), with an open palm covering the area, that matches the accompanying speech “sink ... southward”. The flip of the hand shown in (d) indicates a new shift in deictic center that is present in speech as well: “the far North of England” implies that the reference point is somewhere towards the South of England, and that matches the location of the speaker in London, who is momentarily abandoning the neutral viewpoint of narrator.

Before concluding, let us return to the example given in the title of this paper, and which is reproduced below in (11):

(11) [This area of rain will stick South in the far North].

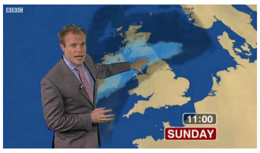


Figure 11: Pointing gesture made in example (11) [19].

While uttering this sentence, the speaker makes a sweeping gesture to and fro along the border separating England and Scotland. While the whole gesture is speaker-oriented and map external, speech functions exactly as in example (10), with a narrative shift from a point in Scotland as the speaker utters “stick South” to a point in the South of England as he utters “in the far North”. The role of the gesture in this example is not to accompany the shift in deictic center, but rather to oppose the two locations implied in speech.

6. Conclusion

This paper presented a preliminary analysis of personal and spatial deixis in weather reports, with a focus on the mapping of gestures with stretches of speech. Beyond the type of map (medium) used in each report, static or dynamic, which has an impact on gesture, creating possible mismatches between the two, it is especially in terms of viewpoint and deictic center that weather reports show real complexity.

In terms of personal deixis, the viewpoint adopted in weather reports corresponds to the “here and now” of the speaker in a relative frame of reference (map-external viewpoint), of a narrator transposed into the absolute frame of reference of the map itself, and/or of viewers (either as people directly sharing the space under description or not). In this Bakhtinian polyphony of voices, gesture and speech may be perfectly congruent as far as viewpoint is concerned, but we have seen as well that a certain viewpoint may be adopted in a modality whereas the viewpoint adopted in the other modality may be different, which is very close to what Fricke ([6]) showed in a face-to-face interaction of a route direction. The externality/internality of viewpoints regarding maps is reminiscent of the distinction made by McNeill ([15] and [16]) for gesture between character or observer-viewpoint.

In terms of spatial deixis, the deictic center is constantly changing as the description of the weather report progresses. Once again, the frequent shifts may result in mismatches in gesture and speech, although this is not necessarily the case.

Because of the already extreme complexity of medium, viewpoint and reference points, the temporal aspects of weather reports were left aside in this paper, but as a weather report is by nature linked with the passing of time, establishing links between past, present and expected weather in the future, all in the condensed form of a one or two-minute presentation, temporal deixis should be included in further studies of weather description. Another possible line of research would be to compare the description of a same event by different channels or in different countries.

7. Acknowledgements

We would like to thank the students of the Discourse Analysis seminar at the University of Nantes, especially Funmbi Fakua and Christophe Gilson for their comments on our presentation of weather forecasts.

8. References

[1] Kita, S. (Ed.). *Pointing. Where Language, Culture, and Cognition Meet*. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey & London, 2003.

[2] Levinson, S.C. *Space in Language and Cognition. Explorations in Cognitive Diversity*. CUP, Cambridge, 2004.

[3] Haviland, J.B. How to Point in Zinacantán. In: S. Kita (Ed.), *Pointing. Where Language, Culture, and Cognition Meet*. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey & London, 139-169, 2003.

[4] Kettebekov, S., Sharma, R. Toward Natural Gesture/Speech Control of a Large Display. *Engineering for Human-Computer Interaction (EHCI'01). Lecture Notes in Computer Science* 2254, 221-234, 2001.

[5] Poddar, I. et al. Toward natural gesture/speech HCI: A case study of weather narration. In *Proceedings of the 1998 Workshop on Perceptual User Interfaces*, 1998.

[6] Fricke, E. Origo, pointing, and speech. The impact of co-speech gestures on linguistic deixis theory. *Gesture* 2(2), 207-226, 2002.

[7] Lyons, J. Deixis, space and time. In *Semantics*, Vol. 2, pp. 636–724. CUP, Cambridge 1977.

[8] Levinson, S.C. Deixis. In: L. Horn (Ed.), *The handbook of pragmatics*. Blackwell, Oxford, 97-121, 2004.

[9] Bühler, K. The deictic field of language and deictic words. In R.J. Jarvella & W. Klein (eds.), *Speech, place and action* (pp. 9-30). New York: Wiley, 1982 [1934].

[10] Sloetjes, H., Wittenburg, P., 2008. Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. http://tla.mpi.nl/tools/tla-tools/elan/

[11] Kendon, A. *Gesture. Visible Action as Utterance*. CUP, Cambridge, 2004.

[12] Pavlovic, V.I. Speech/gesture integration for display control. In *Proceedings of the Army Research Lab Symposium*, Aberdeen, 1-6, 1998.

[13] Grosz, B.J., Sidner, C.L. Attention, Intention, and the Structure of Discourse. *Computational Linguistics* 12(3), 175-204, 1986.

[14] Haviland, J.B. Projections, Transpositions, and Relativity. In: J.J. Gumperz & S.C. Levinson (Eds.), *Rethinking linguistic relativity*. CUP, Cambridge, 271-323, 1996.

[15] McNeill, D. *Hand and Mind: What Gestures Reveal about Thought*. The University of Chicago Press, Chicago and London, 1992.

[16] McNeill, D. *Gesture and Thought*. University of Chicago Press, Chicago and London, 2005.

9. Video references

[17] https://www.youtube.com/watch?v=HSHNkT-V7LY
[18] https://www.youtube.com/watch?v=GsxMhMRtZDw
[19] https://www.youtube.com/watch?v=jIZvHkqTOV4
[20] https://www.youtube.com/watch?v=_XykQpfY_a8
[21] https://www.youtube.com/watch?v=AYGihlgHXog
[22] https://www.youtube.com/watch?v=Qb4RfikG0H4

Analysing the Modifying Functions of Gesture in Multimodal Utterances

Farina Freigang and *Stefan Kopp*

Social Cognitive Systems Group
Faculty of Technology and Center of Excellence "Cognitive Interaction Technology" (CITEC)
Bielefeld University, Bielefeld, Germany

farina.freigang@uni-bielefeld.de, skopp@techfak.uni-bielefeld.de

Abstract

Gestures may contribute to the meaning of an utterance not only by *adding* information but also by *modifying* the gestural or verbal content uttered in parallel. The phenomenon of modification is more common in natural interaction than it has been given attention. We created a corpus of natural communicative gestures and body movements and conducted a study to examine their modifying functions. Results show that index-finger-pointings are most prominent, which emphasise and affirm an uttered content and, thus, are not only used for referencing but also for modifying. Holds emphasise and colour the utterance by showing a stance towards something. Brushing gestures change the utterance in a discounting or downtoning way. A cluster analysis suggests four distinct categories: a focusing category for emphasising an aspect, an epistemic-attitudinal category to convey one's own stance, an epistemic category for uncertainty, and a category where multiple viewpoints are discussed.

Index Terms: gesture, body movements, modifying function, corpus, empirical study, gesture and speech in dialogue, relevance of gesture unit.

1. Introduction

"Sentences are rarely uttered in a behavioural vacuum. We colour and flavour our speech with a variety of natural vocal, facial and bodily gestures, which indicate our internal state by conveying attitudes to the propositions we express or information about our emotions or feelings." [1, p. 1]

Just as prosody adds modal and affective tones to the semantic propositions carried in speech (e.g., [2]), there can be modifying functions of gesture and body movements. These functions may operate on top of the propositional meaning of either speech or the gesture itself. Thus, a gesture may realise a pragmatic modification of the whole utterance meaning. Uncertainty and miscommunication in human interactions may be minimised if one gets hold of which functions appear in natural communication and how they should be interpreted.

In this paper we present an empirical analysis of these, so far under-researched modifying functions of gesture and body movements. Natural for pragmatic modifications or implications are that they depend on the context to which they are added. We want to investigate in particular the modifying functions that gestures can have in different situations, how gestures and movements can be categorised accordingly, and how those can be possibly combined at the same time (multi-functionality). We assume three general classes of functions that express either positivity or negativity related to importance, opinion/emotion and/or knowledge. Besides that, various other interpretations are possible. One of the main goals of this research is to shed light on how modifying functions influence the overall interpretation of an utterance, hence looking more comprehensively at what pragmatic meaning can be communicated by nonverbal behaviour. In the following, we will discuss related work and present our conceptual approach. We then present a rating study on how human observers perceive and interpret modifying functions carried by natural gestures when their verbal context is present vs. non-present. The analysis of the rating study is twofold: we first present descriptive statistics, followed by a cluster analysis.

2. Background

Within the category of pragmatic functions, "Gestures are said to have modal functions if they seem to operate on a given unit of verbal discourse and show how it is to be interpreted." [3, p. 225] Those "modal functions" may be used to express "an hypothesis or an assertion, and the like" [3, p. 159], they are used as "an implied negative" or an "intensifier for an evaluative statement" [3, p. 225]. One gesture that may carry such a modifying function, is the brushing aside gesture, which "usually serves a modal and discursive function: qualifying something as negative and marking the end of a certain discursive activity" [4, p. 1536]. The term "modal function" will be referred to as "modifying function" in this work.

Another gesture category reported to carry modifying functions are open-palm hand gestures [5], in which a hand flip may express epistemicity or a judgemental modality. Also, [6] investigated functions of hand gestures in two Democratic Party primary debates during the 2004 US presidential campaign and observed the following forms: the extended index finger, the slice gesture, the ring (precision grip) gesture, and the power grip. However, besides analysing gestures with a highlighting function, he only focuses on discourse functions. Additionally, [6]'s analysis is based on politicians, which are assumed to perform practised gestures.

In the present work, we are interested in investigating `modifying functions` (MF) in more depth. We concentrate on naturally produced human gestures and body movements that occur in (dyad) interactions, which may carry MF and were accompanied by speech of the same person. The following body movements (BM) are considered: head and shoulder movements, hand and arm gestures and upper BM; and, additionally, coarse facial expressions. We define MF as follows:

If P is the propositional meaning of an utterance (verbal and/or nonverbal), BM or gesture may additionally signal MF which act as an operator F such that F(P) is the combined meaning of the entire multimodal utterance with:

$$F(P) \neq P.$$

Our approach is based on the assumption that, when accompanying speech, BM and gesture may *not only* carry propositional meaning(s) or aspects referring to some content, instance, object, referent or situation of interest in the real word, but that they carry meaning *beyond* any of those propositions. We are interested in exactly all of these BM and gestures.

The BM and gestures meant here belong to the category of ‘pragmatic gestures’, meaning gestures that take up a pragmatic function [3, p. 158]. However, our definition goes a little further: We disregard any so termed ‘pragmatic gestures’ that influence the *structure* (e.g., marking the beginning of discourse (“attention-refocusing” [7]), feedback), or the *timing* (e.g., turn taking) of an utterance, or *refer* to a person or an issue under discussion with a little point or nod [8]. We solely consider MF in pragmatic gestures. More specifically, MF that accompany a propositional message (which is either verbal or nonverbal, i.e., expressed by speech, BM, or gestures) *change* or *add* something to the meaning of the overall message, so that the resulting overall message is different from the ‘purely propositional’ message. Thus, MF in BM and gestures frame the overall meaning of the utterance, namely, they indicate what a person intentionally and non-intentionally communicates and which BM and gestures are used in this process. MF in BM are comparable to modifying words or prosody in speech, which may modify the propositional meaning of a message (e.g., certain acoustic cues are used to convey irony [9]). Although we are interested in the functions of these gestures, we will be using form categories in order to describe how these MF in pragmatic gestures may manifest themselves.

3. Study Design

We conducted a study to investigate and unravel the MF that humans see in BM and gestures. The study was carried out in two parts: first, the corpus creation part included recording and annotating utterances of the candidates and, second, in the rating study part data was presented and rated by naïve participants.

3.1. NIC - Natural Interaction Corpus

The Bielefeld `Natural Interaction Corpus` (NIC) of nonverbal behaviour is comprised of eight dyads (4 female-female, 4 male-male interactions); each dyad consisting of two iterations, one after each stimulus video (the two stimulus videos were shown in alternating order). This results in 16 interactions with three audio recordings (over the participant’s headset and a room microphone) and three video recordings from different angles (each participant recorded from a slight side-angle and both participants from a bird’s-eye view perspective). The total length of all videos is 1 hour and 45 minutes, with an average length of 6 minutes and 30 seconds per interaction, and the recording took place at a CITEC laboratory in September 2014. The stimulus videos consisted of technical instructions with varying complexity and relevance: one was about how to operate a mobile working platform (from which to cut trees, among others) and another video was about how to grout the joints of tiles using silicone. The participants[1] were university students and university staff, all were German native speakers (self-reported), with an average age of 25 (in a range of approx. 20 to 40) years and were paid for 50 minutes of participation in the study. After watching one stimulus video, the two participants talked about the video (with no other person being present in the study room). The participants were informed that we are interested in natural human behaviour in spontaneous dialogues in order to shed light on facets of human communication. In no situation it was referred to BM or gesture. However, we seated the participants on three-legged stools that are a little higher to make it easier for them to use their arms and hands (the rest position was usually the thighs) and which were placed in contiguity and facing one another. The participants performed many natural BM and gestures (although as expected, this depended on the extroversion of the person), among which we also found MF in BM and gesture, mainly pointings, holds and brushings.

In the post-processing of the data, we created manual annotations in ELAN[2] [10] of BM and gestures that we speculated may carry a MF. We define MF to have a *focusing*, an *attitudinal* and an *epistemic* component.

A A **focusing** function highlights or brings into or out of focus an aspect of communication that was communicated by the utterance giver. The utterance giver wants to ensure that the interaction partner perceives the piece in or out of focus.

B An **attitudinal** or an emotional function expresses an utterance giver’s stance, opinion or feeling regarding an aspect of communication. The utterance giver wants to communicate a personal viewpoint and maybe even convince the interaction partner of it.

C An **epistemic** function refers to knowledge or lack of knowledge of an utterance giver regarding an aspect of communication. The utterance giver may want to communicate an assessment or rating of a knowledge content of the same or a different utterance.

BM and gestures that seemed to carry any of these MF were annotated according to three categories: (1) salient movements, those which obviously have a MF and were executed quite clearly, (2) relevant movements which belong to the mainly chosen category, and (3) borderline movements, which showed only very fast, short, small, not easy to recognise MF in movements. BM and gestures were annotated if all of the following criteria could be satisfied: the BM fits the definition of MF (A-C), the BM carries a MF which operates ‘on top’ of a propositional meaning of BM, the BM shapes at least a referent *and* a MF and not a referent alone, the BM is integrated in a person’s utterance and does not stand alone, the BM does not involve any of the following: turn-taking, feedback, word finding, questions, self-adaptors. Additionally, we annotated which of the following body parts were involved in a movement: right/left/both hand(s) (also referring to fingers, e.g., pointing with an index finger), right/left/both arm(s), and (right/left) shoulder(s). We plan to extend the annotation scheme similar to the one created for interpreting the clusters of the cluster-analysis (cf. section 5.2).

3.2. Rating Study

The judgement of uninformed participants in a subsequent rating study had to prove whether other persons also see the MF in BM and gestures. The participants rated the utterances in terms of 14 adjectives that we assumed, first, to be intuitively understandable and, second, correspond to the range of possible combined meanings that can be related back to specific MF.

This study being a proof of concept, we chose not to include fillers and use mostly BM and gestures of the first category

[1] In the following, only pictures of those participants are depicted that agreed to it in a consent form.

[2] EUDICO Linguistic Annotator developed at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

(salient movements) and only in order to balance for various aspects (see below) we added six gestures of the second category (relevant movements). On the basis of the annotations, we automatically extracted video snippets, adding 500 milliseconds before and 150 milliseconds after an annotation. The snippets were between 1194 and 1784 milliseconds long, leaving almost no space for contextual information. Although we are looking at pragmatic functions here, it seemed important to investigate these functions in a minimal time frame in order to extract context independent features.

The experiment consists of two conditions: the first one is the speech-and-gesture (S+G) condition, in which the videos are shown as described above (with speech and with head) and in the second condition, the gesture-only (Gonly) condition, the videos are muted and cropped so that the snippets show only the region between the neck down to the upper legs of the participants (without speech and without head), putting movements of hands, arms, and shoulders into focus (cf. Figure 1). The Gonly-condition provides an isolated view on BM and gestures carrying MF in order to see how much meaning is left. The same 36 video snippets were provided for both conditions. Participant group A watched 18 snippets in the S+G-condition and the other half in the Gonly-condition and vice versa for participant group B, i.e., the videos that a participant saw were all different. We made sure the groups were balanced according to coarse gesture groups (e.g., pointing, hold, brushing), the different participants of NIC and the gender of the participants. Every participant watched the video snippets in the S+G-condition before the Gonly-condition, to mask that this study concentrates on BM and gestural behaviour. For each participant and each condition, the videos were shown in a random order.

The procedure of the rating study was as follows: The participants started with the S+G-condition and every video snippet was played to them three times in a row on the left side of the screen. After these automatic displays, a button could be pressed as often as desired to replay the video. The right side of the screen displayed the heading 'The utterance of the person is ...' with a 7-point Likert scale (excluding forced decision making; 'matches exactly' (1) to 'does not match at all' (7)) and then listing 14 adjectives (displayed in random order with every new video): 'discounting/downtoning', 'revaluing', 'affirmative', 'emphasising', 'classifying', 'emotionally coloured', 'focused', 'critically', 'opinionative', 'negative', 'positive', 'relevant', 'humorous', 'uncertain'.[3] The participants had to rate how much each adjective fitted the expressive behaviour of the person in the video and an answer for every adjective was necessary in order to move on to the next video. No definitions were given for the adjectives, leaving it up to the participants to decide what they mean to them. An optional text field was provided at the bottom of each screen asking for adjectives that would be more characteristic for the utterance. After 18 videos in the S+G-condition had been answered, all participants rated the other set of videos in the Gonly-condition. The final part of the study consisted of definitions for MF and a rating of how the 14 adjectives fit each definition (evaluated by a 7-point Likert scale). The study has been implemented in the Python programming language as a guided user interface, extracting and saving the answers of the participants automatically.

The rating study took place in a seminar room of the university building in March and April 2015. The task was described to the participants as classifying natural utterances of humans nonverbal behaviour; BM and gestures were not mentioned at all. A total of 27 participants took part in the study (13 female, 13 male, 1 other gender). The participants were university students and university staff, all were German native speakers (self-reported), had an average age of 29 (in a range of 21 to 55 years) and were paid for participating in the study (taking from 20 to 50 minutes, depending on answering speed).

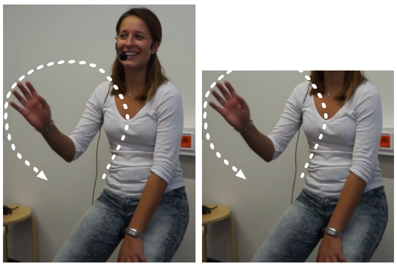


Figure 1: Video snippets of the S+G-condition (left) and the Gonly-condition with muted videos (right). The gesture depicted here is a space holding gesture: the participant performs a circle while saying "I know a lot about this topics".

4. First Rating Study Results

In the following, preliminary results of the video ratings will be presented. Given all ratings for all adjectives, we got a rather normal distribution of all votes with a small tendency towards adjectives that 'do not match' a BM or gesture in a given video snippet. This tendency is a little bigger in the Gonly-condition, when only BM and gestures are observed without sound. In the present analysis, we concentrate on what raters do see in the videos, namely, which adjectives fit the utterance of the person in the video. Tables of the results of the rating study can be downloaded online.[4] In section 5, we will present a cluster analysis based on the same data.

4.1. Adjectives Describing MF of BM and Gesture

The 14 adjectives (as mentioned in section 3.2) were the items of the rating study, which were used with varying frequency to describe the BM and gestures in the videos. Those adjectives that were rated as 'matching a video positively' (*adjectives that describe well what the utterance of the person in the video* does *express*) a lot of times, were 'affirmative' and 'emphasising' and also quite frequent were 'focused' and 'opinionative'. Predicative for 'matching a video negatively' (*adjectives that describe well what the utterance of the person in the video* does not *express*) were 'discounting/downtoning', 'revaluing', 'affirmative', 'humorous', 'uncertain', and also 'critically', 'negative' and 'positive'. In fact, 'humorous' was the most often and clearly rated adjective for describing well what the utterance of the person in the video *does not* express: which on average was the case for every fourth adjective in the S+G-condition and every sixth adjective in the Gonly-condition.

[3] The exact German words were: 'Die Äußerung der Person ist ...', 'abtuend', 'aufwertend', 'bestimmt', 'betonend', 'einordnend', 'emotional gefärbt', 'fokussiert', 'kritisch', 'meinungstragend', 'negativ', 'positiv', 'relevant', 'scherzhaft', 'unsicher'.

[4] Tables of the results of the rating study grouped according to the clusters of the cluster analysis: `http://pub.uni-bielefeld.de/luur/download?func=downloadFile&recordOId=2763501&fileOId=2763503`

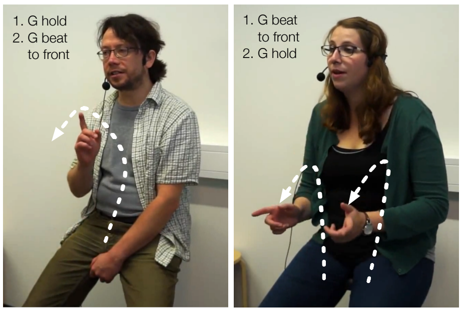


Figure 2: Index-finger-pointings: The ‘baseline’ pointing gesture (left) and a pointing with two hands (right).

4.2. Stable Modifying Functions

As a first analysis, we only considered the video snippets that were rated with maximum clarity and agreement, i.e., with a low standard deviation, namely $\sigma \leq 1.0$, and showing a clear tendency towards one of the poles: $\mu \leq 2.5$ for ‘does match’ and $\mu \geq 5.5$ for ‘does not match’. For now, we will only consider the positive or ‘matching’ cases, that is, the videos with BM and gestures that have been rated with one or more adjectives *matching* what the utterance of the person expresses.

Five videos fulfil these criteria and all of them show pointings with the extended index finger, labelled with the four positive adjectives (cf. section 4.1), but particularly ‘affirmative’ and ‘emphasising’. Three of these video examples are extremely prominent as they have been rated clearly with $\mu \approx 1.6$. Besides the index-finger-pointing, these gestures include an obvious hold of the index finger, a beat before the hold and in one case the index-finger-pointing was done with two hands in parallel (cf. Figure 2). The gestures in the other two videos show only a short index-finger-pointing, in one case the participant snaps his fingers during pointing.

Analysing the data further, and allowing for more uncertainty, the results are not as clear but we can observe a few tendencies. As stated before, the coarse gesture groups are pointings, holds and brushings (example gestures are depicted in Figures 1 to 3). In the group of holdings, some gestures are rated as ‘emphasising’ and ‘opinionative’, but also ‘affirmative’, ‘classifying’ and ‘focused’. Brushing gestures are often seen as ‘discounting/downtoning’ but also ‘emotionally coloured’.

4.3. MF in the Gonly-Condition

The three prominent pointing gestures of the S+G-condition are also most prominent in the Gonly-condition with $\mu \approx 1.6$. In one of the three videos, an index-finger-pointing is rated ‘affirmative’ with even $\mu \approx 1.4$. The other adjectives associated with pointing gestures are ‘emphasising’, ‘focused’ and ‘opinionative’. This suggests a certain amount of communicative ‘self-containment’ of the gestures, even without verbal context. Within the set criteria ($\sigma \leq 1.0$, $\mu \leq 2.5$), another less prominent and quickly performed pointing gesture emerges, rated as ‘emphasising’. This comprises four positive examples in the Gonly-condition with index-finger-pointings. Further analyses of the Gonly-condition shows even weaker but the same tendencies towards ‘emphasising’ and ‘opinionative’ for holding gestures and ‘discounting/downtoning’ for brushings.

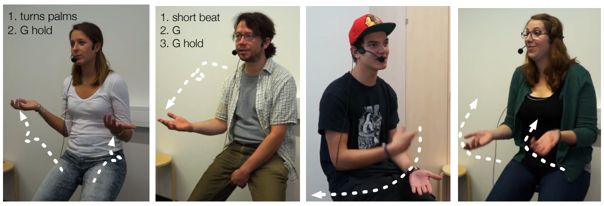


Figure 3: From left to right: First, two holding gestures (one hold open and one hold with one hand) and, second, two brushing gestures (one brushing over the own hand and the other one is brushing/shovelling something away with both hands).

4.4. Interaction Between Modalities

When allocating solely a positive or negative meaning to a BM or gesture and a verbal utterance separately, we observed incongruences between the two modalities. These incongruences are often overwritten by one modality, unless this modality is missing (as in the Gonly-condition). For instance, while performing a shovelling away gesture with two hands (negative), the person has an outstanding positive attitude reflected in the voice and her facial expressions (cf. Figure 3, picture on the right). One prominent rating (according to the criteria above) for this example is that this utterance is ‘not negative’ ($\mu \approx 6.1$), assigning less weight to the gesture while interpreting the mismatching cues in her utterance. In the Gonly-condition, the gesture is interpreted as neutral ($\mu \approx 4{,}5$). In a similar example, the brushing away gesture and the manner of performance (fast, hitting, with a final flap at the end) is purely negative (cf. Figure 3, third picture) just as rated by the participants in the Gonly-condition: ‘not revaluing’ ($\mu \approx 6.0$) and ‘not positive’ ($\mu \approx 5.8$). However, the voice in the video and the facial expressions are rather positive and, consequently, the ratings in the S+G-condition where these features were observed are less negative ($\mu \approx 5.0$ for ‘not revaluing’ and $\mu \approx 4.4$ for ‘not positive’).

5. A Cluster Analysis

In the following, we will present results of a cluster analysis on the rating study data with Ward’s method.

5.1. Method of Cluster Analysis

Ward’s method or *minimum variance method* is a criterion applied in hierarchical cluster analysis. Other clustering methods are used to fuse cluster pairs with the smallest distance (or greatest similarity) in each step. With the Ward criterion, however, clusters and objects are merged step by step in order to reach the smallest increase of heterogeneity within a cluster. Heterogeneity is measured by the sum of variances within the clusters.

We chose a hierarchical clustering method – in contrast to a partitional clustering method like *K-means*[5] – in order to avoid defining the number of clusters before running an algorithm, which we could not know. Then, our goal was to find homogeneous groups, which was perfectly given by Ward’s method. Data outliers cannot be identified but we hypothesise that all pragmatic gestures and BM have a particular meaning or func-

[5] K-means also calculates the sum of variances within the clusters and is therefore quite similar to Ward’s method; although it proceeds differently.

tion and no outliers exist. Several sources confirm that Ward's method is the preferred hierarchical clustering method, citing the comparison of [11].[6] We analysed the rating study (cf. section 3.2) using this method in SPSS.[7]

After applying Ward's method, we used the *Elbow method* to decide on the number of clusters. Here, the percentage of variance is plotted, showing where the slope between data points increases notably and the one data point that forms this 'elbow' indicates the number of clusters. However, the best number of clusters is subjective and not clear-cut. For the S+G-condition an obvious cut exists after four clusters (first 'elbow') but if we are very precise and allow for smaller clusters, we could also agree on seven clusters (second 'elbow'). In order to provide the full picture of the data, we will describe the S+G-condition with four clusters and the according subclusters. Exactly five clusters appeared for the Gonly-condition. For the cluster results compare the tables online.[4]

5.2. Cluster Descriptions and Form Annotations

In a first step, we identified the adjective with highest passing in each cluster. Secondly, we analysed to which extent our categories of MF (focusing, attitudinal, epistemic) were depicted by the clusters; the categories have been annotated in the first post-processing step of the corpus. In order to describe these clusters in more detail, we subsequently annotated the video clips according to form features of nonverbal behaviour. The annotations were carried out without reference to a certain cluster, thus, no similar annotations were made due to neighbouring videos. The annotations included the following categories: 'gesture class', 'hand', 'hand shape', 'hand description', 'palm orientation', 'back of hand orientation', 'arm', 'taken space of movement', 'point in space of movement', 'direction of movement', 'duration of movement', 'additions', 'BM', 'face', 'perspective of movement'. This category system was formed in parts on the basis of two annotation schemes [12] [13] and extended relevant aspects, e.g., shoulders, head and upper BM, and facial expressions.

5.3. Clusters in the S+G-Condition

The four main clusters of the S+G-condition are the following: one cluster with videos depicting primarily a *focusing* MF (`1.A`), one with *epistemic* and *attitudinal* MF (`1.B`), one with negative *epistemic* features (`1.C`) and one cluster with a mixture of *various MF* (`1.D`). For an overview of the clusters of the S+G-condition, cf. Figure 4.

The main gesture class of `cluster 1.A` is 'deictic' including a lot of pointings, carried out primarily by the right arm. The direction of the movement is rather frontward and has an accentuated ending. It may include additions like a beat or a hold or BM like a head nod. Many positive adjectives dominate this cluster such as 'affirmative', 'emphasising', 'classifying', 'focused', 'opinionative' and 'relevant'. Given all of these criteria and considering the previous annotations of MF, we consider this group as representing **(positive[8]) *focusing* MF**, since relevant aspects are marked. When looking closer, two subclusters appear. One resembles the 'deictic' *focusing* category, in that something is 'pointed out', including the nods and a solely frontward and accurate movement with the right arm. The second subcluster forms the 'emphasising' *focusing* category, which contains a lot of hold gestures in addition to the deictic gestures and the gestures are carried out with both hands, in a frontward and upward direction and are realised a little sloppy in a few cases.

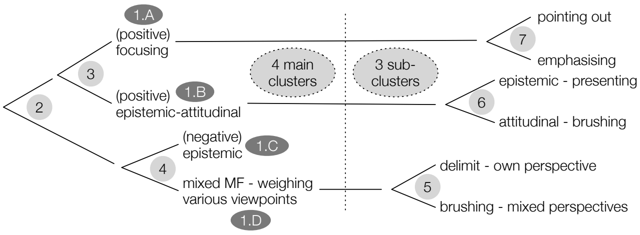


Figure 4: This dendrogram illustrates the cluster partitions and the four main clusters (`1.A` – `1.D`) of the S+G-condition. Note that the distances between the cluster partitions are not accurate.

`Cluster 1.B` accumulates primarily gesture holds and also brushes; both hands are used and negative facial expressions appear. The hands seem to hold up and push down facts, and with these upward and downward movements it is a bidirectional cluster, which is more inherently consistent when looking at the subclusters. Adjectives that *do not* describe this cluster well are 'discounting/downtoning', 'revaluing', 'negative', 'positive', 'humorous' and 'uncertain'. We conclude this to be the cluster of **(positive) *epistemic* and *attitudinal* MF**, since the participants are self-confident about their utterance and have an attitude towards the topic, overall presenting their point of view and statement. The subclusters form two groups. One depicts the rather *epistemic* MF with gestures that present facts on hands, with a neutral face and hold additions. The other group rather accumulates *attitudinal* MF with brushing away gestures, (negative) facial expressions (pinched face, little angry, raised eye brows) and beat additions in some cases.

`Cluster 1.C` consists of gesture holds carried out in vicinity of the initial (or resting) position. The movements contain hedging elements, may be sloppy and without tension and are carried out in an upward direction. Shrugs and head tilts are included and the facial expressions are neutral. Most adjectives have very negative ratings and those which represent the cluster a little are 'discounting/downtoning' and 'uncertain'. We interpret this cluster to show **(negative) *epistemic* MF**, since the persons in the videos seem to be uncertain about what they are uttering. There is no further partitioning.

`Cluster 1.D` consists of very mixed features. It is the only cluster that shows delimiting of something with a movement, e.g., the participants block out space with their hands and arms. However, hold and brushing gestures are similarly prominent. Another interesting aspect is that various perspectives are taken (concluded from the overall utterance of the person): 'my point of view', 'someone else's point of view', 'our point of view'. The movements have a medium to large extend in space and are usually carried out without tension. In some cases circles or wriggles are included. Head shakes and smiles appear with the utterance. There are only minimal trends of adjectives that represent this cluster: 'discounting/downtoning', 'emotionally coloured', 'positive' and 'humorous'; and more often oc-

[6] Bashfield states that Ward's method "clearly obtained the most accurate solutions [...] the minimum variance method is generally preferable" [11, p. 385]

[7] IBM Corp. Released 2013. IBM SPSS Statistics for Macintosh, Version 22.0. Armonk, NY: IBM Corp.

[8] 'Positive' and 'negative' indicate the direction of a MF. An example for a negative *focusing* MF is a brushing away gesture that is used to 'brush' an aspect out of focus.

curring adjectives which *do not*: 'revaluing', 'critically', 'negative' and 'uncertain'. This cluster frames mainly a mixture of ***various (positive)*** **MF**, in particular *attitudinal and epistemic*, but rather no *focusing* MF. We call it the cluster of 'weighing different viewpoints' since it is very diverse and different point of views are taken. Two subclusters exists which divide delimiting and brushing gestures. Delimiting gestures are connected to one's own perspective, beat and hold additions and a very positive attitude. Brushing gestures present rather the mixed perspectives, carrying no additions and are accompanied by a little less positive attitude (in comparison to the other subcluster).

5.4. Clusters in the Gonly-Condition

The five clusters in the Gonly-condition will shortly be described in the following. `Cluster 2.A` consists of similar form annotations and adjectives like `1.A` and represents movements carrying (positive) *focusing* MF. Then, `cluster 2.B` has quite similar adjective ratings and form annotations as `1.C` and, therefore, accumulates (negative) *epistemic* MF. The following three clusters are different from those of the S+G-condition. `Cluster 2.C` is characterised by adjectives like 'affirmative' and 'emphasising' and *not* by 'humorous' and 'uncertain' and by brushing movements and beats. We interpret it as the cluster of (positive) *attitudinal* MF, since a person indicates her stance towards an aspect (only in parts similar to `1.B`). Then, `clusters 2.D` and `2.E` consist of hold gestures with different implications: `Cluster 2.D` carries parts of (negative) *epistemic* MF ('don't know') with a mix of various perspectives. `Cluster 2.E` consists of a mix of *various* MF and various perspectives are discussed. Some videos group similarly in the two conditions, although differences exist already due to the fact that one more cluster emerged in this condition.

6. Discussion and Conclusion

Although the ratings of the first analysis are not clear-cut, they indicate that MF, pinpointed here in terms of a set of adjectives, exist in BM and gesture. In tendency, pointing-like gestures are 'affirmative' and 'emphasising', hold gestures are rather 'emphasising' and brushing gestures are rather 'discounting/downtoning'. So, one important result is that pointing gestures are not solely used to refer to entities in the world, but also have a function of marking an utterance as, e.g., important or meaningful. It is also noteworthy that the most prominent gestures included a beat, which supports the viewpoint that a beat can also have a modal rather than a parsing function [14]. Additionally, the ratings make sense when looking within a gesture: in the prominent cases, if a gesture is rated 'affirmative', it is also rated 'not uncertain'.

However, the roles of the verbal and gestural utterance and their influence on each other are still not clear. It seems that a MF in BM or gesture is not as prominent when being accompanied by speech and facial expressions, as when being perceived on its own, namely in the Gonly-condition. Here, it seem to be interpreted as more negative which could be a result of the increased uncertainty in this unimodal condition.

The results of the cluster analysis suggest four distinct groups to which our MF relate in a plausible way: focusing/emphasising an aspect of the own utterance, conveying an epistemic-attitudinal statement, expressing epistemic uncertainty and discussing and weighing multiple viewpoints. In order to investigate these groups in more detail and to show each function group with its according form features in all its facets, more data is required (36 video snippets were used in this work).

The approach whether to use adjectives to measure MF in BM and gesture is up for discussion. From our point of view this was a viable first step as a proof of concept; the direct matching of the video snippets to the definitions of MF were difficult to realise due to the complexity of the definitions. By performing further analyses, we hope to find more answers regarding the possible modifications of utterance meaning. It would be interesting to observe this change concentrating on differences in modalities and cases when one modality is omitted.

7. Acknowledgements

This work was supported by a scholarship of the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). The authors would like to thank Cord Wiljes for dubbing one of the stimulus videos and Sören Riechers for his support in the rating study setup. Additional thanks goes to the anonymous reviewers for their helpful comments.

8. References

[1] Wharton, T., *Pragmatics and non-verbal communication*, Cambridge University Press, 2009.

[2] Lu, Y., Aubergé, V. and Rilliard, A., "Do You Hear My Attitude? Prosodic Perception of Social Affects in Mandarin", Int. Conf. on Speech Prosody Proc., 685–688, 2012.

[3] Kendon, A., *Gesture: Visible Action as Utterance*, Cambridge University Press, 2004.

[4] Payrató, L., Teßendorf, S., "Pragmatic Gestures", In: C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill and S. Teßendorf (eds.), Body Language Communication: An International Handbook on Multimodality in Human Interaction. Handbooks of Linguistics and Communication Science 38(1):1531–1539, 2013.

[5] Ferré, G., "Functions of Three Open-palm Hand Gestures", *Multimodal Communication*, 1(1):5–20, 2011.

[6] Streeck, J., "Gesture in Political Communication: A Case Study of the Democratic Presidential Candidates During the 2004 Primary Campaign", *Research on Language and Social Interaction*, 41(2):154–186, 2008.

[7] Norris, S., "Three Hierarchical Positions of Deictic Gesture in Relation to Spoken Language: A Multimodal Interaction Analysis", *Visual Communication*, 10(2):129–147, 2011.

[8] Enfield, N. J., Kita, S. and de Ruiter, J. P., "Primary and Secondary Pragmatic Functions of Pointing Gestures", *Journal of Pragmatics*, 39(10):1722–1741, 2007.

[9] Bryant, G. A. and Fox Tree, J. E., "Recognizing Verbal Irony in Spontaneous Speech", *Metaphor and Symbol*, 17(2):99–119, 2002.

[10] Sloetjes, H. and Wittenburg, P., "Annotation by category: ELAN and ISO DCR", Int. Conf. on Language Resources and Evaluation Proc., 2008.

[11] Blashfield, R. K., "Mixture model tests of cluster analysis: Accuracy off our agglomerative hierarchical methods", Psychological Bulletin, 83(3):377–388, 1976.

[12] Lücking, A., Bergman, K., Hahn, F., Kopp, S., Rieser, H., "Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications", Journal on Multimodal User Interfaces 7(1–2):5-18, 2013.

[13] Bressem, J., Ladewig, S.H., Müller, C., "Linguistic Annotation System for Gestures (LASG)", In: C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill and S. Teßendorf (eds.), Body Language Communication: An International Handbook on Multimodality in Human Interaction. Handbooks of Linguistics and Communication Science 38(1):1098–1124, 2013.

[14] Kendon, A., "Gestures as Illocutionary and Discourse Structure Markers in Southern Italian Conversation", *Journal of Pragmatics*, 23(3):247–279, 1995.

The Influence of Clause Structure on Gestural Depiction of Motion Events

Isabella Fritz [1], Sotaro Kita[2], Jeannette Littlemore [1], Andrea Krott[3]

[1] Department of English Language and Applied Linguistics, University of Birmingham, UK
[2] Department of Psychology, University of Warwick, UK
[3] School of Psychology, University of Birmingham, UK

ixf146@bham.ac.uk, s.kita@warwick.ac.uk, j.m.littlemore@bham.ac.uk, a.krott@bham.ac.uk

Abstract

Co-speech gestures are assumed to be shaped by syntactic patterning [1]. Evidence from cross-linguistic studies suggests that how manner and path are packaged in grammatical clauses influences gestural content. If manner and path are split over two clauses, these aspects tend to be gesturally separated, and vice versa [2]. The current study found that increasing the linguistic distance between manner and path components within one clause increases the likelihood that manner and path will be gestured separately. This indicates that complex clause structures can lead to a reconceptualization of motion events in the process of on-line planning for speech-gesture production.

Index Terms: gesture, production, syntax, motion events

1. Introduction

Speech, and the gestures that accompany it, are coordinated on a semantic and on a temporal level [1, 2, 3]. In his Growth Point Theory, McNeill argues that speech and gesture have a common origin and that gestures and speech constitute an integrated process during production. From this starting point, the Growth Point (GP), speech and gesture emerge in the course of a dynamic process [4]. A GP can be identified through the semantic content of a gesture as well as through the gesture's synchrony with speech [5].

How exactly are contents of gestures coordinated with speech? Previous research on motion events has provided insight into this question, by taking advantage of various possibilities in which manner and path are encoded with various constructions. It has been found that a gesture tends to express the information about a motion event that is contained in a clause in concurrent speech. More specifically, if manner and path are linguistically encoded separately in two grammatical clauses as in Turkish and Japanese, these components also tend to be gesturally separated; in contrast, if manner and path are encoded within one clause as in English these components tend to be expressed by one gesture only [1, 2].

It has been argued that these cross-linguistically different gestures are shaped online during speech production based on the linguistic packaging of information, rather than being shaped by language-specific conceptual schemas for motion events. Evidence for this claim derives from Kita et al.'s [6] finding that speakers of the same language (English) use different gestures to encode motion events, depending on the syntactic structures that they are using. The stimulus movies in this study ("the Tomato Man" movies [7]) elicited two different ways of linguistically encoding motion events from English speakers. Speakers encoded manner and path sometimes within one clause (e.g., "he rolled down the hill") and sometimes across two separate clauses (e.g., "he went down as he rolled"). The gestures that they used corresponded to these differences in clausal packaging. Thus these results revealed the same manner and path separation and conflation patterns that were found in the cross-linguistic studies. Hence, differences in gestural encoding appear to be a result of online processes in speech production and are not due to pre-determined schemata based on language typology [6].

Studies so far on motion event gestures, however, have not made it clear why a clause is the linguistic unit in which speech and gesture coordinate their semantic contents. In studies so far, when manner and path were expressed in different clauses, they were also lexicalized differently. Path is encoded as a verb-satellite (e.g., particle such as up or into) in a one-clause description, but as a verb in a two-clause description (e.g., enter or exit). The differences in gesture use that have been found in the literature might therefore be due to differences in lexicalization patterns. For example, the verb-satellite construction may lead speakers to conceptualize manner and path as a single conceptual unit [8]. This gives rise to two possible accounts for the findings: First the lexicalization of the motion event might be tightly linked to the gesture used (which we refer to here as the **lexicalization account**). Second, gesture use might be linked to a planning scope of speech production. A clause may be a good proxy for a planning scope, but when a clause has a complex structure the planning scope can be smaller than a clause. This would extend the ideas expressed in the Information Packaging Hypothesis [9], which states that gesture tends to encode information expressed by a clause because grammatical clauses are important planning units as they roughly represent one processing unit in speech [10, 11]. We refer to this as the **planning unit account**.

The lexicalization account and the planning unit account make different predictions as to how manner and path are gestured when clausal structures differ but when manner and path are lexicalized in the same way. One way in which this can be achieved is through the insertion of an embedded clause between the manner and the path components. Such separation is not possible in English, but is possible in German. When manner and path are separated from each other in such way, they might have to be processed in two different planning units. In such situations, the planning unit account would predict that manner and path should be separately expressed in two gestures. In contrast, the lexicalization account would predict that manner and path should be expressed through a single, conflated gesture, just as in the case of no separation.

2. Present study

This study tested whether speech-gesture semantic coordination is affected by varying clausal structure when the lexicalization patterns are the same. In both languages that we tested (German and English), the preferred encoding of motion events is a one-clause structure where motion is encoded in the main verb and path is encoded outside of the verb in a so-called "satellite" which in many cases is the verb's particle [12]. These events tend to be accompanied by conflated gestures that express both manner and path in a single movement [1]. Satellite-framed languages are often opposed to verb-framed languages (e.g. Turkish, Japanese) where path must be encoded in the main verb. Manner, if indicated in speech, is subordinate to the main verb, and is usually encoded in a new verb or a clause, which results in a multiple clause structure and different lexicalization patterns from satellite-framed languages [13]. These events tend to be accompanied by two separate gestures, one for the manner and one for the path [1].

In this study however, only motion events encoded within a single grammatical clause were considered. Due to word order flexibility in German, elements of particle verbs depicting motion events are ordered differently depending, among other factors, on the clause type. German main clauses have an S-V-O structure where the verb always has to be placed in the second position of the clause (inversion) and the particle comes in the final position. By inserting elements such as prepositional phrases, direct objects (1) or even whole clauses (2) between verb and particle, a distance can be created between the verb and the particle.

(1) Der Elefant *klettert* einen Regenbogen *hinauf*.
“The elephant *climbs* a rainbow *up*.”
(2) Der Elefant *klettert*, wie im Video gesehen, einen Regenbogen *hinauf*.
“The elephant *climbs*, as seen in the video, a rainbow *up*.”

In German subordinate clauses verb and particle are in reverse order compared to main clauses. Furthermore these two elements are contracted in the final position of the clause (3).

(3) Ich sehe, dass der Elefant einen Regenbogen *hinaufklettert*.
“I see that the elephant a rainbow *up-climbs*.”

We took advantage of these word order differences to investigate manner and path separation and conflation within the same clause. As a control group we tested English native speakers since in English motion events are also framed by a “satellite”, but the particle follows the verb directly regardless of the clause type (4 and 5).

(4) The elephant *is climbing up* the rainbow. (Main clause)
(5) I can see that the elephant *is climbing up* the rainbow.
(Subordinate clause)

According to the lexicalization account, manner and path depiction in gestures should be very similar across both clause types in both German and in English because these languages are typologically the same (satellite-framed languages). Hence, it is assumed that the conceptualization of an event does not differ across clause types and across the two languages. However, the linguistic items inserted between the verb and particle in German main clauses might lead to a clause structure that is too complex to be planned within a single unit. Breaking down this clause into smaller planning processes might result in a reconceptualization of the motion event and, according to the planning unit account, this could result in a gestural separation of manner and path.

3. Methods

3.1. Participants

48 participants took part in the study, 23 German native speakers and 25 English native speakers. Six participants had to be excluded because they grew up bilingually and two participants were excluded since they did not follow the task instructions. Consequently the language and gestures produced by 21 English speakers and 18 German speakers were coded and included in the analyses (3 male participants in German, 1 in English). Participants were aged between 18 and 36 (M = 22.1, SD = 4.2). All participants were tested at the University of Birmingham and they either received course credits for their participation or compensation in form of a £3 Starbucks Voucher.

3.2 Material

13 short cartoons taken from the German children’s series “Die Sendung mit der Maus” (“The programme with the mouse”) were used as stimuli [14]. The cartoon sequences ranged from 3-8 seconds and all trials included a character (mouse, duck or elephant) which performed a motion event. To control speech output, participants were given a particle verb to describe the target motion event (see Appendix for the given particle verbs).

3.3 Procedure

Participants came to the lab and they were told that the purpose of the study was to investigate how different sentence structures of a language influence our speech production in narrations. They were instructed to retell the cartoon clips within a single sentence using either a main or a subordinate clause construction. In order to create a more communicative situation and enhance the use of gestures, the participants retold the cartoons to a third person in the room who was not able to see the video clips. In the subordinate clause condition participants had to begin their retellings with the element “I can see in the video that” (German: “Ich sehe im Video, dass”) (6 and 7).

(6) Ich sehe im Video, dass der Elefant in eine Sandgrube *hineinrollt*. (Subordinate Clause Condition)
“I see in the video that the elephant in a sandpit *in-rolls*.”

(7) I can see in the video that the elephant is *rolling into* a sandpit. (Subordinate Clause Condition)

Initiating a sentence with this clause forced the participants to continue with a subordinate clause, in both English and German. The main clause condition aimed to separate verb and satellite in German. To create a complete separation, German participants had to insert the clause “wie im Video gesehen” (“as seen in the video”) between verb and particle (8). With this insertion it was possible to create a so-called “nested sentence”. To ensure that the participants produced this grammatical structure, they were instructed to start the sentence with the subject (the mouse, the elephant or the duck), followed by the verb in second position. However, the total distance between verb and satellite could vary, depending on how many other elements the participants chose to include.

(8) Die Maus *schwebt*, wie im Video gesehen, (mit einem Regenschirm) in den Pool *hinunter*. (Main Clause Condition)
“The mouse *floats*, as seen in the video, (with an umbrella) into the pool *down*.”

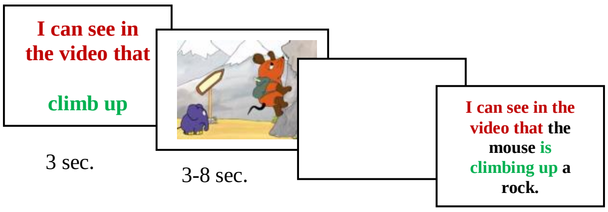
Since it is not possible to insert a clause in between verb and satellite in English, English participants were instructed to place the clause “as seen in the video” at the end of their sentence (9).

(9) The mouse *is floating down* into the pool, as seen in the video. (Main Clause Condition)

In English, the distance between satellite and verb cannot vary; the satellite always follows the verb directly for stimuli such as these. Generally, in English the distance between verb and satellite as well as the word order of the clause remain the same when the sentence structure changes from main to subordinate clause constructions. By adding the clause "as seen in the video" at the end of the sentence, we aimed to keep the speech output and the overall complexity of the sentences in both languages as similar as possible.

The experiment was set up in a PowerPoint presentation and shown to the participants on a laptop. Main clause and subordinate clause conditions were blocked and counter-balanced. Furthermore, within each block the 13 stimuli were semi-randomized. The task was explained to the participants by the experimenter going through an example stimulus (Figure 1). Instructions were given orally. Each condition was explained and introduced with the same example stimulus and practice stimulus. The second condition was not explained until the first condition was completed. Each trial started with a slide containing the particle verb and either the initial main clause in the subordinate condition or the embedded clause (German)/final clause (English) in the main clause condition. This slide was displayed for three seconds before the actual clip started. After the clip had been shown, the screen turned blank and, in order to keep advanced sentence planning to a minimum, the participants were told that they should start their retelling as soon as they saw the blank screen. In the example clip an example answer was shown after the blank slide. This example answer illustrated the correct sentence structure to make it easier for the participants to re-produce it in their own retellings. Concerning the form of the verb (progressive form, tense), no limitations were given and the participants were told that it was up to them which form of the verb they used.

Figure 1 Example stimulus for English


I can see in the video that
climb up
3 sec.
3-8 sec.
I can see in the video that the mouse is climbing up a rock.

Also, participants were instructed to use their hands while describing what the characters are doing, but it was not specified what type of hand movements they should produce. If participants asked how to use their hands or whether a certain type of gesture was correct, they were told that it was up to them what to do with their hands. If participants did not use any gestures twice in a row, they were reminded to do so. For the later analyses, the participants' retellings were video and audio recorded.

3.4 Coding and Analysis

The recordings were coded using the linguistic annotator ELAN [15]. All speech was transcribed, but only responses where the participants produced the trained sentence structure in main and subordinate clauses were considered in the analyses. Hence, manner and path had to be linguistically encoded within one clause for the utterance to be included into the analyses. Other verbs than the ones presented on the slides were included only if they were particle verbs and only if they were semantically similar to the given verb, i.e. participants might have used a different particle (e.g. "emporkriechen" – "crawl upwards"; instead of "hervorkriechen" "crawl out") or a different manner verb ("climb out" instead of "crawl out").

In terms of gesture coding, only strokes depicting the target motion event were considered. The target motion event was the gestural depiction of the given particle verb of each trial. The gestural coding of motion events was based on the "Cross-linguistic Motion Event Project" coding manual [used in 6, 16] and was adapted and elaborated for the stimuli used in the current study. In a first step all target event gestures were classified either as manner, path or conflated. To be classified as a path gesture the gesture could only depict the direction of the event (e.g. for the motion "to float down": this might involve a downward movement with (an) open palm(s) but without any movements to the left or right which would indicate manner). Manner gestures were defined as depicting solely the manner aspect of the motion event (e.g. for the motion "to climb up", this might involve the participant opening and closing their palm(s) without moving their arms upwards). Conflated gestures depicted motion and manner of the motion event in a single gesture (e.g. for the motion "to roll into", this might involve rotating one's wrist(s) with a simultaneous change of location away from the body). Furthermore, in the data a fourth type of gesture occurred that Özyürek et al. [16] termed hybrid gesture. These gestures are a combination of path or manner gestures combined with a conflated gesture within a single stroke.

Next, we classified each response (to each stimulus video) into three types, based on the types of gestures produced: Separated Gestures, Conflated Gestures Only, Singleton Gestures Only. In the category of Separated Gestures, manner and path were both expressed gesturally within a response, and contained two gestures where one aspect of the manner and path information was separated. This included the following combinations: one manner and one path gesture, one conflated gesture combined with either a manner or a path gesture. Hybrid gestures were also classified as Separated Gestures because they separate manner and path in some way. Hybrid gestures combined with a manner or a path gesture were also included in this category. The category of Conflated Gestures Only included responses that contained just conflated gestures. Finally, the category of Singleton Gestures Only included responses in which either manner or path (but not both) were gesturally expressed. It typically contained either one manner only gesture or one path only gesture. In cases where participants combined two manner only gestures in one response (4 instances) or combined two path only gestures in one response (21 instances), these were also classified as Singleton Gestures Only.

For three of the 13 cartoon clips (slide down, jump over, jump into), it was very difficult to encode manner in the gesture, without using whole body gestures, which adults seemed to avoid, and this resulted in the production of path gestures only in the majority of responses for these items. Hence, for the later analyses, these trials were excluded. The responses in which participants failed to follow the instructions (e.g. forgot to include the given clause "as seen in the video" or when they produced a gesture after speech) were excluded from the analyses. The error rate across conditions and participants differed and hence the number of analysed responses in each condition was not equal across participants (see Table 2). German participants sometimes encountered problems in producing the correct structure for main clauses, which led to an especially high error rate in this condition. Responses with an error were excluded from the analysis. Due to differing error rates across conditions, we computed proportion of responses for the analyses.

Table 1 *Error Rates computed on the basis of possible responses (10 per participant for each condition) in the Main Clause Condition and Subordinate Clause Condition in English and German*

Error Rate	Main Clause	Subordinate Clause
German	35 %	17.22 %
English	15.71 %	20 %

For each participant, each clause type and each language, we computed proportion of responses containing the three gesture types (Separated Gestures, Conflated Gestures Only, Singleton Gestures Only). Because Singleton Gesture Only responses are not relevant for our research question, they are not analysed as a dependent variable in the following analyses. Thus, we analysed proportions of Separated Gesture responses and Conflated Gesture responses separately, using 2x2 ANOVAs with Clause Type (Main, Subordinate) and Language (English, German) as independent variables.

4. Results

Results are presented in Figures 2 and 3. The ANOVA for Responses with Separated Gestures yielded a significant main effect of Clause Type, F (1, 37) = 12.85, p = .001, η^2 = .258 and a significant interaction between Clause Type and Language, F (1, 37) = 6.94, p = .012, η^2 = .158. The main effect of Language was not significant (p > .05). Follow-up paired samples t-tests showed that the proportion of Responses with Separated Gestures differed between Clause Types only in German, t (17) = 3.43, p = .003, but not in English, t (20) = .938, p = .359. There was a higher proportion of Responses with Separated Gestures in Main Clauses, where manner and path were also separated linguistically, than in Subordinate Clauses, where manner and path were linguistically contracted. As for Responses with Conflated Gestures, the ANOVA showed no main effects or a significant interaction between Clause Type and Language (all ps > .05). However, paired samples t-tests showed that for German, the proportion of Responses with Conflated Gestures was marginally higher in Subordinate than in Main Clauses, t (17) = 2.09, p = .052, while there was no such difference for English, t (20) = .697, p = .494. Thus, proportion of German Responses with Conflated Gestures showed the opposite pattern than the proportion of German Responses with Separated Gestures.

Figure 2 *Mean Proportions of responses with Separated Gestures Only in Main versus Subordinate clauses in English and German. Error bars represent standard errors.*

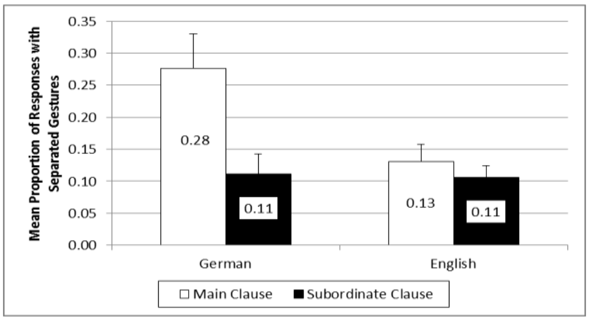
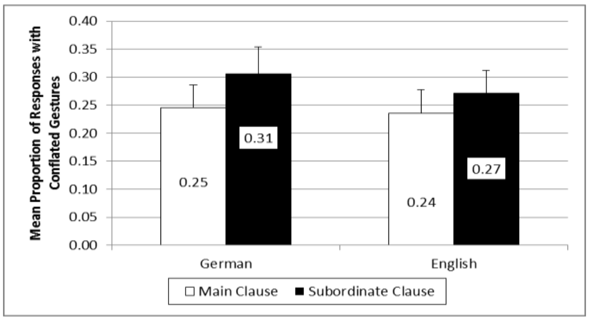


Figure 3 *Mean Proportions of responses with Conflated Gestures Only in Main versus Subordinate clauses in English and German. Error bars represent standard errors.*



5. Discussion

The results show that the usage of gestures was affected differently by clause type in the two languages. In English, where clausal packaging for main and subordinate clauses stays the same, no significant differences were found for the two clause types, for both separated and conflated gestures. In German however, we did find a significant difference of separated gestures for the two clause types. More specifically, the likelihood of a gestural manner and path separation was higher in German main clauses where manner and path were linguistically separated compared to subordinate clauses where these two elements were combined within one lexical item. For conflated gestures we could find a trend towards a higher proportion of responses with conflated gestures in subordinate clauses as compared to main clauses.

Due to these differences found in the German data, our results do not support the lexicalization account. According to this account gesture patterns should neither differ across clause types nor across languages because German and English both frame motion events with a satellite that is encoded outside of the manner verb. Based on this lexicalization pattern, Slobin [8:32] argues that speakers of a satellite-framed language conceptualize manner and path as “a single conceptual event” which would predict a gestural conflation of manner and path when describing motion events.

Our results rather support the planning unit account, suggesting that during online planning for speech and gesture, complexity of a clause plays a role in how we conceptualize a motion event. The clause construction in our German Main Clause Condition might have been too complex to constitute one planning unit. Hence, this complex clause was broken down into smaller planning chunks. This sub-chunking did lead to a reconceptualization of the motion event during online planning. More specifically, our results suggest that the amount of information we can package within one planning unit in speech also translates to gestural planning units.

The conclusion of the current study has important implications for the literature on gestural expression of manner and path. Previous research has shown that when manner and path are linguistically encoded within a single clause, gestures tended to conflate manner and path in a single stroke, and when manner and path are encoded in two separate clauses, gestures tend to separate manner and path [1, 2]. The current results indicate this “clausal packaging effect” is in fact a processing unit effect, as assumed by the Interface Model [1], based on psycholinguistic studies providing evidence for clause as a planning unit [11]. This interpretation is empirically supported, for the first time, by the current study, which demonstrated that within-language differences of gestural representation of motion events can be elicited by manipulating the clause-internal complexity.

Furthermore, we provided further evidence for Kita et al.'s [6] claim that gestures are shaped online during speech production and that they are not bound to a habitual way of gesturing based on language typology. German speakers changed the way they express manner and path in their gesture, depending on the grammatical structure they used for each utterance.

There are a few important issues to be addressed in future studies. First, a more naturalistic production task (including spontaneous co-speech gestures) would give further insight into how gesture and speech interact when lexicalization patterns are the same but clausal packaging differs. Second, the current study design revealed different information packaging patterns within one clause but the scope of these new clausal sub-chunks is still unknown. Since the length of clauses varied across responses depending on how many elements the participants included in their retellings, it can be assumed that a varying length and complexity of clauses led to different information packaging. Evidence for that are the multiple gestural patterns within our Separated Gestures category including hybrid gestures where manner and path are separated within one single gesture stroke.

Finally, the nature of the inserted elements between manner and path might play a role in how speakers package clause-internal information. In our study we inserted a whole clause in between these two elements which did not add any information to the retelling but rather reorganized the clause structure. Future study designs might consider different types of elements which are inserted between verb and particle. This could shed light on the conceptualization of motion events and how they are linked to syntactical encoding.

6. Conclusions

We showed that an increased clause-internal complexity in German main clauses can lead to a reconceptualization of motion events during online planning processes, such that manner and path are conceptualized in separate planning units for speaking and thus manner and path are expressed in two separate accompanying gestures.

7. Acknowledgements

The first author received funding from the College of Arts and Law (University of Birmingham) to conduct this study as part of her PhD project.

8. References

[1] Kita, S. and Özyürek, A., "What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking", Journal of Memory and Language, 48(1): 16–32, 2003.

[2] Özyürek, A., Kita, S., Allen, S., Furman, R. and Brown, A., "How does linguistic framing of events influence co-speech gestures? Insights from cross-linguistic variations and similarities", Gesture, 5(1/2): 219–240, 2005.

[3] Mol, L. and Kita, S., "Gesture structure affects syntactic structure in speech", in N. Miyake, D. Peebles and R. P. Cooper [Eds.], Proceedings of the 34th Annual Conference of the Cognitive Science Society, 761-766, Austin, TX: Cognitive Science Society, 2012.

[4] McNeill, D., Gesture and Thought. Chicago, London: Chicago University Press, 2005.

[5] McNeill, D. and Duncan, S. D., "Growth Points in Thinking-for-Speaking", in D. McNeill. [Ed.], Language and Gesture, 141-161, Cambridge: Cambridge University Press, 2000.

[6] Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., and Ishizuka, T., "Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production", Language and Cognitive Processes, 22(8): 1212-1236, 2007.

[7] Özyürek, A., Kita, S., and Allen, S., "Tomato Man movies: Stimulus kit designed to elicit manner, path and causal constructions in motion events with regard to speech and gestures", Nijmegen, the Netherlands: Max Planck Institute for Psycholinguistics, Language and Cognition Group, 2001.

[8] Slobin, D. I., "Verbalized Events: A Dynamic Approach to Linguistic Relativity and Determinism", in S. Niemeier and R. Dirven [Eds.], Evidence for linguistic relativity, 107–138, Amsterdam: John Benjamins, 2000.

[9] Kita, S., "How representational gestures help speaking", in D. McNeill [Ed.], Language and Gesture, 162-185, Cambridge: Cambridge University Press, 2000.

[10] Levelt, W. J. M., Speaking: From Intention to Articulation. Cambridge, MA: MIT Press, 1989.

[11] Bock, K. J., "Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation", Psychological Review, 89(1): 1–47, 1982.

[12] Talmy, L., Toward a Cognitive Semantics, vol. II: Typology and Process in Concept Structuring. Cambridge, MA: MIT Press, 2000.

[13] Slobin, D. I., "Language and Thought Online: Cognitive Consequences of Linguistic Relativity", in Gentner, D. and Goldin-Meadow, S. [Eds.], Language in Mind: Advances in the study of language and thought, 157-192, Cambridge, MA: The MIT Press, 2003.

[14] WDR (Westdeutscher Rundfunk), Die Sendung mit der Maus. 1971-2014.

[15] Lausberg, H., and Sloetjes, H., "Coding gestural behavior with the NEUROGES-ELAN system", Behavior Research Methods, 41(3): 841-849, 2009.

[16] Özyürek, A., Kita, S., Allen, S., Brown, A., Furman, R. and Ishizuka, T., "Development of cross-linguistic variation in speech and gesture: Motion events in English and Turkish", Developmental Psychology, 44(4): 1040-1054, 2008.

9. Appendix

Table 2 *Given particle verbs in German and English to describe the motion events depicted in the stimuli for the 13 trials used in the study (The particle verb "bore through" was given twice – once it depicted the mouse boring through a globe and once boring through the ground.)*

English	German
climb up (example trial)	hinaufklettern
ride around (practice trial)	herumfahren
bore through	durchbohren
dance around	herumtanzen
jump over	drüberspringen
roll into	hineinrollen
float down	hinunterschweben
jump into	hineinspringen
drill down	hineindrehen
slide down	hinunterrutschen
climb up	hinaufklettern
spin up	hinaufdrehen
bore through	durchbohren
jump around	herumhüpfen
crawl out	hervorkriechen

Individual differences in spatial ability influence the effect of gesturing on navigation and spatial memory

Alexia Galati[1], *Steven M. Weisberg*[2,3], *Nora S. Newcombe* [2], *& Marios Avraamides*[1,4]

[1] Department of Psychology, University of Cyprus, Nicosia, Cyprus
[2] Department of Psychology, Temple University, Philadelphia, USA
[3] Center for Cognitive Neuroscience, University of Pennsylvania, Philadelphia, USA
[4] Centre for Applied Neuroscience, University of Cyprus, Nicosia, Cyprus

galati@ucy.ac.cy, stweis@mail.med.upenn.edu, newcombe@temple.edu, mariosav@ucy.ac.cy

Abstract

Does producing gestures while studying routes facilitate navigation? Thirty-six participants studied route descriptions, producing congruent gestures for one route and keeping their hands still for another. When gesturing, participants made more errors and took longer to navigate the route in a virtual environment. Despite this surprising decrement in navigation performance, gesturing did not impair memory: at least for one route, gesturing actually led to better memory, particularly for navigators with lower spatial ability scores. Overall, the effects of gesturing are selective, depending on the complexity of the described route, the navigators' spatial abilities, and their previous gesturing strategies.

Index Terms: gesture, navigation, spatial memory, route learning, individual differences

1. Introduction

We frequently have to follow route directions to navigate in an unfamiliar environment. These directions may be provided by a friend guiding us over the phone, may be presented incrementally by a GPS device, or may be printed on an invitation. When having the opportunity to study route directions prior to navigation, a number of factors influence how well we represent those routes in memory. Viewing an accompanying map, for example, can help performance, especially when the spatial relations that are encoded visually in the map from a survey perspective are accompanied by linguistic descriptions from a route perspective, which encodes spatial relations relative to the navigator (e.g., "turn left" or "merge right") [1]. Other representational devices, such as arrows, have also been thought to afford schematic features useful to navigation [2] (e.g., for directional turns), and are indeed leveraged by GPS systems to supplement visual information from maps and linguistic directions.

In this study, we examine whether self-generated gestures produced when studying route directions can enhance people's representations of the to-be-navigated route in memory and, by extension, their subsequent navigation performance. Such self-generated gestures may help people construct a spatial representation of the described route by reinforcing, through their schematic features, relevant route information (e.g. directional turns). This facilitative effect could be due to sensorimotor activation from gestures that are schematically congruent with imagined turns or environmental features, resulting in improved situational representations retaining the semantic content or gist of descriptions of the environment (i.e., "situation models), in line with proposals that readers engage in experiential simulation [3].

1.1. The role of gestures in spatial tasks

A confluence of studies underscores the recruitment of gestures in spatial tasks. Speakers often gesture when they provide route directions [4], describe the location of objects in scenes [5], spatial patterns [6], or motion in space [7]. Gestures can reveal the speaker's underlying viewpoint for conceptualizing spatial information [8]. Self-generated gestures can even help in non-communicative contexts without any accompanying speech, as in tasks requiring spatial transformations, including problems that involve mental rotation [9], problems that require making spatiomotor inferences about actions [10] or about spatial relations in described environments [11]. Based on such findings, gestures have been proposed to facilitate spatial visualization and to highlight spatiomotor information for problem solving [10].

Similar to these non-communicative spatial contexts that engender gesturing, gestures in preparation for navigation may improve the spatial representation of the to-be-navigated route. In a recent study, participants who studied routes presented in diagrams recalled more steps of the route when they had gestured during an intervening rehearsal phase compared to other rehearsal conditions that included drawing the route on paper, tracing the route by hand, or mentally simulating it without hand movement [12].

Here, we address explicitly whether gestures confer a benefit to navigation *and* memory performance, not when routes are depicted through diagrams, but rather when they are presented as linguistic directions. We also investigate whether the influence of gestures on navigation and memory performance interacts with the gesturer's spatial abilities. We address these possible interactions in the next section.

1.2. Individual differences in navigation and gesturing

Two relevant lines of research are pertinent here: (i) research examining how spatial and other abilities are related to gesturing and (ii) research addressing the relationship between spatial abilities and navigation performance.

In terms of the former, there is extensive evidence that individual differences in spatial and verbal abilities are associated with differences in gesture production. For example, a combination of spatial and verbal abilities predicts gesture frequency, with the most frequent gesturers being individuals with high spatial visualization ability but low phonemic fluency (i.e., the ability to organize ideas into a chain of linguistic units, associated with executive control) [13]. Differences in fluid intelligence (the ability to select task-relevant information quickly and to focus on a limited set of task-relevant operations) also predict aspects of gesturing:

those with high fluid intelligence are more likely to produce gestures from a non-egocentric perspective in their explanations of geometric analogies compared to those with average fluid intelligence [14]. Moreover, when gestures are recruited in dual tasks, individual differences in working memory capacity influence the extent to which gestures aid performance. When having to recall a series of letters, those with low working-memory capacity benefitted from being able to gesture during an intervening explanation of how they solved a mathematical equation, whereas those with high capacity didn't benefit from gesturing [15]. Finally, differences in expertise in domains with a high spatial component (e.g., neuroscience or meteorology) are also related to differences in producing gestures, with experts being more likely to encode spatial transformations through representational gestures relative to beginners [16].

A second line of research suggests that performance in tasks that require some form of survey knowledge, including navigation, is predicted by individuals' self-reported sense of direction. For instance, sense of direction predicts participants' ability to integrate spatial information from different routes, whether routes are experienced passively while viewing a video [17] or are actively navigated in a virtual environment [18]. Also, individuals with better self-reported sense of direction can more accurately and efficiently navigate routes when they are described from a new perspective (e.g., route vs. survey) that differs from the perspective to which they had been accustomed [19]. Spatial ability may therefore be an important predictor for the effectiveness of gesturing as a learning strategy, since it may render individuals more or less likely to recruit gestures when translating linguistic descriptions into a situational representation. Indeed, beyond spatial abilities, individual preferences for navigation strategies influence the information that navigators encode and can, thus, recall (e.g., [20]).

1.3. Our study

In the present study, we examine whether the self-generated gestures that people produce when studying route directions in preparation for navigation (a) help their initial spatial representation of the to-be-navigated route, thus facilitating navigation performance, (b) help their resulting memory representation for the environment following navigation, and (c) differ in their influence on navigation and memory performance according to the gesturers' spatial abilities.

At the beginning of the study, participants completed self-report and psychometric measures intended to capture individual differences in spatial ability. Next, participants studied directions describing routes from a start point to a destination. This study phase occurred in one of two conditions: for one route participants were instructed to perform gestures congruent with the described path (Gesture), whereas for the other route they were instructed to keep their hands still (No Gesture), with their order counterbalanced across participants. Next, participants navigated those routes from memory in a virtual environment, and finally performed two memory tests that assessed their memory of the environment. This procedure was repeated for the second route.

Based on the reviewed findings that gestures can confer an advantage in spatial tasks, we expected that participants would navigate routes more accurately and efficiently and would remember the navigated environment better when gestures were permitted at study compared to when they weren't. In addition, we predicted that those with better spatial ability would generally perform better during navigation and on the memory tests for the environment. Nevertheless, our investigation of the potential interaction between spatial ability and gesturing was more exploratory. One possibility is that high spatial ability participants would benefit more from gesturing, because they are more likely to gesture spontaneously (e.g., [16]), but another possibility is that those with more limited spatial and related abilities stand to gain more from gesturing (e.g., [18]).

2. Method

2.1. Participants

Thirty-six undergraduate and graduate students from the University of Cyprus (29 female) participated for research credit for a university course or for payment (15 euros).

2.2. Materials

2.2.1. Route descriptions

Routes were described as a series of numbered steps connecting four landmark buildings. Because the two routes connected buildings in a preexisting VE (see [21], and [18]), their descriptions were not fully equivalent. As shown in Figure 1, Route 1 (the red route) was slightly more complex, involving 7 described turns (vs. 5) and 12 distinct segments of text in the directions (vs. 10) relative to Route 2. These differences permitted examining the effect of route complexity, while still matching the routes in terms of the number of landmark buildings, the number of buildings that were intervisible on the route, and the number of spatial locatives (e.g., *left*, *right*, *straight*) in their descriptions.

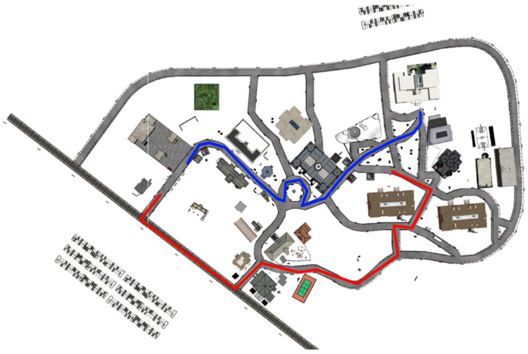


Figure 1: Aerial view map of the layout of buildings with two routes (solid lines) in the VE. Route 1 is shown in red and Route 2 in blue.

2.2.2. Psychometric and Self-report Measures

Participants completed the following self-report and psychometric measures: the Santa Barbara Sense of Direction test (SBSOD; [22]), the Philadelphia Spatial Ability Scale (PSAS; [23]), the Philadelphia Verbal Ability Scale (PVA; [22]), and the Spatial Orientation Test (SOT, [24]).

The SBSOD is a standardized self-report scale of 15 items designed to assess the ability to carry out tasks at the environmental scale of space (e.g., "I am very good at judging distances"). The reported analyses are based on that subset of 10 of those items (the SBSOD-CY scale), due to earlier work suggesting that that only 10 of the SBSOD items are suitable for measuring SOD in the Greek-Cypriot population [25].

The PSAS scale includes 16 items designed to measure how well participants feel they can perform small-scale spatial tasks, such as visualizing and transforming small or medium-sized objects (e.g., "I can easily visualize my room with a different furniture arrangement").

The PVAS scale consists of 10 items designed to measure how strong participants feel their verbal ability is (e.g., "I am good at crossword puzzles"). For these three self-report measures, participants responded on a 7-point Likert scale.

The SOT consists of 12 test items presenting participants with an array of objects and asking them, on a given item, to locate an object from an imagined perspective (e.g., *Imagine you are standing at the car and facing the traffic light. Point to the stop sign*). Participants draw its angle of disparity from their imagined perspective on a circle on the printed page. Participants were timed for 5 minutes to complete as many of the items as they could. Each participant's error score was computed by averaging, across items, the difference between the angle for the correct answer and their response.

2.3. Procedure

Upon giving informed consent, participants completed the self-report and psychometric measures. The SBSOD, PSAS, and PVAS were translated into Greek and were presented on a browser, using SurveyMonkey Inc. services, while the SOT was administered on paper. Next, participants were familiarized with the VE presented on a projection screen, and became accustomed to using the controls for moving and looking around the VE (a mouse and the arrow keys of a numeric keypad). Next, participants moved to an adjacent room to study a route description (study phase). They were told that they would later navigate the described path from memory in the VE, and were asked to ensure that they remembered the names of the route's four landmark buildings in order to recognize them in the VE. Participants were informed that the route description would appear on the computer screen as a series of numbered instructions that would be "similar to the format of directions from Google Maps, but more detailed". Participants could study route directions without a time limit, and were videotaped.

In the *Gesture* condition, participants were instructed to produce compatible gestures while they studied the route directions. They were asked to produce at least one compatible gesture, for each numbered step of the directions, but were otherwise free to gesture in any way the wished. In the *No Gesture* condition, participants were instructed to hold their hands still as they studied the route directions, keeping their right and left index fingers on specific keyboard keys.

After the study phase, participants returned to the original room to navigate the route from memory in the VE. They were reminded of their origin and destination buildings (e.g., "You at Batty House and you want to go to Golledge Hall."). If participants made a navigation error (e.g., taking the wrong turn) and did not readily self-correct, they were interrupted by the experimenter, directed to an earlier correct segment of the route, and prompted for the next instruction. Prompts became more specific only if the participant reported not being able to recall how to continue. If participants asked for confirmation for their navigation choices, the experimenter did not provide feedback, and asked them to proceed as they thought best.

After navigation, participants completed two tests assessing their memory representation for the virtual environment: a pointing task and a model building task. In the pointing task, for each trial, participants were placed directly next to one of the four landmark buildings of the route and were asked to point to one of the other buildings from that location. The prompt appeared at the top of the screen (e.g., "Point to Harvey House", while being next to Batty House). To respond, participants were instructed to move a crosshair that appeared in the center of the screen, until it pointed to where they imagined the front door of the building in the prompt to be. Participants could rotate the crosshair in the horizontal plane by moving their mouse and clicking to log their response. For each trial, performance was assessed by determining the smallest possible angle between the correct answer and the participant's estimate.

In the model building task, participants viewed on the computer screen a blank box with top-down views of each of the four landmark buildings of the route underneath it. Participants were told that the box represented the entire VE they had explored on that route and were asked to place each building where they considered it to be. Participants could drag and drop buildings using their mouse, adjusting their positions as much as necessary. Accuracy on the model-building task was assessed using a bidimensional regression analyses [26], which correct for differences in scale, translation, and rotation, providing the correlation coefficient between the configuration of the target map and the participant's map. The correlation coefficient squared (R^2) was the variable of interest, capturing the proportion of variance explained in the actual layout of buildings by the participant's arrangement of buildings.

Participants then completed the same procedure (study, navigation, pointing task, model building task) for a second block, in which they studied the other route in the other gesture condition (Gesture or No Gesture). After completing this series of tasks for both routes, participants were debriefed. Experimental sessions took about 1.5 hours.

2.4. Coding navigation performance

Navigation videos, captured by Fraps software, were annotated in ELAN [27] to assess the duration to transverse the route, the number of navigation errors made, and the length and frequency of their pauses. The onset of *route duration* was operationalized as the first video frame of movement at the origin of the route, and its offset as the final frame of movement (forward, backward, or lateral) at the destination building. The *navigation errors* of interest were wrong choice point errors, in which navigators deviated from the route at a decision point (e.g., a turn, intersection, crossroad, or forked road). *Pauses* were identified as the segments of the video on which the navigator was stationary, without any movement (forward, backward, or lateral) for two or more frames (i.e., sequences of video frame across which the optic flow either remained unchanged or suggested a change in heading due to rotation but not displacement). In order to control for differences in route duration, we analyzed the proportion of the route's duration that navigators spent pausing (i.e., total duration of all pauses / route duration).

2.2.1. Reliability for navigation coding

Two coders coded uniquely the videos of 15 and 17 participants, respectively, and coded redundantly the videos of another 4 participants. Their estimates of route duration exhibited high reliability in the 8 videos coded redundantly: the single measure of intraclass correlation coefficient (ICC*)* was 1.00, $p < .001$, using type consistency. The coders had a mean difference only of about a video frame (38 msecs) when identifying the end of a route. Their difference in identifying the onset of the route averaged .51 secs, due to one outlying video with a disagreement of 2.80 secs. The two coders also identified the same 18 navigation errors in the routes of the 8 videos. Their only disagreement concerned one instance in which one coder identified a navigation error that the other coder parsed as two consecutive errors. The inter-rater agreement for identifying wrong choice point errors was 95% (Cohen's Kappa = .81, $p <$ 01). The total duration of pauses per route was highly correlated between the two coders as well, ICC= 1.00, $p < .001$. The mean difference in total pause duration was 1.83 secs per route (*SD*= .43 secs), corresponding to only a small fraction of the route's total duration (.68 %).

3. Results

3.1. Navigation performance

Gesturing during the study of route directions did not improve navigation performance overall. In fact, in terms of their efficiency in navigating the route, participants took numerically longer to complete the routes when they had previously gestured (*M*=356.19 secs, *SD*= 153.72 secs) compared to when they hadn't gestured at study (*M*=326.17 secs, *SD*= 136.96 secs), *F* (1, 34)= 1.50, *p*= .23. This numerical difference in route duration can be contextualized by the number of errors participants made during navigation: participants made more navigation errors when they had gestured at study compared to when they hadn't gestured (*M*=2.03, *SD*=1.18 vs. *M*=1.58, *SD*=1.23), although this difference was only marginally significant, *F* (1, 34)= 3.46, *p*= .07. Whether navigators gestured at study did not influence their pausing behavior, *F* (1, 34)= .15, *p*= .70.

When comparing the two routes, navigators took numerically longer (by an average of 34 secs) to complete the more complex route, Route 1, *F* (1, 34)= 1.98, *p*= .17, although they paused for a significantly greater proportion of time in Route 2 (*M*= .36, *SD*= .12, vs. *M*= .29, *SD*= .13 for Route 1), *F* (1, 34)= 11.92, *p* < .01. There were no reliable differences in the navigation errors made across the two routes, *F* (1, 34)= .49, *p*= .49. For none of the above measures did the combination of the gesture condition with route identity or its order influence performance.

3.2. Navigation performance relative to individual differences

Although we didn't find a systematic effect of gesturing on navigation performance, some consistent patterns emerged when individual differences in spatial ability were considered. In general, navigation performance was better for participants with higher spatial ability scores. For instance, the mean duration of both routes taken together was significantly correlated with participants' mean SOT error (Pearson's *r*= .50, *p* < .01) and was marginally correlated with their PSAS score (Pearson's *r*= -.27, *p*= .11). Similarly, the participants' SOT error was significantly correlated with their number of navigation errors (Pearson's *r*= .51, *p* < .01) and with the proportion of the route spent pausing (Pearson's *r*= .37, *p* < .05).

Interestingly, many correlations between spatial ability and navigation performance were reliable in the No Gesture condition but not in the Gesture condition (e.g., SBSOD-CY with route duration and with navigation errors; SOT with navigation errors and with the proportion of time paused; PSAS with the proportion of time paused). That is, constricting gestures had an adverse effect on the navigation performance of individuals with lower spatial ability.

3.3. Memory performance: Pointing task

In the pointing task, gesturing at study did not influence participants' mean pointing error, *F* (1, 34)= 1.73, *p*= .20. Their mean pointing error was 42.45° (*SD*= 37.58°) when they had gestured at study and 45.04° (*SD*= 38.72°) when they had not. Although gesturing did not influence pointing performance on its own, its influence depended on the complexity of the route with which it was paired. As shown in Figure 2, gesturing improved pointing accuracy more so for the less complex route, Route 2, than for Route 1; the interaction between gesturing and the route with which it was paired was significant, *F* (1, 34)= 32.17, *p* < .001. Not surprisingly, participants were overall more accurate on the less complex route, Route 2 (*M*= 38.19°, *SD*= 36.69°) than Route 1 (*M*=49.30, *SD*= 36.39°), *F* (1, 34)= 32.17, *p* < .001.

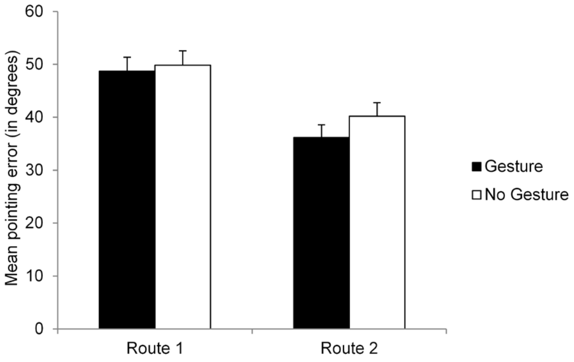


Figure 2: Participants' mean pointing error across the study condition (Gesture vs. No Gesture) and the route studied (Route 1 vs. 2). Error bars represent standard error of the mean.

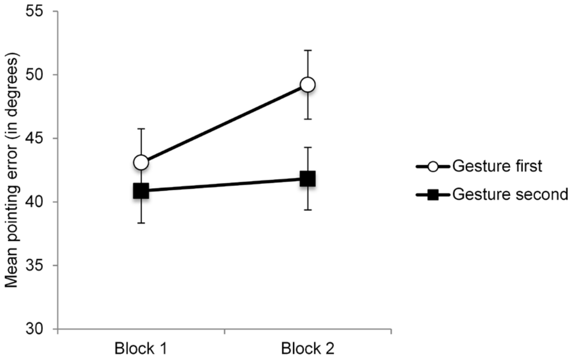


Figure 3: Participants' mean pointing error across the two blocks in the experiment (Block 1 vs. Block 2) and the order of the study conditions (Gesture-first vs. Gesture-second). Error bars represent standard error of the mean.

Finally, the ordering of the study conditions influenced pointing error to some extent, as there was a marginally significant interaction between the study condition and the block in which it took place, *F* (1, 32)= 3.56, *p*= .07. As shown in Figure 3, when the No Gesture condition was in the second block (i.e., the right circle of the "Gesture first" line) performance was worse than when it was in the first block (i.e., the left square of the "Gesture second" line), *F* (1, 34)= 4.05, *p*= .05. That is, participants who couldn't gesture after a block in which they could gesture made larger pointing errors compared to participants for whom that condition came first. This order effect was driven by Route 2, the less complex route. When participants couldn't gesture while studying Route 2, they were about 15° less accurate when this happened the second block than in the first block, *F* (1, 32)= 10.06, *p* < .01. On the other hand, when participants couldn't gesture while studying Route 1, they were comparably accurate whether this happened in the first or second block, *F* (1, 32)= .10, *p*= .76.

3.4. Memory performance: Model building

Altogether, performance in the model building task converged with performance in the pointing task. The correlation coefficients squared (R^2) indicated that gesturing at study (*M*= .72, *SD*=18) did not result in more accurate configuration than not gesturing at study (*M*= .74, *SD*= .20), *F* (1, 32)= .43, *p*= .52. However, the pairing of the study conditions with the routes mattered, as indicated by a significant interaction, *F* (1, 32)= 7.03, *p* < .05. As with the pointing task, when participants couldn't gesture while studying Route 2 they were less accurate, and this was specifically the case when the No Gesture condition was in the second block than the first block, *F* (1, 32)= 4.63, *p* < .05. As with the pointing task, participants were also more accurate on Route 2 (*M*= .78, *SD*= .18) than Route 1 (*M*= .67, *SD*= .18), *F* (1, 32)= 13.04, *p* <. 01.

3.5. Memory performance relative to individual differences

Participants' spatial ability predicted their memory performance in some ways. Performance in the pointing task was marginally correlated with the participants' PSAS and SBSOD-CY scores (Pearson's *r*= -.25, *p*= .15 PSAS; Pearson's *r*= -.27, *p*= .11, respectively): participants with better spatial ability tended to make smaller angular errors in the pointing task. These negative correlations became significant when only pointing performance on Route 2 was considered (for PSAS: Pearson's *r* = - .38, *p* < .05; for SBSOD-CY: Pearson's *r*= -.37, *p* < .05). Interestingly, these significant correlations held when participants couldn't gesture when studying Route 2 (N= 18, for PSAS: Pearson's *r*= -.66, *p* < .01, for SBSOD-CY: Pearson's *r*= -.48, *p* < .05), but not when they could gesture on Route 2.

In terms of model building, participants with higher spatial ability constructed more accurate configurations of the landmark buildings. Specifically, model building performance correlated significantly with PSAS scores (Pearson's *r*= .40, *p* < .05) and marginally with SBSOD-CY scores (Pearson's *r*= .33, *p*= .05). Similar to pointing performance, the correlation between model building performance and SBSOD-CY was significant for Route 2 (Pearson's *r*= .36, *p* < .05) but not for Route 1.

Participants' navigation performance was also correlated with their later performance in systematic ways. Although participants' mean angular error in the pointing task was not significantly correlated with the time they took to complete the route, this correlation became significant when only the No Gesture condition was considered (Pearson's r= .40, *p* < .05). Similarly, in the No Gesture condition, participants constructed less accurate models of the environment as route durations increased (Pearson's *r*= -.36, *p* < .05), whereas this wasn't the case in the Gesture condition. That is, when they hadn't gestured at study, participants made larger pointing errors and constructed less accurate models the longer they had taken to complete the route.

4. Discussion

Contrary to our predictions, gesturing while studying route descriptions did not confer a global advantage to navigation and memory performance. However, when taking the navigators' individual differences in spatial ability into consideration, a more nuanced understanding of the potential benefit of gestures emerges. For instance, although gesturing at study did not improve navigation performance (in fact, it numerically increased navigation errors), those with lower spatial abilities were worse at navigating when gesturing was prevented at study. Measures of spatial ability reliably predicted navigation performance only in the No Gesture condition, with one exception (between SOT and the mean route duration, whose significant correlation held for both study conditions). In other words, preventing gestures during the study phase was especially pernicious to those with lower spatial abilities. For these individuals, constricting gestures at study may have contributed to a less accurate initial representation of the route and the environment compared to when they had been instructed to gesture.

This proposal is supported by the memory tests, on which lower ability navigators performed worse, though in a more restricted context. When gestures were constricted, there was a decrement in pointing performance for the less complex of the two routes (Route 2), on which participants were overall more accurate. It is not fully clear why these effects of gesturing (or not gesturing) are observed only for Route 2, since performance on Route 1 does not seem to reflect a floor effect any more so than performance on Route 2. In both the pointing and model building tasks, performance on this route was worse when gestures were constricted in the second block (i.e., after having used a gesturing strategy), especially for navigators with lower spatial abilities. This order effect is in line with other studies reporting a performance cost when switching from a gesture to a no gesture condition (e.g., [15]).

When interpreting the findings of the navigation and testing phases, it is useful to distinguish the representations of the environment that participants accessed in each of these phases. During navigation, participants presumably accessed an *initial* representation of the environment; this representation had, as its input, the linguistic descriptions provided at study and, in the Gesture condition, self-generated gestures that presumably elaborated or reinforced their initial situation model. During memory testing, participants accessed their *final* representation of the environment, which had been enriched and updated by the visual information experienced during navigation in the virtual environment. This distinction qualifies some patterns that may appear perplexing otherwise: for instance, that Route 2 exhibited worse navigation performance in some ways but better memory performance. The fact that participants paused proportionally longer when navigating Route 2 (vs. Route 1) may have enabled them to create a more accurate representation of the environment along that route, resulting in more accurate pointing judgments and model reconstructions later on.

However, what remains puzzling is that, although constricting gestures impaired the navigation and memory performance of low ability navigators, gesturing led to overall more navigation errors compared to not gesturing (albeit, this was a marginally significant difference) and numerically longer route durations. One possibility for this counterintuitive finding is that *forcing* participants to gesture taxes their cognitive or attentional resources, and thus impairs their encoding of the environment. Other researchers have not found evidence in support of this claim, reporting no reliable differences between the effects of forced and spontaneous gesturing (e.g., [28] in a dual task). Still, in our task, it is possible that there could be an adverse effect of forced gesturing.

Another possibility is that, by gesturing while encoding route descriptions, readers may reinforce somewhat inaccurate inferences about an unfamiliar environment. Linguistic directions convey spatial information through discrete units that do not capture analogue or gradient spatial relationships. For example, readers may interpret the description of a "left turn" as a canonical 90° left turn, when in fact it may refer to a more oblique turn (say, 75°) in the environment. Thus, when readers are trying to construct situation models for unfamiliar environments, asking them to produce compatible gestures could reinforce more canonical representations for some aspects of the environment that may conflict somewhat with their perceptual experience during navigation, sufficiently so to result in navigation errors. This suggests that, although constricting gestures when studying route directions may be particularly harmful to low spatial ability individuals, forcing navigators to gesture on each instruction may not be an ideal encoding strategy overall. Although in this experiment we did not code for the type and frequency of gestures produced (having simply checked that participants behaved as instructed in the Gesture and No Gesture conditions), we are doing so in a follow-up experiment.

In this new experiment, we are aiming to shed light on the reported decrement in navigation performance here by letting participants gesture spontaneously instead of instructing them to gesture. If spontaneous gesturing *is* beneficial to navigation performance, the numerical decrement in navigation performance observed here could have been due to forced gestures taxing participants' cognitive resources. However, if spontaneous gesturing continues not to improve navigation performance, then it may be a non-ideal strategy for encoding routes in unfamiliar environments, as it could contribute to more schematic and slightly inaccurate spatial representations. With 2/3 of these new data coded, more than half of the gestures encode representational features of the environment or route (8.41 out of the 15.25 gestures produced on average per minute), involving a route or survey perspective, or their combination. Examining the distribution of gesture types and their frequency will be useful to unveiling the strategies that navigators employ at study and their relationship to the navigators' spatial abilities and subsequent performance.

So far, the effect of gesturing on navigation and memory performance appears to be selective, depending on the complexity of the described route, the spatial abilities of the navigators, and their previous learning strategies (e.g., the prior availability of gestures).

5. Acknowledgements

This material is supported by a SAVI grant from NSF through the Spatial Intelligence and Learning Center (SILC)'s Thematic Network in Spatial Cognition (TNSC). We are grateful to Philippos Panagiotou and Georgia Leonidou for assistance with data collection and coding.

6. References

[1] Brunyé, T. T., Rapp, D. N., & Taylor, H. A. (2008). Representational flexibility and specificity following spatial descriptions of real-world environments. *Cognition*, *108*, 418–443.

[2] Tversky, B. and Lee, P.U. (1998). How space structures language. In Freksa, C., Habel, C., and Wender, K.F., Eds., *Spatial Cognition: An Interdisciplinary Approach to Representation and Processing of Spatial Knowledge*. Berlin: Springer-Verlag, pp. 157-175.

[3] Zwaan, R.A. (2004). The immersed experiencer: toward an embodied theory of language comprehension. In B.H. Ross (Ed.), *The Psychology of Learning and Motivation*, *44*, (pp. 35-62). New York: Academic Press.

[4] Allen, G. (2003). Gestures accompanying verbal route directions: Do they point to a new avenue for examining spatial representations? *Spatial Cognition and Computation*, *3*, 259–268.

[5] Tutton, M. (2013). A new approach to analysing static locative expressions. *Language and Cognition*, *5*, 25-60.

[6] Melinger, A., & Kita, S. (2007). Conceptualisation load triggers gesture production, *Language and Cognitive Processes*, *22*, 473-500.

[7] Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*, 16-32.

[8] Emmorey, K., Tversky, B., & Taylor, H. (2000). Using space to describe space: Perspective in speech, sign, and gesture. *Spatial Cognition and Computation*, *2*, 157-180.

[9] Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General*, *140*, 102-115.

[10] Alibali, M. W., Spencer, R. C., Knox, L., & Kita, S. (2011). Spontaneous Gestures Influence Strategy Choices in Problem Solving. *Psychological Science*, *22*, 1138-1144.

[11] Jamalian, A., Giardino, V., & Tversky, T. (2013). Gestures for thinking. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 645-650). Austin, TX: Cognitive Science Society.

[12] So, W.C., Ching, T. H-W., Lim, P. E., Cheng, X., & Ip, K. Y. (2014). Producing Gestures Facilitates Route Learning. *PLoS ONE*, *9*(*11*): e112543. doi:10.1371/journal.pone.0112543

[13] Hostetter, A. B., & Alibali, M. W. (2007). Raise your hand if you're spatial: Relations between verbal and spatial skills and representational gesture production. *Gesture*, *7*, 73–95.

[14] Sassenberg, U., Foth, M., Wartenburger, I. & van der Meer, E. (2011). Show your hands—are you really clever? Reasoning, gesture production, and Intelligence. *Linguistics, 49*, 105-134.

[15] Marstaller, L., & Burianova, H. (2013). Individual differences in the gesture effect on working memory. *Psychonomic Bulletin & Review*, *20*, 496-500.

[16] Trafton, J. G., Trickett, S. B., Stitzlein, C. A., Saner, L., Schunn, C. D., & Kirschenbaum, S. S. (2006). The relationship between spatial transformations and iconic gestures. *Spatial Cognition and Computation*, *6*, 1-29.

[17] Muehl, K. A., & Sholl, M. J. (2004). The acquisition of vector knowledge and its relation to self-rated direction sense. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 129–141.

[18] Weisberg, S. M., Schinazi, V. R., Newcombe, N. S., Shipley, T. F., Epstein, R. (2014). Variations in cognitive maps: Understanding individual differences in navigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 669-682.

[19] Ishikawa, T. & Kiyomoto, M. (2008). Turn to the left or to the west: Verbal navigational directions in relative and absolute frames of reference. In T. J. Cova, H. J. Miller, K. Beard, A. U. Frank, & M. F. Goodchild (Eds.), *Geographic information science: LNCS 5266* (pp. 119-132). Berlin: Springer.

[20] Marchette, S.A., Bakker, A., & Shelton, A.L. (2011). Cognitive mappers to creatures of habit: Differential engagement of place and response learning mechanisms predicts human navigational behavior. *Journal of Neuroscience*, *31*, 15264-15268.

[21] Schinazi, V. R., Nardi, D., Newcombe, N. S., Shipley, T. F., & Epstein, R. A. (2013). Hippocampal size predicts rapid learning of a cognitive map in humans. *Hippocampus*, *23*, 515–528.

[22] Hegarty, M. Richardson, A. E., Montello, D. R., Lovelace, K & Subbiah, I. (2002). Development of a Self-Report Measure of Environmental Spatial Ability. *Intelligence*, *30*, 425-447.

[23] Hegarty, M., Crookes, R. D., Dara-Abrams, D., & Shipley, T. F. (2010). Do all science disciplines rely on spatial abilities? Preliminary evidence from self-report questionnaires. In D. Hutchison et al. (Series Eds.), *Lecture Notes in Computer Science: Spatial Cognition VII: International Conference, Spatial Cognition 2010, Mt. Hood/Portland, OR, USA, August 15-19, 2010: Proceedings* (Vol. 6222, pp. 85–94). Berlin, Germany: Springer.

[24] Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective- taking spatial abilities. *Intelligence*, *32*, 175-191.

[25] Shimi, A., Avraamides, M.N., & Fanti, K. (2008). The construct validity of the Santa Barbara sense of direction scale in the Greek-Cypriot population. *XXIX International Congress of Psychology*, Berlin, Germany.

[26] Friedman, A., & Kohler, B. (2003). Bidimensional regression: Assessing the configural similarity and accuracy of cognitive maps and other two-dimensional data sets. *Psychological Methods*, *8*, 468-491.

[27] Brugman, H., & Russel, A. (2004). Annotating Multi-media / Multi-modal resources with ELAN. *LREC.*

[28] Goldin-Meadow, S., Nusbaum, H., Kelly, S., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*, 516-522.

Eyebrows in French talk-in-interaction

Aurélie Goujon [1], *Roxane Bertrand* [1], *Marion Tellier* [1]

[1] Aix Marseille Université, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France

Goujon.aurelie@gmail.com Roxane.bertrand@lpl-aix.fr Marion.tellier@lpl-aix.fr

Abstract

Conversational facial gestures can be considered as co-verbal thanks to their timing with words and context. In particular, a relationship between eyebrow movements and speech, on both prosodic and conversational levels has been found. The originality of the present study is twofold: it deals with eyebrow movements in three different corpora of French talk-in-interaction and it involves the same pair of speakers in each condition. The aim of this study is 1/ to compare the production of eyebrow movements in different speaking tasks specific and 2/ to establish the location of eyebrow movements according to speaking turns. Our results show that the different corpora exhibit a significant difference related to the production of eyebrow movements while the latter significantly co-occur with the beginning of speaking turns, whatever the type of corpus.

Index Terms: Eyebrow movements, conversational facial gestures, interaction, multimodality.

1. Introduction

Facial expressions have been regarded since [1] as expression of emotions [2]. However, in 1979, [3] introduced the distinction between emotional facial gestures and conversational facial gestures. This distinction is updated by [4] with four criteria: conversational facial gestures are context-dependent, speech-dependent, without "stereotype" (no fixed forms), and appear in a social process. This function of facial expressions in conversation is consistent with a multimodal perspective of speech as developed in our corpora. Therefore, some gestures such as hand gestures, head movements and facial expressions can be considered as co-verbal since they occur during speech and they cannot be analyzed without it.

In this preliminary study, we focus on eyebrow movements in three types of talk-in-interaction corpora in French. We adopt an approach based both on interaction studies ([5], among others) and gesture studies. We analyze the production and location of eyebrow movements of the same pair of participants in three interactional tasks. Although studies focusing on eyebrow movements are scarce, we know that they are heavily connected to speech. Moreover, to our knowledge, there is no previous study (such as ours) aiming to compare eyebrow movements in three types of interactional data.

Eyebrow movements are strongly connected to speech, both on the prosodic level and on the conversational level. On the prosodic level, [6] claimed that 93.75% of the eyebrow movements of their data were associated with accentuating intonation contours. Indeed, they showed a tiny link between fundamental frequency and eyebrow movements. On the conversational level, [7] showed that a raised eyebrow structures the start, the continuity and the end of a topic in a conversation. More precisely, eyebrow movements have been predicted to occur more frequently at the start of a new segment in the structure of the dialogue [8]. In French, [6] and [9] argue that eyebrows movements are associated with the beginning of a new speaking turn. The authors demonstrated these results thanks to time measurement between speech and gestures. However, few studies have focused on the French language. One of the purposes of our paper is to confirm the results of [9] by using an automatic request procedure and by expanding the study to other kinds of interactional activities.

Moreover, studies concerning eyebrow movements were conducted on a single type of interaction (interview) involving different speakers. The main interest of our corpora is that they involved the same pair of speakers across three different types of talk-in-interaction. Thus, our first research question is: **Does the interactional task impact the production of eyebrow movements in terms of number of occurrences?** Our second research question is focused on the timing: **Does the link between the beginning of the speaking turn and the occurrence of eyebrow movements remain the same whatever the interactional task?** For the first question we hypothesise that there will be a difference in the production of eyebrow movements since there is 1/ a difference in the various activities engaged in tasks and 2/ a difference in the way the various tasks involve the participants (personal opinions vs. factual information for example). For the second question we will attempt to corroborate the results of [9] about the link between the beginning of the speaking turn and the eyebrow movements. We therefore hypothesize that the different types of corpus will have no effect on the location of eyebrow movements. Whatever the interactional task, the function of eyebrow movements in the structuring of the turn taking organisation will remain the same and thus eyebrow movements should be more present at the beginning of a speaking turn than in another position.

2. Methodology

2.1 Participants and experimental settings

Participants are a pair of two male speakers (AG and YM). The two speakers work in the same laboratory and have known one another for years. They are friends. They are also familiar with the anechoic chamber where the interactions were recorded. In the CID condition, participants take place in three-quarters view. In the DVD and MTX conditions, participants are in a face-to-face interaction.

2.2 Corpus

Our study is based on three corpora that were recorded at the LPL (Laboratoire Parole et Langage, Aix-en-Provence). Corpora investigated here were the CID (Corpus of Interactional Data) [10] (visible at *sldr.org/sldr000720/en*), the DVD Corpus (CoFee Project, [11]) (visible at *sldr.org/sldr000773/en*) and the Maptask corpus (MTX) [12] (visible at *sldr.org/sldr000732/en*). The CID is a corpus of conversational data in which the participants had to freely discuss an imposed topic: unusual moments in life. Given this instruction, the CID exhibits a large narration activity. The DVD corpus is a corpus of argumentative speech about movies. This interaction exhibits a large activity in which participants have to express personal opinions. The DVD corpus also contains an activity of negotiation required by the global task of demanding that each speaker choose two DVDs to take home at the end of the task. The MTX corpus is a collaborative task in which speakers are directors or followers round after round and they have to map out a path on paper [13]. Participants realized seven map tasks. The MTX corpus exhibits a large explanation activity. Given the different interactional tasks particular to our different corpora - the DVD corpus is typically conducive at the expression of opinion; the CID corpus requires narrative and descriptive speech; and the MTX corpus presents factual information concerning the direction and the location of elements on the map - we can test whether eyebrow movements from the same pair of participants are different according to these particularities.

The selected extracts for this study were of equal duration (i.e. 30 minutes for each corpus).

2.3 Unit of analysis

All three corpora were segmented in Inter-pausal Units (IPU) which are speech blocks separated by silent pauses (>200ms). Following [14] we use IPU as a unit of turn. Other units based on syntactic or prosodic cues can be used to refer to a turn but their identification remains difficult, especially on spontaneous data. Our choice of IPU as a unit of analysis is based on its objective nature, which makes its identification easier. We therefore considered each IPU produced by a speaker as a new speaking turn.

2.4 Annotation and request

The study of eyebrow movements requires the use of multimodal annotation software. We chose ELAN (v4.7.3 [16]) and SPPAS [17]. The methodology used in this paper is comprised of 3 stages:

2.4.1 Pre-segmentation of eyebrows

The pre-segmentation of eyebrow movements was realised with the segmentation mode of ELAN. Only eyebrow raises and frowns were annotated. It is necessary to watch the video at a slowed rate to find the exact image that corresponds to the eyebrow movement. If it is an eyebrow-raising movement, the eyebrows gradually rise up and go back to a neutral position on a vertical axis. If it is a frowning movement, the eyebrows move on a horizontal axis. Eyebrows move toward each other near the center and a bend appears between them. It is important to note that the annotated movements can be on a single eyebrow.

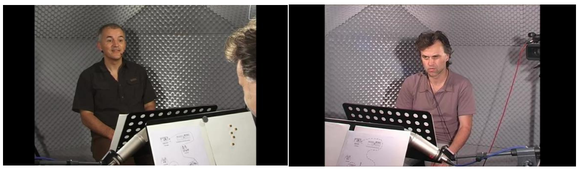
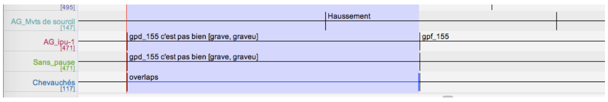


Figure 1: Example of eyebrow movements: raised (on the left) and frown (on the right)

2.4.2 Transcription of speech

The transcription of verbal speech in ELAN (see Figure 2) takes place after the segmentation of eyebrow movements. This is a second stage to avoid the influence of speech when annotating eyebrow movement.

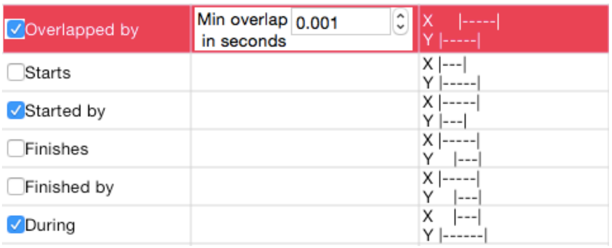


[495]
AG_Mvts de sourcil [147] — Haussement
AG_ipu-1 [471] — gpd_155 c'est pas bien [grave, graveu] — gpf_155
Sans_pause [471] — gpd_155 c'est pas bien [grave, graveu]
Chevauchés [117] — overlaps

Figure 2: Example of an overlap

2.4.3 Filters

The silent pauses and overlaps of eyebrow movements with IPUs were filtered in SPPAS in order to see how many eyebrow movements were produced along with speech. For this, we customised the overlap criteria, and chose between: "*Overlaps*", "*overlapped by*", "*starts*", "*started by*", "*finishes*", "*finished by*", "*during*" and "*contains*". Each time an eyebrow movement occurs in these positions, SPPAS creates an annotation on a new tier named "Chevauchés" (Figure 2). In order to see the link between eyebrow movements and the beginning of a speaking turn, we created a new tier named "Chevauchés_deb" containing only "*Overlapped by*", "*started by*", and "*during*" to show the co-occurrence of eyebrow movements and IPUs). In the example given in the Figure 3, we consider that X corresponds to the IPUs tier and Y corresponds to eyebrow movements.



☑ Overlapped by	Min overlap in seconds 0.001	X |-----| Y |-----|
□ Starts		X |---| Y |-----|
☑ Started by		X |-----| Y |---|
□ Finishes		X |-----| Y |---|
□ Finished by		X |-----| Y |---|
☑ During		X |---| Y |------|

Figure 3: SPPAS filter

3. Results

3.1 Production of eyebrow movements

In order to analyse the production of eyebrow movements in the three selected corpora, descriptive statistics and proportion tests were performed.

3.1.1. Descriptive results

The following table (Figure 4) shows the descriptive statistics of the corpus for each of the three corpora: the activity of narration (CID), the activity of expression of opinion (DVD) and the activity of explanation (MTX). It shows the number of IPUs, the total number of eyebrow movements, the eyebrow movements co-occurring with overlapped IPUs and the ones appearing only at the beginning of the IPUs for each speaker (AG *vs* YM).

Speaker AG	DVD	CID	MTX
TOTAL_IPU	531	499	732
TOTAL_EYEBROW	214	172	76
TOTAL_IPU_OVERLAP	159	138	63
TOTAL_IPU_OVERLAP_BEG	83	58	31
IPU_OVERLAP/TOTAL_IPU	10	9.67	3.39
Speaker YM	DVD	CID	MTX
TOTAL_IPU	583	503	652
TOTAL_EYEBROW	210	154	60
TOTAL_IPU_OVERLAP	180	125	55
TOTAL_IPU_OVERLAP_BEG	98	33	22
IPU_OVERLAP/TOTAL_IPU	9.93	8.67	3.09

Figure 4: Table of results

In Figure 4, IPU_OVERLAP/TOTAL_IPU refers to the mean eyebrow movement production per minute. The eyebrow rate is obviously dependent on the number of occurrences but it is important information: for instance if we look at speaker AG in the DVD corpus, he produced a mean of 10 eyebrow movements overlapped with an IPU for 1 minute and speaker YM in the DVD corpus produced a mean of 9.93. The amount of TOTAL_IPU is almost the same in the DVD corpus (AG = 531 vs. YM = 583).

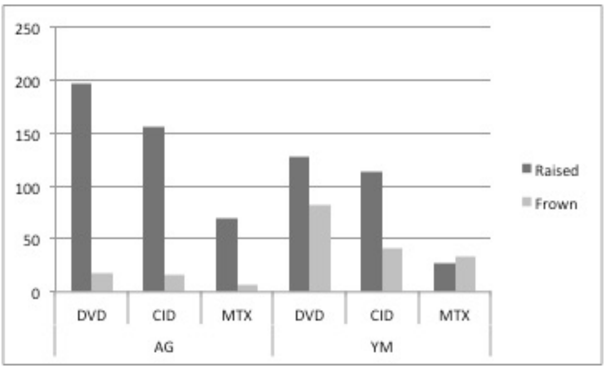


Figure 5: Production of eyebrow movements

Nonetheless, as we can see in Figure 5, the repartition of the two types of eyebrow movements (raising and frowning) is very uneven. AG always produces more eyebrow-raising movements than frowning movements in the three corpora. The production of YM is more balanced than that of AG: he produces a fair amount of eyebrow raises but the frowning is more frequent than for AG.

Figure 5 shows the difference between raised eyebrows (in dark) and frowns (in light), for each corpus and each speaker. The graphic illustrates a difference between each corpus, with a more important frequency of movements in DVD > CID > MTX, whatever the speaker, and whatever the type of movements (raised or frown). Even if YM produces fewer eyebrow movements than AG, the proportion between corpora is constant.

3.1.2. Statistical analysis

In order to test the link between eyebrow movement occurrences and the interactional task, proportion tests were performed [17].

The first proportion test is about the difference of proportion in eyebrow movement occurrences on the total number of IPUs between the three corpora, i.e. TOTAL_IPU_OVERLAP/TOTAL_IPU. One proportion test by speaker was performed. (AG = 159/531; 138/499; 63/732; YM = 180/583; 125/503; 55/652).

	DVD	**CID**	**MTX**
AG	0.30	0.28	0.09
YM	0.31	0.25	0.08

Figure 6: Table of proportion test score

AG (X-squared = 108.5312, df = 2, p-value <2.2e-16*); YM (X-squared = 101.7436, df = 2, p-value < 2.2e-16*). As expected, the proportion of eyebrow movements significantly differs: we can note that the MTX corpus very strongly differs from the DVD and CID which exhibit similar proportions.

3.2. Link between eyebrow movements and speaking turn

In order to confirm the link between eyebrow movements and the beginning of speaking turns, and as we did for the production test, descriptive statistics and proportion tests were performed.

3.2.1. Descriptive results

Our second research question deals with the moment of appearance of eyebrow movement on IPUs. Our hypothesis is that they should appear at the beginning of an IPU, to be consistent with [9]'s findings. We have considered how the total number of eyebrow movements is distributed in Fig 7.

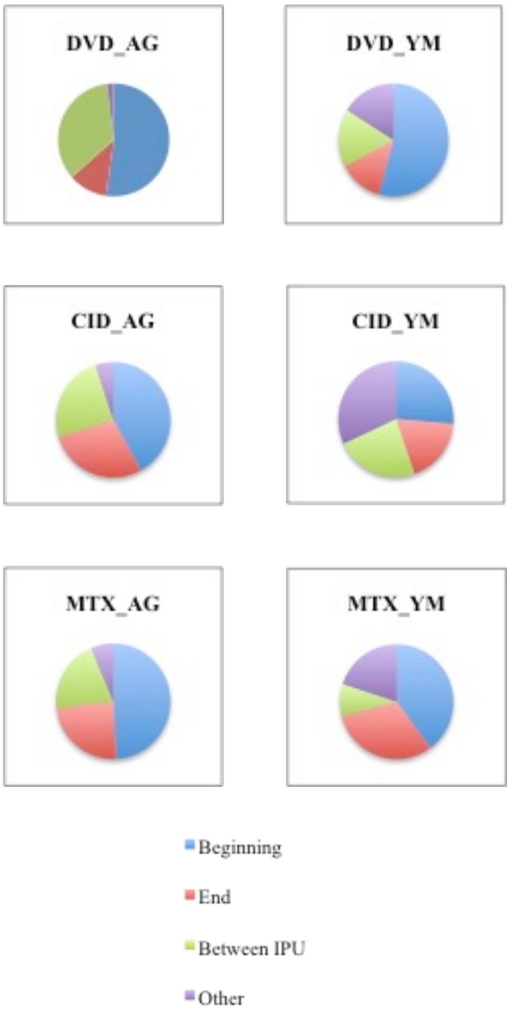


Figure 7: Movements associated with IPU

3.2.2. Statistical analysis

We performed a second test of proportion to confirm that there are no differences between the corpora at the beginning of IPUs. The proportion test is about the difference of proportion about overlaps in beginning of an IPU on the total number of overlapped IPUs between the three corpora, i.e. TOTAL_IPU_OVERLAP_BEG/TOTAL_IPU_OVERLAP. One proportion test by speaker was run. (AG = 83/159; 58/138; 31/63; YM = 98/180; 33/125; 22/55).

	DVD	CID	MTX
AG	0.52	0.42	0.49
YM	0.54	0.26	0.40

Figure 8: Table of proportion test score

AG (X-squared = 3.1264, df = 2, p-value = 0.2095); YM (X-squared = 23.9082, df = 2, p-value = 6.433e-06*). IPUs overlapped at the beginning concern nearly 50% of the total number of overlapped IPUs for AG but not for YM. As expected, the proportion of IPUs overlapped at the beginning is not significant in the three corpora for the speaker AG, but this is not the case for the speaker YM. We can only partially confirm our second hypothesis.

3.3. Qualitative observations

While exploring our data, we noted a link between eyebrow movements and some linguistic phenomena in speech. In a first step, we systematically analysed data in order to generalise some effects of eyebrow movements on speaking turns. In a second step, we thought it would be interesting to analyse more precisely what was going on concerning other discursive effects.

3.3.1. Example of feedback

Once one establishes the link between IPU and eyebrow movement, one can analyse more precisely what happens in terms of discursive role.

Eyebrow movements can be associated with a mark of feedback produced by the listener. In this figure (Figure 9) we can see that speaker YM, in the follower role, has produced an eyebrow movement in the beginning of his turn. More precisely it is a frowning movement associated with a confirmation request. As we saw in Figure 5, speaker YM regularly produced the two types of eyebrow movements (raised and frowning movements), which explains this association.

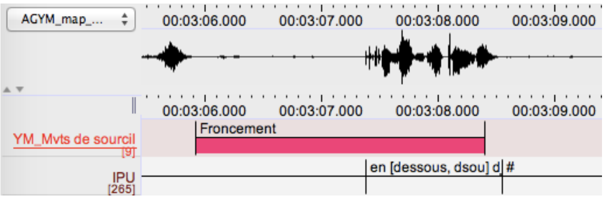


Figure 9: Example of feedback of speaker YM in MTX corpus

3.3.2. Disfluencies

Previous studies on the CID corpus were done on speech disfluencies. We used these annotations to explain some of the results of our study. When we selected criteria in order to detect eyebrow movements produced at the beginning of IPUs in

SPPAS, we chose overlaps as criteria (Figure 3). This process of automatic annotation can skew our results when disfluencies appear.

Disfluencies appearing at the beginning of a speaking turn are comparable to tests. These attempts are used to recover and help maintain the speaking turn. We noted that an eyebrow movement is produced after the altered portion of speech (Tier YM_DRMFI, annotated "D"). In this way we can suggest that eyebrow movement produced by the speaker starts when disfluencies end. The eyebrow movement could be a mark of structuring the discourse.

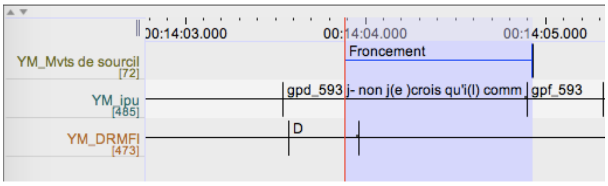


Figure 10: Example of disfluencies

4. Discussion

The goal of this study was to examine eyebrow movements in talk-in-interaction from a double point of view (based on interactional linguistics and gesture studies). Our first goal was to estimate the amount of eyebrow movements in three different French corpora and our second was to confirm eyebrow movement as a cue of turn taking.

Our findings reveal that the production of eyebrow movements (in terms of number of occurrences) significantly differs between the three different interactional corpora. This suggests that the interactional task impacts its production. DVD exhibits a large amount of eyebrow movements, followed by CID and then MTX. Given the task, i.e an exchange of personal opinion, participants in DVD are in a symmetric role and may hold the floor in a more regular exchange of turns while CID (narrative) is mainly composed of storytelling known as an asymmetrical activity, in which the main speaker needs several turn-construction units (or turns) to reach the end of his/her story ([18], [19], among others). Concerning the MTX, given the task, the role of director and follower are pre-established at the beginning of the interaction. As for DVD and CID, turn-taking organisation is indeed determined by these roles with their key task being the realisation of the concrete goal (reconstruct a way on a map) in a collaborative way. This task is the most asymmetric, the director always being the main speaker and the follower being the recipient. We explain the hierarchy between the three corpora as follows: the great amount of eyebrow movements in DVD could be associated with the frequent turn-taking from each participant while the movements in CID could be associated with the beginning of each storytelling (less frequent). This confirms the role of eyebrow movement as a structuring cue as shown by [8] and [9]. This also shows the non one-to-one relationship between eyebrow movements and the level of organisation since the eyebrow movements can indicate a level of turn or a level of activity (narrative). Furthermore, our results can corroborate [20] findings about prosodic cues. As high pitch onset, eyebrow movements can be seen as a relevant cue for indicating the beginning of 'big packages' that refers here to storytelling activity. This confirms the tiny link mentioned earlier between eyebrows movements and prosody. Concerning the MTX, we already mentioned the predefined role as director and follower that could have an impact on the occurrences of eyebrow movements. In fact, we suggest that eyebrow movements could be mainly produced by the follower when there is a problem in the explanation. The two participants have to find the right way on the map, thus they do not look at each other. When they are looking at each other, it is mainly the listener who asks confirmation or expresses a doubt or surprise [4]. Given that, the number of occurrences of eyebrow movements in the MTX is lower than in the DVD corpus or CID corpus.

Our second research question was about the location of eyebrow movements at the beginning of a speaking turn, whatever the condition. Despite the different interactional task, eyebrow movements seem to appear mostly at the beginning of an IPU. This effect confirms the role of eyebrows as a relevant cue in turn-taking organisation. The proportion of IPUs overlapping at the beginning is almost always the same, except for one case: CID_YM. One of the reasons of this failure with CID_YM may concern the unit of analysis. We chose to consider a speaking turn as an IPU, so each IPU has been analysed. We think that the designation of an IPU as a speaking turn is pertinent with a few adjustments. An IPU can be considered as a speaking turn if we take into account only IPUs that are alternated with the other interlocutor's speech (like in a speaking turn). If two IPUs of AG are following each other, the second IPU cannot be considered as a speaking turn, because the listener (YM) has not interrupted AG's speech. On a subsequent analysis we will take this criterion into account.

On the other hand, we noted that disfluencies could play an important role in automatic detection. They could blur the location of eyebrow movements at the beginning of a speaking turn. Eyebrow movement is synchronised with the real start of a speaking turn and not only with the simple fact of taking a speaking turn. In this way, we confirm the role of eyebrow movements as a cue of turn taking.

In further studies, we can improve our results by taking into account not only this type of phenomena (i.e. disfluencies) but also discursive roles for analysing feedback phenomena for example. We know that discursive roles and the type of production have an impact on speaking turns.

5. Conclusions

The question that we raised about the production of eyebrow movements according to the type of interaction has a response. In this study, with these corpora, the number of occurrences of eyebrow movements seems to be conditioned by the interactional task. According to our findings, the more a corpus allows for the expression of personal opinion, the more the participants will produce eyebrow movements.

Concerning the hypothesis about eyebrow movements occurring at the beginning of IPUs, we cannot confirm it for our two speakers. However, we can say that eyebrow movements tend to appear at the beginning of a speaking turn.

6. Acknowledgements

We would like to thank Laurent Prévot and Jan Gorish for lending me these corpora and for their assistance concerning the technical aspect of their utilisation. Thanks as well to Brigitte Bigi for training in using her software (SPPAS). Finally, we would like to thank Robert Espesser for his statistical explanations and comments and Benjamin Holt for proofreading this text.

References

[1] C. Darwin, The expression of the emotions in man and animals. *University of Chicago Press* vol. *526*, 1965.

[2] P. Ekman & W. V. Friesen, The repertoire of nonverbal behaviour: Categories, origins, usage, and coding. *Semiotica*, 1(1), pp. *49-98*, 1969.

[3] R. E. Kraut & R. E. Johnson, Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology* vol.*42*: pp. *853-863*, 1979.

[4] J. Bavelas *et al*, "Hand and facial gestures in conversational interaction". *The Oxford Handbook of Language and Social Psychology*, pp.*111*, 2014.

[5] E. Couper-Kuhlen, Towards and interactional perspective on prosody and a prosodic perspective. *Prosody in conversation: Interactional studies*, vol. *12*: pp.*1,* 1996.

[6] C. Cavé *et al* "About the relationship between eyebrow movements and Fo variations". In *Spoken Language. ICSLP 96. Proceedings. Fourth International Conference on* (Vol. 4, pp. 2175-2178), (1996).

[7] N. Chovil, "Social determinants of facial displays". *Journal of Nonverbal Behavior*, vol.*15*(3),pp.*141-154*, 1991.

[8] M. L. Flecha-García, "Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English". *Speech Communication*, vol.*52*(6), pp.*542-554*, 2010.

[9] C. Cavé *et al,* "Eyebrow movements and voice variations in dialogue situations: an experimental investigation". In *Seventh International Conference on Spoken Language Processing*. (2002).

[10] R. Bertrand *et al*, "Le *CID - Corpus of Interactional Data* - Annotation et Exploitation Multimodale de Parole Conversationnelle" *Traitement Automatique des Langues*, vol.*49-3*, pp.*105-134*, 2008.

[11] L. Prévot & R. Bertrand, "Cofee-toward a multidimensional analysis of conversational feedback, the case of french language". In *Proceedings of the Workshop on Feedback Behavior,* 2012.

[12] J. Gorisch *et al*, "Aix Map Task corpus: The French multimodal corpus of task-oriented dialogue". In *LREC,* 2014.

[13] A. Anderson, *et al,* "The HCRC map task corpus". *Language and speech*, vol.*34*(4), pp.*351-366*, 1991.

[14] H. Koiso et al, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs". *Language and speech*, vol.*41*(3-4), pp.*295-321*, 1998.

[15] H. Sloetjes & P. Wittenburg, "Annotation by Category: ELAN and ISO DCR". In *LREC*. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, the Netherlands, 2008. http://tla.mpi.nl/tools/elan/

[16] B. Bigi, "SPPAS : a tool for the phonetic segmentations of Speech", Language Resources and Evaluation Conference, ISBN 978-29517408-7-7, pages 1748-1755, Istanbul (Turkey), 2012.

[17] R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2014. http://www.R-project.org/

[18] M. Selting, the construction of units in conversational talk. *Language in Society*, vol.*29* (04), pp *477-517*, 2000.

[19] M. Guardiola & R. Bertrand, Interactional convergence in conversational storytelling: when reported speech is a cue of alignment and/or affiliation. *Frontiers in psychology*, vol. *4*, 2013.

[20] E. Couper-Kuhlen, Interactional prosody: High onsets in reason-for-the-call turns. *Language in Society*, vol. *30*, pp *29-53,* 2001.

Full-Body Gesture Recognition for Embodied Conversational Agents: The UTEP AGENT Gesture Tool

Ivan Gris, Adriana Camacho, David Novick

Department of Computer Science, The University of Texas at El Paso, El Paso, TX USA
{igris, accamacho2}@miners.utep.edu, novick@utep.edu

Abstract

Recognition of body gestures has long challenged developers of interfaces for real-time interaction between humans and embodied conversational agents (ECAs). In this paper we present a computationally simple approach to full-body gesture recognition along with an example of a human-agent application that makes use of it. We discuss how developers can use the tool to create pose libraries and how it works across different applications. And we evaluate gesture recognition implemented with the tool in the context of the human-agent application.

Index Terms: full-body gesture recognition, embodied conversational agents

1. Introduction

Full-body gesture recognition provides natural human-computer interaction in applications such as embodied conversational agents (ECAs). However, this approach to interaction remains difficult to achieve due to low recognition accuracy, distant sensor positioning, performance issues in real-time processing, intrusive interactive tracking technology, and the expense of capturing motion for representation of gestures.

For developers of ECAs, agent gestures can be animated or represented for purposes of recognition via hand-drawing or motion capture. For example, one agent with hand-drawn animation is a virtual nurse for hospital patients with low health literacy [1]. But hand-drawn animation is time-consuming and represents an artistic rather than naturalistic approach to gesture generation. And for recognition of gestures by human conversants, hand-drawn animation is highly problematic, in large part because each animation represents a particular movement path rather than a robust representation that accounts for variability in human motion.

Other ECAs use gestures generated via motion capture. [e.g., 2, 3]. This approach provides gestures that are more plausibly realistic, although it is certainly possible to capture and produce gestures that are idiosyncratic and unconvincing. Yet the motion-capture approach also can be time-consuming. For example, developing the relatively simple gestures for the ECA in the “Escape from the Castle of the Vampire King” game [4] took many weeks. And capturing gestures for purposes of recognition, which involves recording and processing the multiple examples of gestures needed for robust recognition, can require great effort.

To speed radically the process of capturing human gestures for purposes of generating ECAs’ gestures and of recognizing the gestures of the ECAs’ human conversational partners, we developed a tool that is capable of recognizing full-body gestures in real time and that can generate pose libraries for recognition across applications. In this paper, we review methods of gesture recognition that target different parts of the body, discussing the advantages and disadvantages of these methods. We present our gesture tool, explain how it works, and briefly describe the mathematical principles of full-body gesture recognition on which the tool is based, discuss the tool’s potential applications. We discuss how we use the tool to aid with gesture annotation in real time and how the tool connects with our ECA system to enable real-time responses to gestures. We conclude with a discussion of the tool’s limitations and how future updates will address these.

2. Background

To increase the believability and naturalness of human-agent interactions, developers seek to build agents capable of representing and interpreting traits that humans seem to do effortlessly. This includes the recognition of speech and gesture.

There are many commercial and research solutions to gesture recognition. Some target the face and focus on detecting emotions through facial features [5] or skin color [6], and others focus on gaze patterns. These systems, though, target specific body parts and usually require people to sit in front of a camera or sensor and maintain a relatively static position. With sensors like the Kinect, a device that is able to track user’s body position and movements, users and developers alike have greater flexibility in terms of distance and gesture types. These sensors can be used at a short range to perform head [7], gaze [8] or hand tracking [9], while at greater distances they can cover the full body. This often involves a tradeoff, where detection at a short distance cannot be performed with a full-body setup, and vice-versa, leaving it to the developer’s priorities to choose between full-body tracking versus head, gaze or hand tracking.

Although applications are often controlled through a computer screen and a traditional keyboard and mouse setup, some ECAs, such as those developed in our lab [e.g., 4] are life-sized projections of virtual human characters whose interaction instead aims for a more naturalistic approach using speech commands. The goal of these agents is to perform conversational tasks, often involving user-agent collaboration. To maintain the naturalness of the conversation, agents often need to react to the user’s physical behavior, such as facial expression, gaze, and gesture, just like humans would. The more detailed the information about the user’s non-verbal actions are, the better the agent can interpret and more accurately react to them [10, 11]. This enables a better interactive storytelling application, a domain of choice for full-body gesture recognition, as users can interact with objects contained in the same virtual space as the agent [12, 13].

These systems provide real-time full-body tracking in 3D, often including information about the hands and the face concurrently. However these systems can be costly and intrusive, meaning they often require users to wear special suits or markers to be detected by a set of several cameras positioned across an empty room. This sort of elaborate setup and its associated costs are not the only barriers to interaction and gesture recognition. Even though they work with much greater accuracy, this information is usually processed and applied to 3D characters, meaning that the tracking information is translated directly into a virtual character to make the character replicate the actor's movements as closely as possible [14]. This means that there is no further analysis of the gesture-capture data, which makes impractical the identification of gestures and reactions in real-time to these gestures. And if a system does not identify full-body gestures automatically, this means that analysis of gestures will require, manual annotation of videos (e.g. [15]). Even though video annotation is nonintrusive and can be encoded on an abstract level, this is still a burdensome and time-consuming process.

To address these problems, we built a tool using Microsoft's Kinect sensor that suits specifically the full-body gesture recognition scenario while standing at a distance of six to eight feet away from the sensor. This tool is capable of generating poses for libraries that can be used for recognition through applications. Using this pose library, this tool identifies users' full body gestures in real time which enable the capability of analyzing the gestures performed by the user.

A similar system was recently developed that included similar functionality, although its current applications are game-oriented and is not actively maintained [16]. This system provided a full-body gesture recognition solution for existing applications, but this addressed only part of the challenge. When translating existing controls to gesture recognition, subjects are often required to perform the same gesture repetitively, and although the gestures can be metaphors of real-life gestures, they might not be ergonomic.

Accordingly, we designed our tool for detecting large sets of unique gestures and for users to create, export and import these gesture sets. Another key difference is that our tool can be used not only to interact with different applications but also to generate log files as spreadsheets that present the users' behavior across time, presumably facilitating researchers to analyze this data rather than the painful long process of annotating gestures manually.

Our approach sought to lower significantly the computational cost of gesture recognition. As discussed in Section 3, to make real-time recognition computationally feasible, our approach converts the 3D rendering to a 2D representation. An alternative approach involved using only depth information [17]. Again, to simplify recognition to reduce computation, our approach used a finite-state model for gesture recognition (see [18] for a review of alternative approaches generally, and see [19 and 20] for reviews of alternative approaches using the Kinect), although our approach is even simpler than the FSM model of [21] because it relies on pose sequences without timing information.

We connected our tool to a markup language and interpreter [22], a middleware system that enables external applications to access pose libraries and gesture detection. In addition, we created a user interface (see Figure 2) that enables developers to build pose libraries based on screenshots of the desired poses and that has additional features aimed at improving accuracy through basic statistical analyses. Figure 1 shows a human performing a “hi-five” gesture that is recognized and interpreted by an ECA.

3. Tool implementation

In this section we delve deeper into the implementation, features, and use cases of the UTEP AGENT gesture tool. The

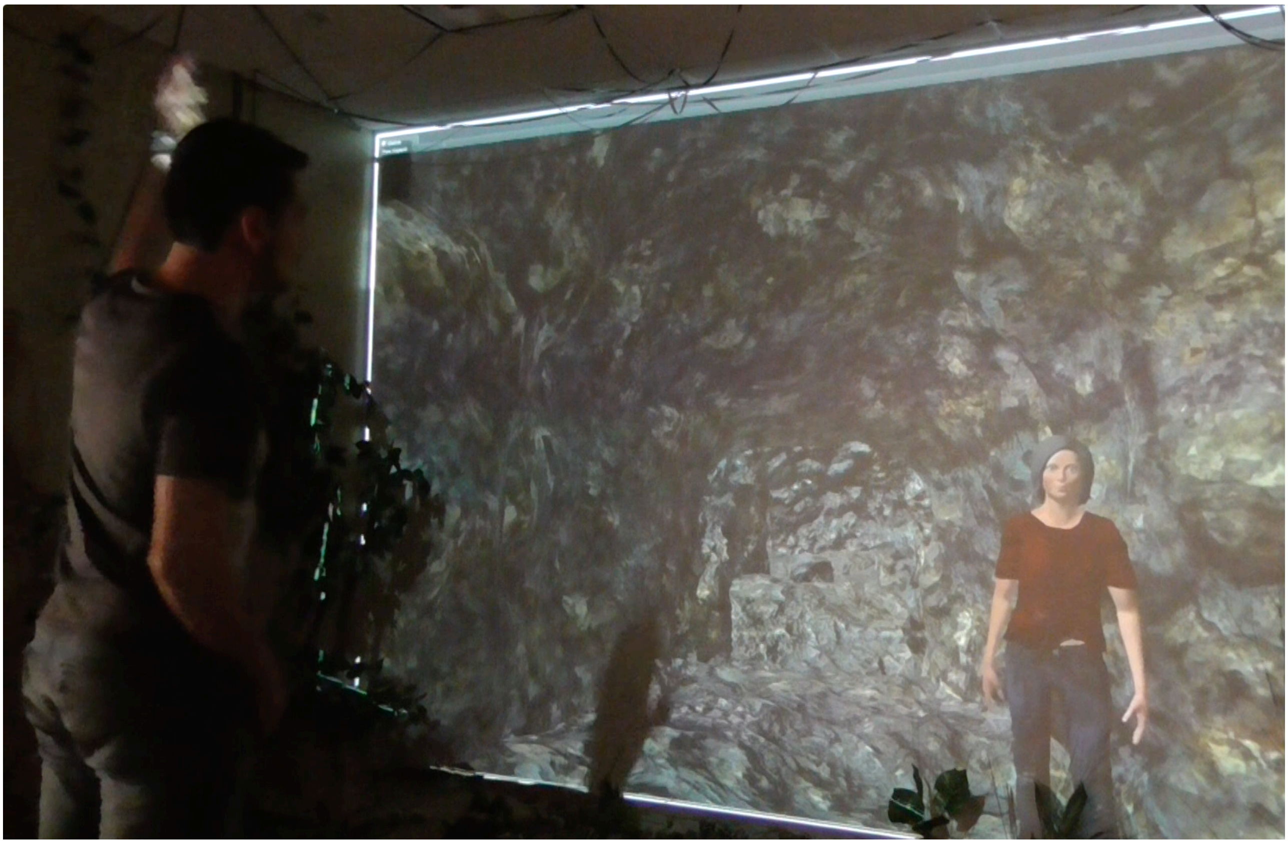
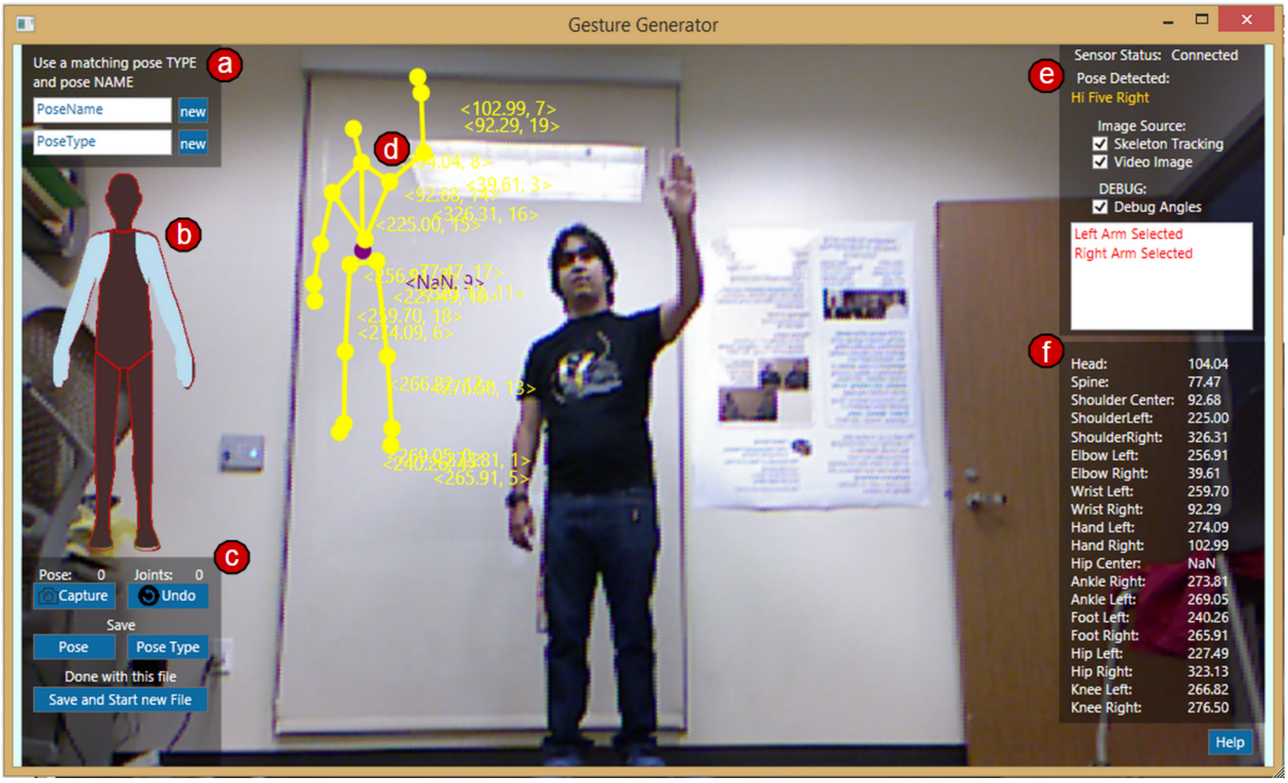


Figure 1. *Human, interacting with ECA, performing a "hi-five" gesture. The human's gesture is sensed by a Kinect just in front of the projection wall and is interpreted via the gesture tool.*



Use a matching pose TYPE and pose NAME

PoseName new

PoseType new

Pose: 0 Joints: 0

Capture Undo

Save

Pose Pose Type

Done with this file

Save and Start new File

Sensor Status: Connected

Pose Detected:
Hi Five Right

Image Source:
☑ Skeleton Tracking
☑ Video Image

DEBUG:
☑ Debug Angles

Left Arm Selected
Right Arm Selected

Head:	104.04
Spine:	77.47
Shoulder Center:	92.68
ShoulderLeft:	225.00
ShoulderRight:	326.31
Elbow Left:	256.91
Elbow Right:	39.61
Wrist Left:	259.70
Wrist Right:	92.29
Hand Left:	274.09
Hand Right:	102.99
Hip Center:	NaN
Ankle Right:	273.81
Ankle Left:	269.05
Foot Left:	240.26
Foot Right:	265.91
Hip Left:	227.49
Hip Right:	323.13
Knee Left:	266.82
Knee Right:	276.50

Help

Figure 2. *UTEP AGENT gesture tool interface tracking a "hi-five" pose. (a) specify the name (e.g., high five) and type (e.g., right/left) of a pose; (b) selection of specific body parts for capture; (c) capture controls; (d) 2D rendering stick figure of a person with every dot representing each joint; (e) debugging tools showing the recognized pose (if any), record of your activity and turning section "f" (joint angles) on/off; (f) list of all 20 major joints recognized by the Kinect with its angle value.*

tool is built as a standalone Windows application that can be connected to Unity3D, a game engine that renders and animates our ECAs and the virtual environment in which these appear.

Based on depth information captured by the infrared camera from the Kinect, the tool renders a skeleton consisting of lines connecting 20 major joints of the human body (see Figure 2f). These skeleton is a 2D rendering in stick-figure style of the person recognized, as shown in Figure 2d, which shows a person performing a hi-five gesture like that shown in Figure 1. Although the Kinect is able to recognize human figures and track their joints in real time, it cannot differentiate between poses. In fact, the sensor produces only a visual representation of lines connecting the joints and updates them according to their position in 3D Cartesian coordinates. Microsoft's Kinect SDK [23] enabled pose detection.

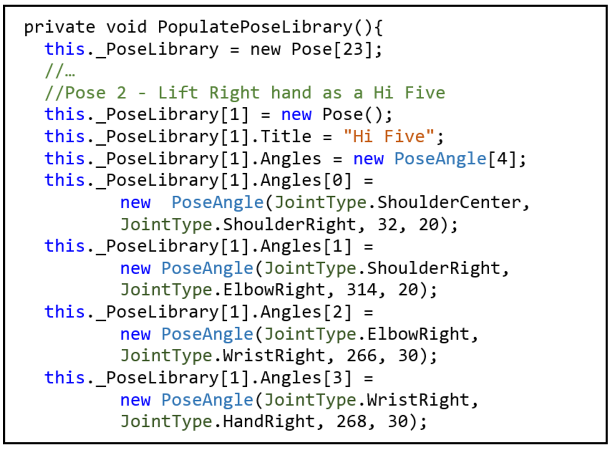
The configuration of the subject's body joints, their position and posture, defines a pose. However, creating a pose recognizer from coordinates presents several problems during translation, rotation, and scaling of the skeletons. First, coordinates of the tracked joints change depending on the base position. That is, doing a pose while standing to the left of the screen will render different coordinates than doing the same pose on the right edge of the sensor's tracking field. This could be solved by using an offset parameter that checks for the same pose across different locations, but this approach can become computationally expensive, depending on the size of the pose library. A solution to this would be to calculate the offset based on an anchor point, in this case the hip joint, that controls translation, but this approach would remain ineffective for rotation and scale.

Scale is an issue, not because people come in all shapes and sizes but because they move. People do not grow and shrink in a few seconds, but they do change their distance from the sensor, which looks like a growing and shrinking effect to the sensor. In other words, when people translate along the z-axis, they appear larger or smaller on screen. This, combined with the x-axis translation, can make the process computationally expensive and unmanageable in real time.

To resolve these issues, first we eliminated the depth information. Because the rendering occurs in 2D regardless of the 3D information contained in the coordinates, and because in our research settings users are always located directly in front of the sensor at a relatively constant distance of eight to nine feet, the 3D information does little to help the gesture tool accuracy but does slow our system considerably.

Second, once the coordinates are transformed to 2D, each joint is triangulated using the parent joint (in this case the hip center) as base and creating a right triangle. We then use this triangulation to switch from location information to angles between joints to avoid normalizing position information in real time and to improve the accuracy of our measures and enable a more intuitive margin of error. Because positions are relative to the standing position of the person interacting with the tool, different coordinates could mean the same gesture, making it hard to classify or differentiate gestures that occur at a different standing position. By using angles, we can instead guarantee that they will remain constant regardless of the user's starting position. However a limitation still remains in our tool because it does not completely remove the ambiguity of angles. To address this ambiguity, pose capture and recognition have to be done in the same room with the same angle position of the Kinect performing the recognition.

The third step is to recognize a pose. However to do this there must already be information about the pose to be recognized. To address this, we created a pose library that contains an array of pose objects; Figure 3 presents sample code. These objects contain a subset of joint pairs and the



```
private void PopulatePoseLibrary(){
  this._PoseLibrary = new Pose[23];
  //…
  //Pose 2 - Lift Right hand as a Hi Five
  this._PoseLibrary[1] = new Pose();
  this._PoseLibrary[1].Title = "Hi Five";
  this._PoseLibrary[1].Angles = new PoseAngle[4];
  this._PoseLibrary[1].Angles[0] =
         new  PoseAngle(JointType.ShoulderCenter,
         JointType.ShoulderRight, 32, 20);
  this._PoseLibrary[1].Angles[1] =
         new PoseAngle(JointType.ShoulderRight,
         JointType.ElbowRight, 314, 20);
  this._PoseLibrary[1].Angles[2] =
         new PoseAngle(JointType.ElbowRight,
         JointType.WristRight, 266, 30);
  this._PoseLibrary[1].Angles[3] =
         new PoseAngle(JointType.WristRight,
         JointType.HandRight, 268, 30);
```

Figure 3. *Pose library sample code*

angles between them. For example, to recognize a “hi-five” pose we would be interested in the angles that form between the shoulder, the elbow, the wrist, and the hand joints. This is still not enough, however, because we need to mirror the values to enable gestures to be executed symmetrically for both left and right side of the subject’s body.

Initially, this process required manually taking screenshots from the rendering of the angles of a person on screen performing the desired pose for integration into the library. These angles were then passed to Excel sheets and processed manually to calculate the average and a proper margin of error to populate the pose library. To automate this process, we created the gesture capture tool as a separate module. The capture tool enables the developer to select the relevant body parts for the pose capture, as shown in Figure 2b. Then it enables the capture of those joint angles, using the controls in Figure 2c. The process requires the developer to click a capture button while the pose actor is representing the pose in front of the sensor. In addition, the tool enables developers to capture several times the same pose from the same or different pose actors to improve accuracy. As different people do the same gestures differently, or the same person might slightly change posture between one attempt and another, the capture tool collects the data, analyzes it by calculating the maximum, minimum, and average angles, and estimates a range parting from the mean of the angles required to recognize these gestures in the majority of cases. This is effectively calculating a margin of error, depending on the variety of poses that were captured.

There is a tradeoff between multiple captures and few or single captures. The more captures of the same pose (or gesture) that are taken, the more accurate the recognizer becomes. But the increased precision may prevent users from being recognized properly due to the reduced margin of error. In contrast, smaller sets of training data might lead to over-coverage and large margins of error due to frequent outliers. To examine the results and identify these outliers we generate two files. One file contains the values of all captures for each joint angle, making it easier for us to find these outliers and, if necessary, to recalculate the margin of error. The second file is in xml format and contains tags for every joint identifier, its average angle, and its margin of error calculated by getting the smaller value of either the difference of the maximum and the average, or the minimum and the average.

When the gesture capture tool has defined the angles and their respective margin of error, the pose is then added to the pose recognition library and can now be named and detected. The resulting string of the detected pose can then be used to trigger events in other applications or simply collect the data of common gestures (e.g., hands on hips, arms crossed, hand on face). Once the poses are stored in the library, we can build gestures from them. Because gestures require movement, we define a gesture as a sequence of poses. Once a pose of the collection of poses that constitute a gesture is detected, the system then expects to detect a second pose (or some number of poses) that will integrate a gesture and only then be detected as such. In other words, when the user follows the pose sequence, the tool detects the gesture.

4. Evaluation

Currently, we are using the UTEP AGENT pose tool for several studies, including analysis of the amplitude or extraversion of gestures and poses. We also use the tool as part of an immersive jungle-survival application in which we evaluate the level of rapport between humans and virtual agents as a function of their non-verbal behaviors. These behaviors are recognized to enable physical interaction with the ECA and its virtual world.

For the jungle game application we defined two types of gestures in the pose library to be detected: task gestures and background gestures. Task gestures were performed where users had to accomplish a certain task (e.g., lift hand, strike, throw spear) to advance through the story. The background gestures were performed by the user but were not necessary to advance through the story (e.g., crossed arms, normal stance, hands in front, hand on shoulder, hand on face). At the same time, we automatically capture and annotate, in a log file, the background gestures so that we can avoid manually annotating hours of paralinguistic behaviors.

The annotation includes gestures from both the human and the ECA, because we know when the agent will change poses from the animations that are specified in scripted interaction. For the human, the gesture tool detects when the subject does a certain gesture and adds a corresponding time-stamped annotation. This results in a graph like the one in Figure 4, which shows the changes in gestures of both the agent and the user across time.

Each interaction session of the jungle game where these gestures were recorded lasted for about 40-60 minutes; we expected the user to perform eight task gestures to advance in the scripted story. Users were not instructed as to how to perform the gestures, which resulted in a longer period of people trying to figure out how to perform the gesture resulting in some variance of gesture performance. For

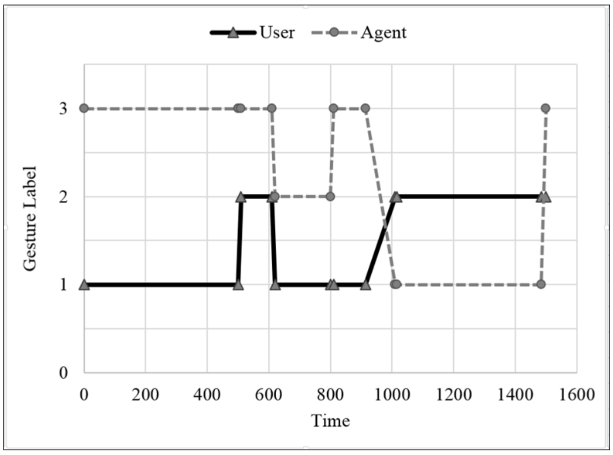


Figure 4. *User-agent gesture timeline. The numbers on the y-axis are labels of different gestures.*

example, they were asked to "strike a magnesium bar to light a fire." This gesture might not be as intuitive and does not have a standard way to be performed, which resulted low recognition rates for this gesture.

To evaluate the success/failure rate for recognition of the task gestures, we annotated the performance of these task gestures in the interaction of 30 users in the jungle game. The accuracy percentage was calculated with those task gestures that were performed correctly and recognized without a problem over the false negatives (when users performed the correct gesture and the gesture tool had a hard time recognizing them) and the false positives (when users did not perform the correct gesture and the gesture tool recognized them anyway) plus those gestures that were correctly recognized. For most task gestures, recognition accuracy ranged from about 50 to 80 percent; recognition was much lower for the unintuitive "strike 2" task. Table 1 reports these results. The recognition rates reflect multiple tries by the users; usually the users were able to achieve gesture recognition eventually.

Table 1. *Accuracy percentage of recognition of task gestures*

Gesture	**Accuracy percentage**
lift hand	70.73%
strike 1	51.02%
throw spear 1	49.02%
throw spear 2	65.79%
throw spear 3	70.27%
ventilate	77.78%
lift hand	65.85%
strike 2	25.00%

5. Discussion

Before the tool, generating a pose library took days or weeks of manually screening participants, getting their joint angle information, filtering the joints to remove the non-necessary joints for each poses, and collecting and aggregating the data from the different participants. With the tool, the process has been largely automated, and we now only need to have participants line up, stand in front of the sensor, and get scanned once per person per pose. For example, we used the tool to generate in less than an hour a library containing over 20 poses by scanning 12 members of our research group several times, with each person enacting a pose at a time. Participants did not receive any additional instructions apart from where to stand and what pose they had to enact. Each pose took about 15 seconds, and the 240 total poses from the 12 subjects took an hour to collect.

The tool, however, has several major limitations. Principally, aspects of the tool designed to reduce computational cost correspondingly eliminated consideration of information about depth and speed, our pose definitions average across all joints rather than focusing on the most meaningful joints, and our recognizer relies on context constraints to reduce confusion among gestures.

With respect to speed, the tool's gesture recognition in its current implementation is based on pose sequences that are insensitive to time. This means that a gesture will be recognized when the human follows a pose sequence regardless of the speed with which it is executed. This is not optimal, as gestures can vary in meaning depending on speed of execution [24]. We plan to integrate timers that can be set between poses to add greater precision to the gesture recognition.

We note, too, that the sequence of poses to detect a gesture can vary depending on users' performance and the accuracy of the Kinect in detecting a pose, making it difficult for a gesture to be recognized even if the user has performed the correct gesture.

In terms of joint angle accuracy, currently, our tool simply averages over different pre-labeled gesture instances and gives the developer the liberty to decide which body parts are relevant for a specific pose, treating all joint angles of selected body parts alike. Additional features, such as machine learning algorithms, could have been integrated for further refinement of the pose generation. A clustering approach, for example, could increase accuracy of pose generation by focusing only on relevant joint angles. In this case, a cluster of joint angles would represent a predefined pose in the tool. However, this approach would be limited by its inability to remove overlap within poses, an issue that is handled appropriately in our current implementation.

Another concern involves confusion among gestures. When gestures are not well defined, their margin of error might be higher than usual. If this happens across several gestures, there might be subsets of coordinates that fall between one or more gestures, making the recognizer unable to decide which gesture was actually executed. To avoid this, we activate and deactivate poses or gestures based on our expectations, much in the same way that we create contexts for speech recognition. By lowering the number of poses that can be recognized at the same time we decrease the overlap risk. This approach can be problem, however, when the user does not know what poses to expect or when two poses that overlap are expected. Moreover, our technique of reducing the joint positions to a 2D plane significantly increases the risk of confusion.

The tool has other limitations related to its implementation. Indeed, one of the tool's main advantages is also a disadvantage: it can perform all the data gathering and analysis in real time, but only in real time. This means that the tool cannot analyze a video recording after it has been captured. In contrast, motion-capture systems can store the 3D data and can be used at a later time for tweaking and post-processing to adjust for different physical traits among actors and the characters they represent. For some studies, we have been able to capture and store 3D information as rendered by the depth sensors of two Kinects. However, the data sets become large, making it infeasible to record several hours of conversation for further analysis or automated annotation. Moreover, the analysis cannot be executed in real time, and as it provides 3D depth data rather than a 2D skeletal representation, our tool cannot convert or interpret the data in these formats.

Although the tool is limited in terms of dimensional space and post-processing data handling, it has proven to be useful and reliable for our current applications. Provided that there is post-processing of the pose library to minimize overlap, the tool performs well even though it is a lightweight application in comparison to commercial motion-capture systems or other recognizers that are unable to process information in real time. As it is, with our ECA front-end applications, the tool can be applied to real-time interaction, real-time video annotation, and pose analysis.

In the future, we plan to implement the recognizer with 3D coordinates on a more powerful computer, to include timing information for gestures, and to update the recognizer and capture tool to work with the Kinect ONE, which offers greater accuracy and additional capabilities.

The UTEP AGENT gesture tool is available from the authors.

6. Acknowledgments

For their contributions to the studies through which we developed the gesture tool, we thank Diego A. Rivera, Mario Guitierrez, Alex Rayon, Laura Rodriguez, Paola Gallardo, Alfonso Peralta, Victoria Bravo, and Brynne Blaugrund.

7. References

[1] T. Bickmore et al., “Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents,” in *Proc. Conf. on Human Factors in Computing Systems*, Boston, MA, 2009, 1265-1274.

[2] M. Thiebaux et al., “Smartbody: Behavior realization for embodied conversational agents,” in *Proc. 7th Intl. Joint Conf. Autonomous Agents and Multiagent Systems*, vol. 1, Richland, SC, 2008, 151-158.

[3] C. Bregler et al., “Turning to the masters: Motion capturing cartoons,” *ACM Transactions on Graphics* 21, no. 3 (2002): 399-407.

[4] D. Novick and I. Gris, “Building rapport between human and ECA: A pilot study.” In *Proc. HCI Intl.*, Crete, Greece, 2014, 472-480.

[5] C. Busso et al., “Analysis of emotion recognition using facial expressions, speech and multimodal information,” in *Proc. 6th Int. Conf. Multimodal Interfaces*, State College, PA, 2004, 205-211.

[6] G.A. Ramirez et al., “Color Analysis of Facial Skin: Detection of Emotional State,” in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conf. Computer Vision and Pattern Recognition*, Columbus, OH, 2014, 474-479.

[7] L.P. Morency et al., “Head gestures for perceptual interfaces: The role of context in improving recognition," *Artificial Intelligence* 171, no. 8 (2007): 568-585.

[8] D. Bohus and E. Horvitz. “Facilitating multiparty dialog with gaze, gesture, and speech,” in *Int. Conf. Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, 2010, doi: 10.1145/1891903.1891910.

[9] P. Trigueiros et al., “Hand Gesture Recognition System Based in Computer Vision and Machine Learning,” in *Developments in Medical Image Processing and Computational Vision*, Springer International Publishing, 2015, 355-377.

[10] J. Gratch et al., “Virtual rapport,” in *Intelligent virtual agents*, pp. 14-27. Springer Berlin Heidelberg, 2006.

[11] J. Bailenson and N. Yee, “Digital chameleons automatic assimilation of nonverbal gestures in immersive virtual environments.” *Psychological Science* 16, no. 10 (2005): 814-819.

[12] D. Thue et al., “Interactive storytelling: A player modelling approach,” in *Proc. Artificial Intelligence for Interactive Digital Entertainment Conf.*, 2007, 43-48.

[13] U. Spierling et al., “Setting the scene: playing digital director in interactive storytelling and creation.” *Computers & Graphics* 26, no. 1 (2002): 31-44.

[14] E. Bevacqua, I. Stankovic, A. Maatallaoui, A. Nedelec and P. De Loor, 'Effects of Coupling in Human-virtual Agent Body Interaction', in *Intelligent Virtual Agents*, Boston, MA, 2014, pp. 54-63.

[15] M. Kipp, M. Neff and I. Albrecht, 'An annotation scheme for conversational gestures: how to economically capture timing and form', *Lang Resources & Evaluation*, vol. 41, no. 3-4, pp. 325-339, 2007.

[16] E.A. Suma et al., “Adapting user interfaces for gestural interaction with the flexible action and articulated skeleton toolkit.” *Computers & Graphics* 37, no. 3 (2013): 193-201.

[17] K.K. Biswas and S.K. Basu, “Gesture recognition using Microsoft Kinect®” in *5th Intl Conf. on Automation, Robotics and Applications (ICARA*), 2011, 100-103.

[18] S. Mitra and A. Tinku, “Gesture recognition: A survey,” in *IEEE Tran. on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* 37.3 (2007), 311-324.

[19] O. Patsadu et al., “Human gesture recognition using Kinect camera," in *2012 Intl Joint Conf. on Computer Science and Software Engineering (JCSSE),* 2012, 28-32.

[20] J. Han et al., “Enhanced computer vision with Microsoft Kinect sensor: A review," in *IEEE Trans. on Cybernetics,* 43.5 (2013): 1318-1334.

[21] P. Hong et al., “Gesture modeling and recognition using finite state machines,” in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recogn.*, Grenoble, France, Mar. 2000, 410–415.

[22] D. Novick et al., “A mark-up language and interpreter for interactive scenes for embodied conversational agents,” in *Proc. HCI Intl. 2015*, Los Angeles, CA, in press.

[23] J. Webb and J. Ashley, *Beginning Kinect programming with the Microsoft Kinect SDK*. [New York]: Apress, 2012.

[24] M. Neff, N. Toothman, R. Bowmani, J. Fox Tree and M. Walker, 'Don't Scratch! Self-adaptors Reflect Emotional Stability', in *Intelligent Virtual Agents*, Reykjavik, Iceland, 2015, pp. 398-411.

A modality axis in gesture space?

The Vertical Palm and construal of negation as distance

Simon Harrison [1]

[1] University of Nottingham Ningbo China

simon.harrison@nottingham.edu.cn

Abstract

Locations and movements in gesture space reflect both interactional and conceptual constraints. This paper presents evidence that suggests gesture space is at least partially structured by a conceptual axis projecting outwards from the speaker's body, along which gestures are positioned or moved to reflect the construal of negation as distance – hence, a modality axis (as in 'epistemic mood'; [1, 2]). The Vertical Palm gesture described by Kendon (2004) offers an interesting case in point because it illustrates how the interactional and modality axes inter-relate. Integrating form-based gesture study with discourse analysis and conceptual semantics [3, 4], this paper analyses Vertical Palm gestures in a corpus of spoken English to shed light on the intersection between negation, gesture, and cognition.

Index Terms: gesture space, negation, modality axis, recurrent gesture

1. Introduction

Where a speaker locates a gesture depends on numerous factors. One factor is the position of addressees in the interactional context, and another is the conceptual content motivating the gesture. By demonstrating how a particular gesture operates simultaneously in relation to both an interactional and a conceptual axis, this paper sheds light on the intersection between negation, gesture, and conceptualisation.

The gesture in focus was termed the 'Vertical Palm' gesture by Kendon (2004) in his context-of-use study of the Open Hand Prone gesture family. The Vertical Palm gesture occurs when the open hand or hands are raised vertically into space with the flat palm(s) oriented away from the speaker's body (Figure 1). The Vertical Palm is a semi-conventionalised gesture form that expresses the general "semantic theme of stopping or interrupting a line of action that is in progress" (pp.248-249).

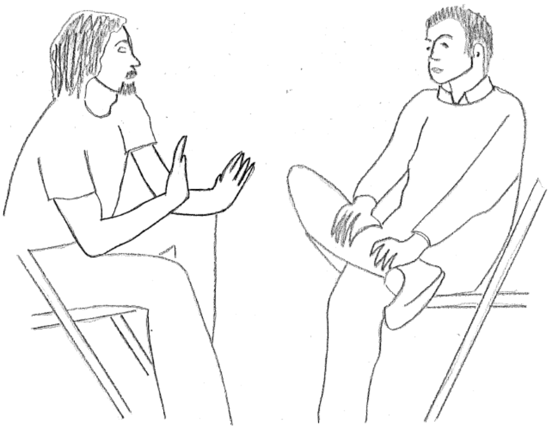


Figure 1: The Vertical Palm gesture

The Vertical Palm may occur in linguistically diverse contexts, but much research now documents its connection with the overt expression of negation. Calbris (2011) pioneered this connection by establishing analogical links between negative concepts (e.g. removal, exclusion, separation, opposition), structural, cognitive and etymological analyses of verbal forms of negation, and gesture symbolism (see 'les variantes gestuelles de la négation'—gestural variants of negation—which included the Vertical Palm). Harrison (2010, 2015) explored how such links are manifest in the moment-to-moment temporal unfolding of negative utterances and established that the form and organisation of gestures like the Vertical Palm were integrated with linguistic structures of negation, including negative node, scope, and focus.

The Vertical Palm shows the hallmark characteristics of a 'recurrent gesture' – it is a relatively "stable form-meaning unit [that] recurs in different contexts of use over different speakers in a particular speech community" ([5] pp.1559-1560). Recurrent gestures were absent from the original 'Kendon's continuum' classification scheme [6, 7], but have since been inserted between gesticulation and emblems [8]. Recurrent gestures characteristically exhibit a number of form-meaning variants; they operate on speech and achieve meta-communicative, interactive, and pragmatic functions; and they exhibit schematised re-enactments of everyday manual actions [5, 9-12].

Through the semiotic mode of enacting ('the hand acts' – [13]), when speakers perform a basic Vertical Palm gesture they reenact 'stopping' (hence the semantic theme of 'stopping or interrupting'; [10]). The hand is shaped so that the surface of the palm is displayed as if it is to come into contact with someone or something. Conceptualised counterforce evokes the stopping action, which may be applied in relation to the speaker or the addressee – this determines where the gesture will be located [10]. Supporting previous research into personal and interpersonal gesture spaces (e.g. [14]), the location of the Vertical Palm gesture thus highlights an interactional axis structuring the gesture space (Figure 2).

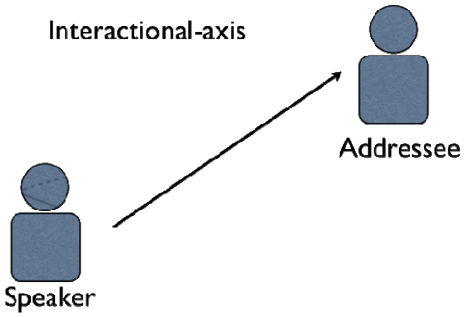


Figure 2: An interactional axis structuring gesture space

In addition to the interactional axis, where a speaker locates a gesture may be motivated by conceptual content i.e. conceptualisation [15]. For example, it is well known that locations and axes in gesture space may be used metaphorically to iconically encode the source domains of conceptual metaphors [16] (cf. contributions to [17]). Regardless of location, the form of recurrent gestures often metaphorically map an instrumental action onto a communicative action [5, 9, 11, 12]. Gestures associated with the expression of negation have been connected metaphorically to actions that serve to clear the space around the body, such as 'holding away' for the Vertical Palm gesture [18]. Additional derivations proposed for the Vertical palm include "creating a barrier..., or pushing, or sweeping something away" ([10] p.283).

As a gesture associated with negation, contextual, linguistic, and conceptual observations of the Vertical Palm are consistent with linguistic models of epistemic mood and negation theorised in Cognitive Linguistics. Thus, Langacker's (1991) basic epistemic model situates *irrealis* as a 'region' of conceptual space 'outside' the cognizer's immediate reality or 'epistemic center' [19]. In Chilton's (2014) geometric model of conceptual space, cognizer's judgements of discourse referents are plotted along a modality axis ranging from 'real/true/right' at the deictic center to 'unreal/untrue/wrong' at the axis endpoint [1, 2]. Furthermore, Johnson's (1987) image schema theory presents negative speech acts and epistemic mood dynamically in terms of a subset of underlying force schemas, including BLOCKAGE, COMPULSION, and COUNTERFORCE ([20] pp.41-64).

Taken together, gesture space usage of the Vertical Palm gesture (and presumably other gestures associated with negation) appears to be motivated by conceptualisations of negation in which locations further away from the body map onto conceptualisation of negation as distance. In this brief paper my argument is that the modality axis that Chilton described in abstract conceptual space [1, 2] may be mapped metaphorically onto the physical gesture space as a way of modelling certain uses of the Vertical Palm gesture (Figure 3). The aim is to establish the modality axis as a sub type of conceptual axes structuring the gesture space.

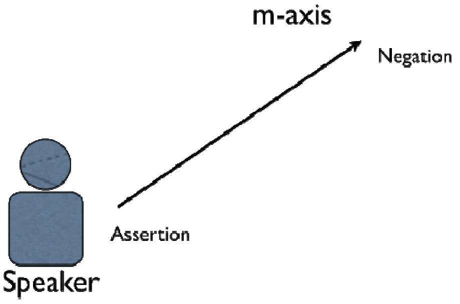


Figure 3: A modality axis structuring gesture space

Research questions include, first of all, can positing such a modality axis help interpret the Vertical Palm gesture data? And secondly, comparing Figure 2 and Figure 3, how do speakers locate gestures in relation to both interactional and modality axes? A series of Vertical palm examples will now be analysed to argue for the presence of a modality axis and illustrate how the two axes interrelate.

2. Methods

13 examples of the Vertical Palm with negative utterances have been identified from conversations among pairs of English speakers, elicited using conversational topics on a board game as a stimulus [21]. Methods of analysis integrate form-based gesture study with discourse analysis and conceptual semantics, as developed within the *Towards a Grammar of Gesture* research group [3, 4].

My application of the ToGoG methodology for this study meant first describing specific form variations of the Vertical Palm, then based on form properties, seeking to identify modes of representation, underlying actions, and salient conceptual motivations; the spoken discourse of which the gestures were part was then examined to determine timing of gestures in relation to linguistic structures (focusing on negation in this case) and discursive processes (pragmatic implications, speech acts, interactional moves). Tools from conceptual semantics were finally (re-)applied to interpret the discourse multimodally in terms of speaker's dynamic conceptualisation (including image and force schema analysis, as well as conceptual metaphor analysis), thereby verifying whether initial interpretations based on observable gesture features alone were borne out by the full contextual data.

As befitting of the current data set and research questions, this process was predominantly qualitative involving a continual back and forth between preliminary coding categories, analysis, and interpretation—however the method is applicable within both qualitative and quantitative research paradigms [3, 4]. The data was viewed and coded in ELAN and an Excel file was used to organise the data.

3. Examples and Analysis

The analysis is divided into the three main gesture variants observed, which are distinguished primarily by the form parameter 'movement':

- Zero-movement (VP+Ø) – there is no movement, the Vertical Palm is placed and held in the gesture space; the dominant underlying action is 'stopping' and the salient image schema is BLOCKAGE (6 examples).
- Lateral movement (VP+Lateral) – performance of the Vertical Palm stroke includes a single abrupt movement along the lateral axis; the dominant underlying action is 'throwing aside' and the silent image schemas are BLOCKAGE and COMPULSION (3 examples). When the movement is slow and steady rather than abrupt, the mode of representation may be 'molding' a barrier rather than enacting a stoppage.
- Oscillation (VP+Oscillate) – performance of the Vertical Palm stroke includes repeated side-to-side movements along the lateral axis; the dominant underlying action is 'erasing/rubbing out' and the silent image schemas are BLOCKAGE and COMPULSION. (4 examples)

Form variants of the Vertical Palm that do not encode movement as part of the stroke (VP+Ø) derive from 'stopping' and operate primarily in relation to the interactional axis. They occur in contexts where either the speaker or the addressee are conceptualised as the origin of a discursive force. The enactment of 'stopping' motivating the gesture is applied by locating the gesture accordingly along the interactional axis, either immediately at the speaker's body (Example 1) or in a location situated towards the addressee (Example 2).

In example 1, the speaker has digressed from the topic 'what I'd like to learn' with a long story about a time she met strangers in a bar and managed to communicate with them even though they were Deaf. Although she begins to introduce new information ("and I used to know"), she decides the digression needs to end. As she interrupts her story with a pause and the discourse marker 'anyway', she locates a Vertical Palm gesture in her personal space.

<u>Example 1</u>

S1: and I found it all out using like symbols and sign language because they were all Deaf
S2: really?
S1: ...all of them, were Deaf
yeh and I used to know [......] anyway...*erm* what i'd like to learn *ehumha ha*.
VP-Ø
S2: *ha ha*

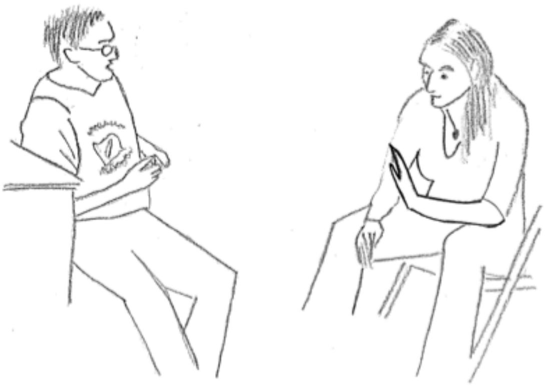


Figure 4: Personal space along interactional axis

The speaker's performance of the VP gesture in her personal space reflects her construal of her digression as a communicative force of which she is the origin. In combination with the discourse marker 'anyway', the VP performance effectively stops that force and she proceeds to the relevant topic ('*erm* what I'd like to learn').

In the next example, Speaker 1 has reported that he would like to meet the Queen and Speaker 2 is offering his understanding of that report. However when as part of that understanding he says "stay the night" he immediately interrupts himself to negate the non-literal interpretation of his expression (i.e. sleep with the Queen). As he says "not with her", he locates a Vertical Palm gesture along the interactional axis towards his addressee (in this instance the orientation of the palm is lateral because of assimilated features from a previous gesture, however the palm is still oriented outwards towards the addressee).

<u>Example 2</u>

S1: have a cup of tea with her, i think go round and have a cup of tea, have a conversation
Time to get to know one another
S2: Stay the night, **[not with her]** but stay the night where they dinner... breakfast
VP-Ø
S1: ha hahaha in Buckingham palace yeh
S2: Yeh in the palace
S1: yehyeh that'd be fine

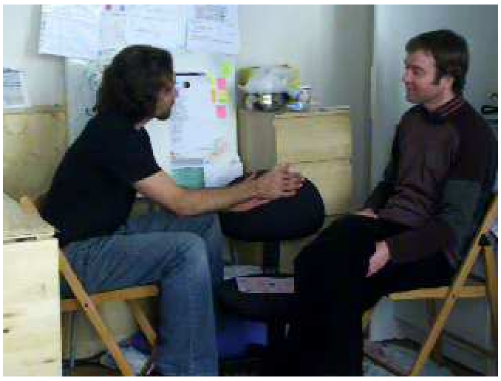


Figure 5: Inter-personal space along interactional axis

Although the speaker interrupts himself, his performance of the Vertical Palm towards the addressee indicates that it is the addressee's potential objection that is construed as a force, to which the act of stopping re-enacted by the Vertical Palm is applied. These examples establish the interactional axis but so far are unaffected by the modality axis.

In other contexts, it is a conceptualised referent that may be construed as the origin of a force e.g. an unwanted implication or a false assumption to be negated. In this case, the force is conceptualised as located away from both the speaker and the addressee, and accordingly the gesture is located diagonally away from the interactional axis, along what I am now arguing is the modality axis (Example 3).

The conversational pair in example 3 have been discussing the topic 'a job you'd like to have'. When Speaker 2 wrongly assumes that Speaker 1 aims to be a politician, Speaker 1 explicitly negates his proposition by saying "my position is not to be a politician". As he negates, he performs a Vertical Palm gesture with lateral movement of both hands (VP-Lateral), which he performs in a location outwards and diagonally away from both himself and his addressee.

<u>Example 3</u>

S2: But if you didn't like politics… you wouldn't talk about it… so much.
S1: I don't like what's happening... I don't like the way politics work, which is why…
S2: OK, you disagree with some of the … sorry, some of the
S1: Yeh, I disagree definitely with certain things that are happening...
So my position is [not to be a **politici**<u>an…</u> my pos]ition is to make sure…
VP-Lateral
It can get kind of complicated
B: Sounds a bit like…

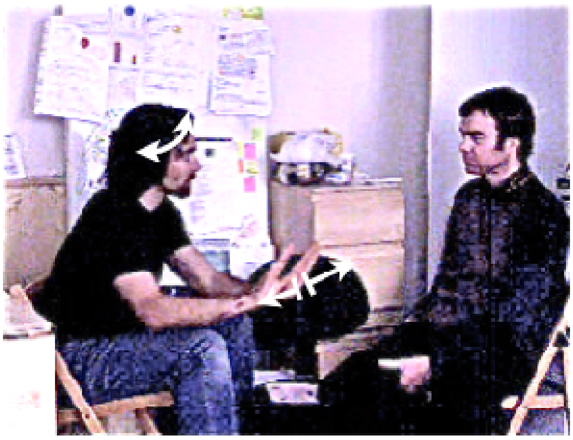
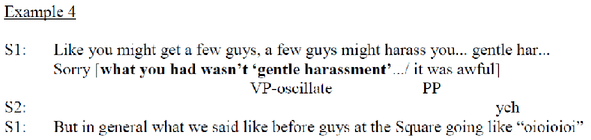


Figure 6: Location diagonally away from interactional axis

Here it is the implied proposition 'speaker/be a politician' that constitutes a force, while the gesture's lateral movement creates an extended barrier construed as a blockage to that force (sometimes VP-Laterals have an abrupt movement, which enacts 'throwing off' rather than extending a barrier via 'molding'). By negating the proposition linguistically and simultaneously locating the gesture in a space away from the interactional axis, the speaker evidences a construal operation in which the negated object is not only 'distanced' from both speaker and addressee but also blocked from view (cf. [22]). In this case it is the location of the gesture that reflects a modality axis.

Finally, cases arise where the location of speakers' gestures are motivated by both the interactional and modality axis. In this case the location of the gesture is determined by the interactional axis while movement incorporated into the gesture stroke encodes negation as distance metaphorically via the enactment of 'rubbing away' (VP-Oscillate). The position of the gesture along the interactional axis reflects to whom the negation is most relevant (Example 4).

In Example 4 the speaker is apologising to the cameraman (S2 – not the person pictured next to him) for referring to his experience of being mugged as “gentle harassment”, then goes on to explain he was talking more in general terms. As part of his apology he says “what you had wasn’t gentle harassment”, and with this (meta-linguistic) negation he performs a Vertical Palm gesture which he oscillates to encode negation and locates towards the cameraman.



Example 4

```
S1:   Like you might get a few guys, a few guys might harass you... gentle har...
      Sorry [what you had wasn’t ‘gentle harassment’.../ it was awful]
                                VP-oscillate              PP
S2:                                                                ych
S1:   But in general what we said like before guys at the Square going like “oioioioi”
```

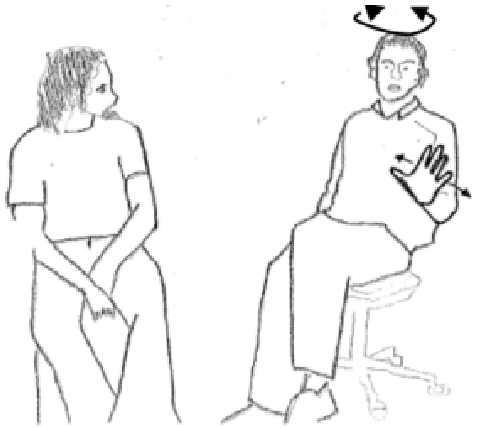


Figure 7: Interactional and conceptual axes combined

As a potentially inappropriate term, ‘gentle harassment’ is reified by the speaker and erased by the oscillation of his VP (derived from ‘rubbing away’). Location of the gesture towards the addressee reflects the intended recipient of the broader speech act of apology, while the horizontal axis effectively becomes a modality axis by providing metaphoric meaning of negation to the movement via the underlying action of rubbing away.

4. Discussion

This paper only offers a brief exploration of the Vertical Palm along with its interactional and conceptual complexity. Nonetheless, I hope to have shown how an interactional and modality axis inter-relate with a gesture associated with negation. The interactional axis is structured by the presence of an addressee (or addressees), while the modality axis reflects construal of negation as distance from the speaker. These empirical observations converge with models of negation in cognitive linguistics [1, 2, 19].

Movements of the Vertical Palm that encode conceptual distance occur horizontally or diagonally away from the speaker’s body, and importantly for an ‘interactive’ gesture they occur obliquely to the interactional axis (I have not observed examples of negations being expressed as distance towards the addressee along the interactional axis). The underlying action of ‘holding away’ [18] may account for movements and locations along the modality axis (examples 3 and 4), but it does not always account for locations along the interactional axis so well. When speakers perform the Vertical Palm as they interrupt themselves or each other i.e. along the interactional axis, this imposes a barrier rather than holds away (lest they be holding themselves or others away from action, rather than the other way round). At least in the examples considered here, locations along the interactional axis reflect construals of the origin of actions or thoughts to be blocked, stopped, negated, etc.

The logic of the form-meaning variations of the Vertical Palm gesture is thus derived from conceptualisation of negation in terms of conceptual force and distance mapped onto physical actions encoded in gestures (stopping, holding away, pushing, etc.), while at the same time the gesture operates along an interactional axis that is characteristic of such recurrent gestures. Models of gesture space thus need to take into account at least both these axes.

To conclude, the inter-relation of interactional and conceptual axis (in this case, negation as distance or a modality axis) emerges as an important feature of gestures associated with negation. Since previous research shows intimate connections between gesture and negation at the level of linguistic organisation too [23-28], we see how a traditionally ‘linguistic’ concept is an embodied system—conceptually, grammatically, and gesturally.

5. Acknowledgements

I wish to thank Daryl Johnson and Margaret Dowens for useful comments on an earlier draft, as well as the University of Nottingham Ningbo China for a Travel Grant to attend the conference. I am also grateful for two anonymous reviews from the GESPIN review board.

6. References

1. Chilton, P., *Negation as maximal distance in discourse space theory*, in *Negation: Form, figure of speech, conceptualization*, S. Bonnefille and S. Salbayre, Editors. 2006, Publications universitaires François Rabelais: Tours. p. 351-378.
2. Chilton, P., *Language, space and mind. The conceptual geometry of linguistic meaning*. 2014: Cambridge University Press.
3. Bressem, J., S. Ladewig, and C. Mueller, *Linguistic Annotation System for Gestures*, in *Body - Language - Communication (HSK)*, C. Mueller, et al., Editors. 2013, de Gruyter. p. 1098-1124.
4. Mueller, C., J. Bressem, and S. Ladewig, *Towards a grammar of gestures: A form-based view*, in *Body - Language - Communication (HSK)*, C. Mueller, et al., Editors. 2013, de Gruyter. p. 707-733.
5. Ladewig, S., *Recurrent gestures*, in *Body - Language - Communication (HSK)*, C. Mueller, et al., Editors. 2014, de Gruyter. p. 1558-1574.
6. McNeill, D., *Hand and Mind. What gestures reveal about thought*. 1992, Chicago: University of Chicago Press.
7. McNeill, D., *Gesture and thought*. 2005, Chicago: Chicago University Press.
8. Cienki, A., *Usage events of spoken language and the symbolic units we (may) abstract from them*, in *Cognitive processes in language*, K. Kosecki and J.

Badio, Editors. 2012, Peter Lang: Frankfurt am Main. p. 149-158.

9. Mueller, C., *Forms and uses of the Palm Up Open Hand. A case of a gesture family?*, in *The semantics and pragmatics of everyday gestures*, R. Posner and C. Mueller, Editors. 2004, Weidler Buchverlag: Berlin. p. 234-256.

10. Kendon, A., *Gesture. Visible action as utterance*. 2004: Cambridge University Press.

11. Calbris, G., *Elements of meaning in gesture*. 2011, Amsterdam/Philadelphia: John Benjamins Publishing Company.

12. Streeck, J., *Gesturecraft. The manu-facture of meaning*. Gesture Studies (GS), ed. C.M. Adam Kendon. 2009, Amsterdam/Philadelphia: John Benjamins Publishing Company.

13. Mueller, C., *Gestural modes of representation as techniques of depiction*, in *Body - Language - Communication (HSK)*, C. Mueller, et al., Editors. 2014, de Gruyter. p. 1687-1701.

14. Sweetser, E. and M. Sizemore, *Personal and interpersonal gesture space: Functional contrasts in language and gesture*, in *Language in the context of use: Cognitive and discourse approaches to language and language learning*, A. Tylet, Editor. 2008: Mouton de Gruyter. p. 25-51.

15. Langacker, R., *Foundations of cognitive grammar. Vol 1, Theoretical prerequisites*. 1987, Stanford: Stanford University Press.

16. Sweetser, E., *Regular metaphoricity in gesture: bodily-based models of speech interaction*, in *Actes du 16e Congrès International des Linguistes (CD-ROM)*1998, Elsevier.

17. Cienki, A. and C. Mueller, eds. *Metaphor and gesture*. 2008, John Benjamins: Amsterdam/Philadelphia.

18. Bressem, J. and C. Mueller, *The family of Away gestures: Negation, refusal, and negative assessment*, in *Body - Language - Communication*, C. Mueller, et al., Editors. 2014, de Gruyter. p. 1592-1604.

19. Langacker, R., *Foundations of cognitive grammar. Vol 2, Descriptive application*. 1991, Stanford: Stanford University Press.

20. Johnson, M., *The body in the mind: The bodily basis of meaning, imagination, and reason*. 1987, Chicago: The University of Chicago Press.

21. Harrison, S., *Grammar, gesture, and cognition: The case of negation in English*, 2009, Université de Bordeaux 3.

22. Kryshtaliuk, A., *The image-schematic dimension of English negation*, in *Cognitive processes in language*, K. Kosecki and J. Badio, Editors. 2012, Peter Lang. p. 99-108.

23. Harrison, S., *Evidence for node and scope of negation in coverbal gesture.* Gesture, 2010. **10**(1): p. 29-51.

24. Harrison, S., *Organisation of kinesic ensembles associated with negation.* Gesture, accepted; 2015.

25. Harrison, S., *Grammar, gesture, and cognition: The case of negation in English*, 2009, Universite de Bordeaux 3.

26. Harrison, S. and P. Larrivee, *Morphosyntactic correlates of gestures: A gesture associated with negation in French and its organisation with speech*, in *Negation and negative polarity. Experimental and cognitive perspectives*, P. Larrivee and L. Chungmin, Editors. accepted; 2015, Springer: Dordrecht.

27. Harrison, S., *The expression of negation through gesture and grammar*, in *Studies in language and cognition*, J. Zlatev, et al., Editors. 2009, Cambridge Scholars. p. 421-435.

28. Harrison, S. *The temporal coordination of negation gestures in relation to speech*. Proceedings of TiGeR 2013 - Tilberg Gesture Research Meeting, June 19-21, 2013., 2013.

From establishing reference to representing events independent from the here and now: A longitudinal study of depictive gestures in young children

Vivien Heller [1] *& Katharina Rohlfing* [12]

[1] CITEC, Emergentist Semantics, Bielefeld University, Germany
[2] University of Paderborn, Germany
vheller@techfak.uni-bielefeld.de, kjr@uni-bielefeld.de

Abstract

Based on longitudinal video data of picture-book reading routines, we investigated the employment of depictive practices in young children. Our sequential analysis focused on the interplay between a child's mastering of the sequential organization of the activity and the acquisition of referential practices. Taking into account both the caregiver's and the child's actions, we examined what bodily and verbal resources young children employ for labeling objects and what role the caregiver's interactive demands and support play for the child's achievement of referential communication. Our findings not only demonstrate how depictive gestures are used in early adult-child interaction; they also shed light on how different methods of representation emerge developmentally.

Index Terms: representational gestures, depictive practices, reference, language acquisition, social interaction, sequential analysis

1. Introduction

From early on, children participate in interactionally organized activities. Especially routines such as picture-book reading are geared towards involving the child in referential communication [1-3]. For the child, the challenge is to discover that communication is meaningful in the sense of getting things done with words [4]. First, establishing reference is heavily supported by the adult participant: the caregiver ascertains the child's visual attention, points to pictures and labels them. Yet soon the child is asked to take on tasks, e.g. to localize depicted objects by pointing and to identify referents by labeling. While it is obvious that caregivers refer to objects and events by introducing novel words to their children, it is still an open research question how *children* come to grips with establishing reference.

One strand of research investigating how children learn to refer to objects is represented by an early study by Ninio and Bruner [1]. Based on video recordings of one mother-child-dyad (8 to 18 months), they examined the interplay between the child's mastering of the sequential order of routines, i.e. highly "structured exchange(s) on non-concrete objects" (p. 6), and the ability to make reference to objects. The authors argue that the acquisition of reference does not only include "mastering a relationship between sign and significate, but […] an understanding of social rules for achieving dialogue in which that relationship can be realized" (p. 15). They suggest that it is not so much imitation, but rather the reciprocity of the dialogue structure that is reinforcing the child. This early study has convincingly shown that the acquisition of a lexicon is situated in interactive contexts and based on the *pragmatic* learning of their sequential order. Yet, albeit mentioning non-verbal resources, the study did not pursue a systematic analysis of bodily resources and their role in young children's achievement of reference.

Bodily resources in referential practices became a topic of research in studies interested in cognitive functions of gestures. These studies have highlighted the role of representational gestures as a transitional device to two-word speech [5-9]. On the basis of correlational analyses it is claimed that combinations in which gesture and speech convey different information predict the onset of two-word speech [7]. Supplementary combinations of gesture and speech are assumed to enable children to express increasingly complex ideas and construct more complex utterances, with gestures adding a predicate or an argument to speech [10]. Albeit often based on naturally occurring data, the interactive contexts in which children employed gestures have received comparatively little attention in these studies. Taking a structural perspective, the focus was mainly on the child's gestural and linguistic system. Non-verbal and verbal utterances were often examined in isolation and abstracted from the communicative action-sequences.

In contrast, Liszkowski [11] discusses the role of action contexts for the child's transition from nonverbal to verbal communication. He suggests that a context is formed by preceding action as well as act-accompanying characteristics (such as prosody and gesture shape), which helps children to enter symbolic communication. His argument is that preceding actions provide a background against which, already in young children, communicative means are interpreted or used. It remains to be shown, however, *how* exactly action contexts enable children to make meaning of others' behavior and to productively contribute to referential communication. For this purpose, a sequential analysis of interaction is needed.

In the present study, we will thus address the following questions: What bodily and verbal resources do young children employ for labeling objects depicted in a book? And what methods of representation do they accomplish with their hands when labeling by gesture? What role do the caregiver's interactive demands play for the child's deployment of representational gestures? And (how) does the caregiver's uptake of the child's non-verbal utterance provide resources that can be reused or even combined for achieving more complex communicative actions?

The present study addresses the questions above by drawing on sequential analysis and describing in detail the embodied practices of labeling, depicting objects and activities. It also pays attention to what ways or methods of representation children accomplish with their hands. Our analysis will also shed light on the question whether children pick up gestures that have been previously used by the adult or whether they create gestures 'on the spot' [12] and derive them from actions [13] and [14].

2. Methodology

The methodology draws on the analysis of embodiment-in-interaction and CA. Thus, verbal and nonverbal resources are analyzed with regard to the exact moment of production within a sequentially organized activity, in relation to the ongoing utterances and the actions of the co-participants [15]. The present study explores, first, what kind of semiotic resources are used by child and caregiver, and how they are coordinated with other means to fulfill particular communicative functions. Second, we compare sequences of talk about the same pictures of a picture-book longitudinally to illuminate if and how the child reuses, combines and replaces semiotic resources for participating in the activity and eventually acquires both interactive as well as representational abilities.

Developmental studies have often distinguished two types of gestures, namely deictic gestures on the one hand and symbolic or iconic gestures on the other hand. What is termed "iconic gesture" may encompass, however, a whole range of *depictive practices* [16]. Streeck uses the term "depictive" instead of "iconic" in order to emphasize that gestural depiction is not necessarily grounded in visual resemblance. He suggests that a depictive gesture does not mirror the denoted object but rather offers a construal or analysis of the referent. From this perspective, depictive gestures are "acts of showing the addressee by movements and postures of the hand what something *looks like* or *is like*" (p. 289). A range of depictive practices have been described, for instance acting, handling, abstract motion, modeling, scaping, model-world making [16-18]. Our analyses suggest that some of these practices play a crucial role in the child's emerging ability to establish reference and to depict events.

3. Data

The longitudinal analysis presented here is based on video-recordings of dyadic interactions taking place between four typically developing children aged between 10 and 24 months and their caregivers. The mothers' educational background was comparable; all of them had university degrees. All mothers spent a min. of four hours per day with their child.

Each family was visited at home once every six weeks (12 points in time). Two different activities were video-taped, free play (lasting 20–25 min.) and picture-book reading (lasting 5–10 min.). For the latter activity, the dyads were provided a book; each page showed photographs of two related objects (a spoon on a mug) or a person and an object (e.g. a child on a swing). Using the same book across the 12 sessions allowed us to compare conversations about the same pictures longitudinally. For space reasons, only excerpts from one dyad are presented and analyzed here.

All interactive sequences in which child and/or caretaker employed depictive practices were transcribed in ELAN, following the notation conventions of *Gesprächsanalytisches Transkriptionssystem 2* (GAT2) [19]. The transcripts depict participants' verbal, non-verbal and para-verbal actions in their sequential order. Each intonation unit is notated in a separate line; an intonation unit is realized in one cohesively perceived intonation contour (symbols for rising pitch movements are "?", "," and for falling pitch movements: ";", ".") and shows at least one focus accent (indicated by capital letters and exclamation marks for outstandingly prominent accent). Prosodic features are notated (e.g. "::" for lengthened sounds or syllables) and aspects of turn-taking such as overlap ([ ]) or latching (=) are shown, too. The alignment of nonverbal and verbal behavior is transcribed in separate lines and indicated by vertical bars. All transripts have been checked by two research assistants. On average, the transcribed corpus per dyad encompasses 18.5 minutes.

4. Results

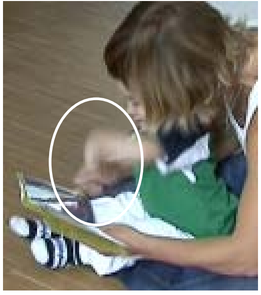
The following section presents episodes in which children use different practices for establishing reference (4.1) and depicting events (4.2). *Acting* is used first (4.1.1); the same manual schema is then later reused, however with a different method of representation, namely as *handling* (4.1.2). The second subsection shows how *acting* (4.2.2) is combined with other semiotic resources and later replaced by *modeling*.

4.1. From pointing and *acting* to *handling* and labeling: establishing reference

Acting is the first depictive practice used by children in all three dyads. It starts to evolve around thirteen months. The following episode, however, stems from a picture-book reading at 16 months. The child, Ole, sits on the mother's lap. The picture both are talking about shows a mug with a spoon.

4.1.1. Extract 1: Stirring/spoon I (Ole: 16 months)

```
01  O   ((turns page))
02      |((points to spoon in the book))|
        |mh::;                          |
03  M   LÖFfel und?
        spoon and
04  O   ((promptly points to mug))
05  M   TASse;
        mug
06  O   fu:p;
07  M   chu::p !AH::!;
08      geNAU;
        exactly
09      chu:p AH::;
10  M   |mit dem LÖFfel     | was kann man damit MAchen;
         with the spoon       what does one do with it
11      |((points to spoon))|
12  O   ((stirring movement))
```



```
13  M   |rüh da REIN und rühren;|
          sti   put inside and stir
        |((stirring movement))  |
14      geNAU.
        exactely
15      da rein und RÜHren.
        put inside and stir
```

The episode is structured into two parts: First, participants establish reference to two objects (l. 2–5). The sequence is then expanded, and the usage of the objects is the conversation's topic (mug: l. 6–9; spoon: l. 10–15).

Establishing reference is an interactive achievement. It requires interpersonal coordination for establishing a) a joint perceptual focus on both the body(part) of the pointing participant as well as the target of the pointing and b) the interactive constitution of the meaning of the referent [20]. Thus, pointing alone is not sufficient for establishing reference. Rather, the act of pointing needs to be tied to "the construals of entities and events provided by other meaning making resources" [21]. In the present episode, reference is established within the context of a labeling sequence which is initiated by Ole (l. 2). By pointing and vocalizing, he draws the mother’s attention to one particular element depicted on the page. The mother orients to his action as asking for a label. With her next turn, she does two things at once: First, she provides the label and thus identifies the object Ole points at (l. 3). Now, joint reference to the spoon is established. Second, she attaches a coordinating conjunction (“and?”). The rising pitch movement turns her response into a “designedly incomplete utterance” [22] and establishes a conditional relevance for Ole to identify the other object. Ole quickly points to the mug, thus completing the turn non-verbally. The label is then again provided by the mother (l. 5). Thus, in both cases, verbal reference is established collaboratively, with a certain division of tasks. While Ole establishes the objects as *perceptual* targets of both participants’ joint attention, the mother *identifies* the objects. This division is designed to steer Ole to take on particular tasks at particular sequential positions – first by means of pointing, and later, as we will see in the following section, by means of verbal resources.

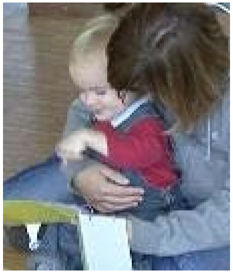
Joint verbal reference to the depicted objects provides the basis for further talk about them. The expansion of the sequence is again initiated by Ole. He produces a vocal gesture (l. 6: “fu:p;”) that depicts the movement and sound of drinking. Thus, he deploys a depictive practice that Streeck terms *acting:* “the gestural action of the hand shows the practical action of a hand” [16, p. 295], [18] and evokes an action. In this case, it is not the hand, but the mouth which represents itself in the action of drinking.

This practice began to evolve at the age of thirteen months. At that time, Ole realized a similar drinking movement with an empty mug – presumably to initiate pretense play. Yet by asking “what is that?” and verbalizing the action (combining noun and verb: “a mug, you can drink out of it”), his mother contextualized this activity as “talking about things” and turned his action into a gesture. In the present episode, taking place three months later, Ole observably performs the movement as a gesture: It is not only realized without an object but within the interactive context of “talking about pictures”. The mother repeats (l. 7: “chu::p”) and expands the depiction which now also represents the result of the action (AH::;” for the quenching of thirst). Finally, she also confirms Ole’s answer (l. 8: “exactly”).

The second expansion is initiated by the mother. Her question (l. 10: “with the spoon, what does one do with it;”) deserves closer analytical attention. It asks for an explication or demonstration of the spoon’s usage on the side of the child. Note that the mother could well have done this herself. Yet her actions are oriented to increasing the child’s participation opportunities. The formatting of her question is recipient-designed to fit the child’s receptive and productive abilities: The syntactical structure (with the prepositional phrase in the pre-front field, a maximally distinct slot, and two focus accents on the most important linguistic elements) is geared to facilitate the child’s understanding of the question. Also the topic of the question is tailored to fit the child’s productive abilities: It asks for the usage of the spoon and prompts Ole to deploy the depictive practice of *acting*. Indeed, he produces the relevant next action by pretending to hold a spoon and to stir it in a virtual mug. The mother displays her understanding of Ole’s action by repeating the movement and labeling the action. In the following episode, which takes place six weeks later, the stirring movement is used as a resource again, albeit with a different semiotic strategy.

4.1.2. Extract 2: Stirring/spoon II (Ole: 17 months)

```
01  O  ((turns page))
02     |((points to mug in the book))|
       |pf::                         |
03  M  °h:::;
04     was ist DAS?=
       what is that
05     [=ne TASse    ] mit EInem?
         a mug          with a
06  O  [points to mug]
07     |ÖFfel;              |
        oon
       |((stirring movement))|
```



```
08  M  LÖFfel;
       spoon
09     [geNAU::;                     ]
        exactely
10  O  [((repeats stirring movement))]
11     !(H)EIß!;
         (h)ot
12  M  ja ist HEIß;=
       yes it’s hot
13     =muss man PUSten?
        do you have to blow
14  O  f::
15  M  f: f:
16     dann kann man die milch TRINken;
       then you can drink the milk
```

The sequence is launched in exactly the same way as the first one (4.1.1): Ole points to a detail in the book and produces a vocalization which resembles either the sound of blowing or drinking (l. 2). Taking a deep breath (l. 3), the mother displays her attention and excitement and then produces a question (l. 4) which takes the recurrently used format “what is that?”. The second part of the mother’s turn is again “designedly incomplete” (l. 5) and formulates the beginning of the answer that Ole is expected to provide. In contrast to the first episode, Ole ‘joins’ this turn. In overlap with his mother, he points to the mug in the book, thus making the turn – and the task of establishing reference – a collaborative achievement which is now accomplished ‘at one go’. This fine interpersonal coordination shows how skilled Ole has become in anticipating and exploiting the sequential order. This also becomes evident in the further process: Without delay, Ole continues the incomplete turn.

For his response, Ole employs semiotic resources that have been used in the previous episode (section 4.1.1), but a) by different participants, and b) successively, i.e. in different sequential positions. These are assembled now and recomposed into a simultaneous construction [17, p. 3]: while Ole utters the lexical label “ÖFfel;”, he realizes the stirring movement (l. 6). Note that before, the movement was employed to depict the usage of the object. In this instance, however, the movement does not depict the action of stirring but the object itself. Thus, the spoon is “indirectly represented by a schematic act that ‘goes with’” it, a practice that Streeck [16, p. 293] terms *handling*: “a motor schema or prehensile posture is coupled with an affordance of the referent”.

Remarkably, this different method of gestural representation is accompanied by verbal labeling. Both semiotic resources mutually elaborate each other. The gesture enhances the intelligibility of the still incomplete phonological form; at the same time, the verbal label contextualizes the gesture as a *handling* that depicts the spoon (and not as an *acting* that depicts the stirring). This raises the question whether the co-occurrence of verbal resources and the new depictive practice is a coincidence or reflects new abilities. We argue that both resources not only mutually elaborate each other but also have in common a particular method of representation: In *acting*, an action is represented by a pretense action. In *handling*, however, the method of representation makes use of conventional knowledge, in this case knowledge about the common usage of the object. Here, the relation between the gesture and the denotatum is an indirect one. Likewise, the relationship between a verbal sign and a denotatum is based on conventional knowledge.

How could Ole acquire this knowledge? We suppose, it could be achieved in finding an answer to the question “what does one do with it?” (extract 1). Note that conventionality was already inscribed in this question in which the neutral pronoun “man” (“one”/“you”) was deployed. Furthermore, the objects were labeled as exemplars of a class (*a* mug, *a* spoon) and characterized in terms of their common usage. In the second episode, the mother consistently uses generic pronouns (l. 13, 16) and stresses conventional uses and procedures. Likewise, Gelman et al. [23] have found that mothers typically produced generic phrases in the context of book-reading. Thus, the mother’s interactive demands as well as the formatting of her question provide the infrastructure for Ole to develop a new method of meaning making.

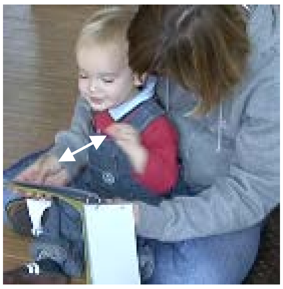
The new semiotic strategy – *handling* in concert with verbal labeling – enables Ole to represent an object independently from the picture. Pointing allowed him to direct his mother’s attention to an object but not to specify himself in which regard the target of pointing was relevant. This had to be achieved in a separate action by the mother’s labeling and her subsequent questions. In the second instance, Ole’s share in establishing the meaning of an object is much larger: He labels the referent and at the same time depicts in which regard the spoon is meant to be relevant. On this basis, new aspects can become the topic of talk. In this case, Ole mentions another feature that typically ‘goes with’ the object (or its content): “hot” (l. 11). The mother’s subsequent question “do you have to blow?” (l. 13) leads to an enactment of further behaviors. Thus, a series of actions – stirring, blowing, drinking – is represented (albeit not in this order). This achievement is taken further when Ole starts to deploy yet another depictive practice: *modeling*.

4.2. From *acting* and labeling to *modeling* and depicting events: constituting contexts transcending the here and now

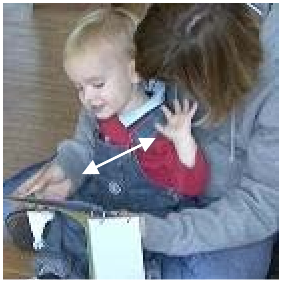
The next episode stems from the same reading session when Ole is 17 months old. The picture shows a child on a swing.

4.2.1. Extract 3: Swinging I (Ole: 17 months)

```
01  M  !OH!
02  O  |((points to picture))|
       |ATa;                 |
03  M  !SCHAU!kel;
        swing
04  O  oh NOno;
       (a boy)
05  M  ein JUNge;=
       a boy
06     geNAU:;
       exactely
07     der junge SCHAUkelt, (---)
       the boy swings
08     /HIN (.)  /
        back
       /und he:r;/
        and forth
09  O  |HIN-                     |
         back
       |moves hand back and forth))|
```



```
10  M  /HIN     /
        back
       /und HER;/
        and forth
11  O  |!ATS!  !ATS!                   |
       |((moves open hand back and forth))|
```



```
12  O  hin hnd;
       back
```

Overlapping with the mother’s display of exitement (l. 1), Ole points to the picture and simultaneously realizes a vocalization (l. 2), which is maybe an attempt to articulate “swing”. Both the mother and Ole label in quick succession of turns what is visible on the picture (l. 3–5). No questions are used here to stepwise foreground individual elements. Reference is established without effort, and mainly by verbal resources.

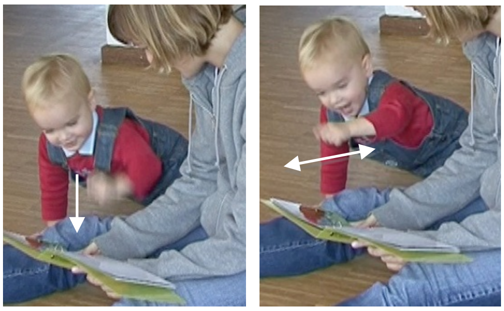
After the protagonist is introduced, the mother expands the labeling sequence (l. 7–8) by *describing* the depicted action in the book. She does so both verbally and prosodically by realizing „back and forth” rhythmically, with the two metrical feet /back (.)/ and /and forth/. The speech rhythm evokes a sense of the swinging movement and thus provides a cue to the meaning of the words. Ole immediately repeats one part of the mother’s verbal utterance (“back”) and simultaneously displays his understanding by moving his left hand back and forth (l. 9). Thus, he again realizes a simultaneous construction [17, p. 3]. His hand takes the shape of a pincer grip, thus depicting the action of holding fast (from the child’s perspective) or moving the swing (from the adult’s perspective). The depictive practice he deploys here is again *acting*: the hand represents the hand in performing the action of holding fast or moving the swing. After the mother has repeated the rhythmic description “back and forth” once more (l. 10), Ole reenacts the activity of pushing, this time opening the hand and realizing two stronger movements (l. 11). The latter are temporally aligned with two vocalizations which contain elements of the German word “anschubsen”, thus representing ‘stronger pushing’ with another multimodal action package. Note that the static picture only shows a child on a swing. Thus, Ole depicts an action that has to be inferred from the picture. This action precedes the visible event.

Finally, he also repeats the verbal phrase describing the event of “swinging”, this time articulating two parts of it (l. 12).

To sum up, Ole has again created a depictive gesture ‘on the spot’ [12]. Yet it is important to note that the gesture occurred *after* the mother had described the action verbally. Thus, the particular context, i.e. the mother’s verbal description, provided the decisive infrastructure for picking up the verbal resource and developing another acting gesture. In contrast to the first episodes, verbal and nonverbal means are now acquired as a *package*. They are reused in a second round of reading that takes place only a few minutes later.

4.2.2. Extract 4: Swinging II (Ole: 17 months)

```
01  M  guck mal;
       look
02     was MACHT der junge da;
       what is the boy doing there
03     |ups-                   |
       |((places Ole on her lap))|
04     der ist im GARten der junge.
       he is the garden          the boy
05  O  at
06     ((crawls away))
07     DA:te;
08     |$^1$da !EI!        $^2$ha                 |
        in there            (forth)
       |(($^1$moves fist downwards, $^2$back/forth))|
```



```
09  M  |ja HIN und HER schaukelt der;      |
         yes back and forth     he swings
       |((moves hand palm-down back/forth))|
```

```
10  O  HIN:-
       back
11  M  HIN und HER;
       back and forth
12  O  es !EK!;
       (it/is gone)
13  M  |und er fällt     | NICHT runter von der schaukel;
       and he does not fall from the swing
       |((points to boy))|
14  O  hm:;
15  M  der kann das schon GUT;
       he’s already good at that
```

The second instance of talk about the child on the swing is initiated by the mother. Her question, taking the typical form “what is the boy doing there?” (l. 2), displays her expectation that Ole is now able to contribute to the description of the picture. As he is busy with being placed in her lap, she bridges the transition relevant place [24] by producing another turn that provides general information (l. 4) but does not deny him the opportunity to answer the open question himself. Ole repeats one element (l. 7: “DA:te” for „garden“) and then produces a complex turn comprising two verbal building blocks that are temporally aligned with two gestures (l. 8). First, his hand forms a fist. The verbal utterance “in there” instructs the mother how to perceive Ole’s hand in this context: as a token for the protagonist depicted in the book.

This practice has been described as *modeling*: “the hand places some generic object, specified by the moment‘s talk, before the interlocutor, who is invited to include it in her visual imagination of the topical scene.” [16, p. 292]. Modeling uses a different representational method which enables Ole to perform new communicative actions. The cupped hand embodies the boy and represents it independently from the picture-book. In contrast to *acting* and *handling*, which depict transient actions, *modeling* allows Ole to represent an object or protagonist for a longer stretch of time. The embodied protagonist can now perform actions that are not depicted in the book: First, Ole moves his fist downwards and shows how the protagonist is seated on the swing. The fist is then moved back and forth, now simulating the motion of swinging. Thus, embodying the protagonist more enduringly, Ole’s hands have decomposed the event into a series of actions. After his mother has reformulated his utterance (and repeated the movement as an *abstract motion*, with flat hand palm-down, l. 9), Ole even depicts another event (l. 12: “it/is gone”): the child falls down. With this, he clearly transcends the visible context of the picture-book and begins to constitute a (imagined) context [25], independent from the here and now.

Thus, Ole has again developed a different method of representation. In contrast to the methods used for *acting* and *handling*, the manual acts do not make use of the usual actions of the hand anymore. Instead, Ole uses his hand creatively and develops a *hand shape* that enables him to represent the protagonist more enduringly and to depict series of actions. These achievements form the building blocks of more complex discursive practices such as narrating and explaining. In our data, *modeling* is often combined with other depictive practices and coordinated with more complex verbal resources, often comprising two or more words.

5. Summary and discussion

Our analyses suggest that at first, depictive practices are used to identify referents. At the beginning of the picture-book reading routine, the child deploys pointing to make an object the *perceptual* target of joint attention. Using pointing, he relies on the mother’s verbal labeling in the subsequent turn. Mastering the sequential organization of labeling sequences and having available *acting* practices, reference is soon established ‘at one go’. This provides the basis for more complex communicative actions such as describing actions and depicting series of events by *modeling*.

It could be observed that the child assembled and combined semiotic resources which had been formerly distributed among different participants and different turns. Tracking talk about one and the same picture longitudinally demonstrated that the child increasingly produces simultaneous constructions. Whereas new verbal resources were picked up from the mother’s contributions, depictive practices were developed by the child. Even more remarkable, the child did not merely assemble semiotic resources but reused them with new methods of representation: In extract (2), the stirring movement that was initially used as *acting* was adapted to represent the object *(handling)*; in extract (4), the movement of pushing was applied to the *modeling* of the protagonist. Thus, in both cases, new depictive practices – which included more conventional ways of representation – were abstracted from pretense action (extract 1 and 3) or developed on the basis of *acting practices* (extract 2, 4). This finding provides first evidence not only for the fact that different methods of representation exist in early childhood [26] but also for how these different methods emerge. In addition, this finding also supports the claim that depictive practices evolve as an abstraction from actions [13] and, as our analysis demonstrates, on the basis of acting practices itself.

Thus, in depicting by gesture, Ole increasingly uses his hands instrumentally. With *modeling*, he deploys a particularly versatile way of meaning making that allows him to represent objects or protagonists over a longer stretch of time. The analysis suggests that he uses this capability to tackle more demanding communicative tasks. He decomposes actions, depicts series of events and even constitutes contexts of talk that are independent from the here and now. Given the more demanding tasks, it comes as no surprise that he combines *modeling* with other depictive practices and also begins to realize more complex verbal resources.

The finding that the use of depictive practices does not rely on imitation but reflects the child's ability to constructively develop new ways of establishing meaning does not imply, however, that these achievements happen *in vacuo*. On the contrary, our analysis shows that the adult's sensitivity to the child's actions, her question design, her multimodal descriptions and reformulations provide the decisive infrastructure for the development of new semiotic resources.

6. Conclusions

The study presented episodes from picture-book reading. Although depictive practices in our data also occur in free play, picture-book reading seems to provide an especially fertile context for developing new methods of representation [3]. Indeed, the activity of 'reading pictures' could have played an important role in these achievements. 'To see pictures', Streeck [16, p. 286) reminds us, basically means to see something *in them*. The same holds for depictive practices, although to different degrees. In *acting* and *handling* we still see the hand acting. When performing and observing *abstract motions* and *modeling*, however, we disattend the fact that they are actions of the hand and instead see something *in them*. To better understand the role of picture-book reading for developing new methods of representation, different activity contexts should be compared systematically.

7. References

[1] Ninio, A. and Bruner, J., "The achievement and antecedents of labelling", J. Child Lang. 5(01):1–15, 1978.

[2] Marcos, H., "How adults contribute to the development of early referential communication", European Journal of Psychology of Education 6(3): 271–282, 1991.

[3] Rohlfing, K.J., Grimminger, A. and Nachtigäller, K., „Gesturing in joint book-reading", in B. Kümmerling-Meibauer et al., K. Nachtigäller and K.J. Rohlfing [Ed], Learning from picturebooks. Perspectives from child development and literacy studies. Routledge, 99–116, 2015.

[4] Bruner, J., "Child's talk. Learning to use language", Norton, 1983.

[5] Capirci, O., Iverson, J.M., Pizzuto, E. and Volterra, V., "Gestures and words during the transition to two-word speech", J. Child Lang. 23(03):645–673, 1996.

[6] Butcher, C. and Goldin-Meadow, S., "Gesture and the transition from one- to two-word speech: when hand and mouth come together", in D. McNeill [Ed], Language and gesture. Cambridge Univ. Press, 235–257, 2000.

[7] Goldin-Meadow, S. and Butcher, C., "Pointing toward two-word speech in young children", in S. Kita [Ed], Pointing. Psychology Press, 85–107, 2003.

[8] Pizzuto, E.A., Capobianco, M. and Devescovi, A., "Gestural-vocal deixis and representational skills in early language development", Interaction Studies 6(2):223–252, 2005.

[9] Pizzuto, E. and Capobianco, M., "The link and differences between deixis and symbols in children's early gestural-vocal system", Gesture 5(1):179–199, 2005.

[10] Özçalişkan, S. and Goldin-Meadow, S., "Do parents lead their children by the hand?", J. Child Lang. 32(03):481–505, 2005.

[11] Liszkowski, U., "Two sources of meaning in infant communication: preceding action contexts and act-accompanying characteristics", Phil. Trans. Royal Soc. B 369(1651):1–9, 2014.

[12] Behne, T., Carpenter, M. and Tomasello, M., "Young children create iconic gestures to inform others", Developmental Psychology 50(8), 2049–2060, 2014.

[13] LeBaron, C., Streeck, J., "Gestures, knowledge, and the world", in D. McNeill [Ed], Language and gesture, Cambridge Univ. Press, 118–138, 2000.

[14] Capirci, O., Contaldo, A., Caselli, C. and Volterra, V., "From action to language through gesture. A longitudinal perspective", Gesture 5(1):155–177, 2005.

[15] Streeck, J., Goodwin, C., LeBaron, C.D., "Embodied interaction in the material world: an introduction", in ibid. [Ed], Embodied interaction. Language and body in the material world, Cambridge Univ. Press, 1–26, 2011.

[16] Streeck, J., "Depicting by gesture", Gesture 8(3):285–301, 2008.

[17] Kendon, A., "Semiotic diversity in utterance production and the concept of 'language'", Phil. Trans. R. Soc. B 369(1651):1–13, 2014.

[18] Müller, C., "Redebegleitende Gesten. Kulturgeschichte — Theorie — Sprachvergleich", Arno Spitz, 1988.

[19] Selting, M., et al, "A system for transcribing talk-in-interaction: GAT 2", Gesprächsforschung 12:1–51, 2011.

[20] Stukenbrock, A., „Referenz durch Zeigen: Zur Theorie der Deixis", Deutsche Sprache 37:289–315, 2009.

[21] Goodwin, C., "Pointing as situated practice", in S. Kita [Ed], Pointing. Psychology Press, 217–241, 2003.

[22] Koshik, I., "Designedly Incomplete Utterances: A Pedagogical Practice for Eliciting Knowledge Displays in Error Correction Sequences", Res. on Language & Social Interaction 35(3):277–309, 2002.

[23] Gelman, S. A., Chesnick, R. and Waxman, S.R., "Mother-child conversations about pictures and objects: Referring to categories and individuals", Child Development 76:1129-1143, 2005.

[24] Sacks, H., Schegloff, E.A. and Jefferson, G., "A Simplest Systematics for the Organization of Turn Taking for Conversation", Language 50:696–735, 1974.

[25] Heller, V. and Rohlfing, K.J., "Coming to grips with narrating: Depictive practices as a springboard into constructing fictional contexts", Paper presented at the 37. Jahrestagung der DGfS, Leipzig, March 4–6, 2015.

[26] Cartmill, E. Novack, M. and Goldin-Meadow, S., "The development of embodied iconicity: representational gesture over the first 5 years of life", Paper presented at the 37. Jahrestagung DGfS, Leipzig, March 4–6, 2015.

The use of teaching gestures in an online multimodal environment: the case of incomprehension sequences

Benjamin Holt[1]*, Marion Tellier*[2]*, Nicolas Guichon*[3]

[1]Université Lumière Lyon 2, Labex ASLAN, Laboratoire ICAR and Aix-Marseille Université, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France
[2] Aix Marseille Université, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France
[3]Université Lumière Lyon 2, Laboratoire ICAR

Abstract

This study aims to describe the gesture use and multimodal behavior of future French teachers in Lyon, France during videoconferenced pedagogical interaction with learners of French in Dublin, Ireland. These multimodal conversations were recorded, transcribed and annotated using an annotation scheme that was designed for this multimodal corpus. The use of gesture was measured and compared quantitatively during sequences of incomprehension and sequences in which there was no manifest incomprehension between participants to measure the extent to which gesture was used to repair incomprehension. Results show that contrary to expectations, teacher trainees did not use gestures to repair incomprehension.

Index Terms: exolingual communication, incomprehension, teaching gesture, gesture rate, computer-mediated communication, videoconferencing, multimodality

1. Introduction

A project called *français en première ligne* [1] started in 2002 that allowed distant learners to interact asynchronously with native French speakers. With increased bandwidth and the ubiquity of webcams, the project became synchronous in 2006 [2], which allowed interlocutors to see each other, including their hand gestures. It has already been shown that gestures play an important role in the teaching and learning of foreign languages, [3]–[6] but what happens when the pedagogical interaction takes place in a videoconferencing environment where the gesture space is drastically reduced [7]? This case study, devised to prepare a PhD, will focus on webcammed pedagogical interaction between master's students in France and undergraduate learners of French in Ireland. More specifically, the use of hand gestures during incomprehension sequences will be quantitatively and qualitatively measured.

2. Theoretical Framework

In foreign language acquisition, the understanding of input is of the utmost importance and is preliminary to production [8], [9]. Therefore we must take a look at moments that favor the modification of input. Such moments occur during *exolingual conversation*, which has been defined by Porquier [10] as interaction that takes place between interlocutors who do not share equal linguistic proficiency in the language that is being spoken, and who consciously make adaptations according to the disparity. Native speakers use *foreigner talk*, altering and simplifying the way they speak [9]. Porquier & Py [11] think that *exolingual competence* is probably learned and not innate, and Sarré [12] agrees that negotiation of meaning is an important interactional skill.

Input is modified most during negotiation of meaning and incomprehension sequences [11], [13] where the interlocutors strive to facilitate comprehension. As Sarré [12] points out, videoconferencing environments are propitious for negotiation of meaning, and for this study we will use the four-step model devised by Varonis & Gass [13]:

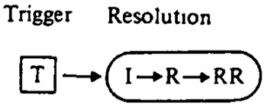


Figure 1: Incomprehension sequence model [13, pp. 74]

A *trigger* (T) is an utterance from the native speaker that is not understood by the non-native speaker, and is usually found at the beginning of the sequence [13]. The trigger can come from many things including lexical items, the context, the difficulty of the task, etc. [8], [14]. The trigger never exists out of context, and thus does not become a trigger unless it elicits a reaction from the non-native. Until this happens, it remains a "potential trigger" [13, pp. 78]. If the non-native gives feedback [15], this is considered to be the *indicator* (I), which "halts the horizontal progression of the conversation and begins the downward progression, having the effect of 'pushing down' the conversation rather than impelling it forward" [13, pp. 75]. If the native speaker does not ignore the indicator, s/he gives a *response* (R), which is the most important step because it reveals the negotiation strategies used by the teacher. Yanguas [14, pp. 82] explains that:

> "responses are perhaps the most vital element in the negotiation routines because, on the one hand, they include the feedback provided to the interlocutor to fix the communication problem and, on the other, they are pushed output on the part of the speaker."

Videoconferenced interactions have been shown to produce more negotiation sequences than other types of online conversation [12]. All of the reparation strategies enumerated by Long [9] are strictly verbal and prosodic, completely leaving out nonverbal means of repairing incomprehension, such as the use of coverbal gestures. Coverbal gestures, which have been defined as "the movements of the hands and arms that we see when people talk", [16, pp. 1] "offer themselves as a second channel of observation of the psychological activities that take place during speech production – the first channel being overt speech itself" [17, pp. 350]. Coverbal gestures are part of exolingual competence [11], [18] and aid in foreign language comprehension [5], [19], [20]. Furthermore, the act of teaching influences and changes nonverbal behavior [21]. In a pedagogical context, gestures and body language are used more consciously to fulfill various functions, which incites us to consider this type of gesture as *teaching gestures* [3].

Teaching gestures, which are strongly related to the speech of the teacher, facilitate foreign language comprehension and memorization of lexis [4], [22]–[25]. Just as speech is modified during foreigner talk [9], [15], so too are gestures [26]. An experimental study by Tellier & Stam [6] examined the same future French teachers in two different situations: either sitting across from a native French speaker or from a non-native. As in a game of Taboo, these master's students had to explain a word to their partner without saying it. Tellier and Stam's analysis showed that when explaining words to non-native partners, future French teachers made gestures that lasted longer and were more iconic. Furthermore, the gesture rate increased and the gesture space grew larger when speaking to non-natives. Wagner *et al.* [27] mention that if the use of gesture is the best way to clarify ambiguity, an interlocutor will tend to use it. For these reasons, analysis of gesture during incomprehension sequences is indispensable.

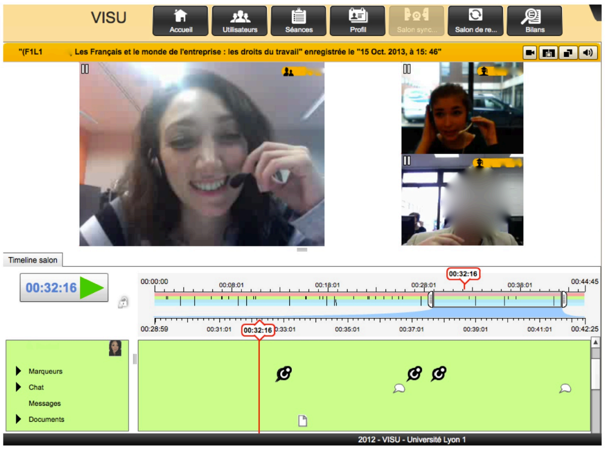
There are certain particularities of webcammed interaction that can have an effect on communication and gesture production. In general, temporal delay, poor audiovisual quality and inadvertent disconnections are prone to making synchronous foreign language teaching more difficult, and teachers must learn to anticipate these technical problems [28]. Direct eye contact is impossible, because in order to give the impression of direct eye contact, one must look directly at the camera and off screen, therefore paradoxically not really looking at one's partner [7], [29]. Whether or not the webcam enhances or facilitates online communication is still up for debate. Wang [30] found that "video has been greatly appreciated by the distance participants and its pedagogical values are indispensable to language learning at a distance"; Develotte, *et al.* [31] found that visual clues provided learners with supplementary psycho-pedagogical, cultural and linguistic information; Develotte *et al.* [2] found that the video contributed to a more fluid interaction, as nonverbal clues helped to complete oral instructions. However, Develotte *et al.* [31, pp. 309] note that it can be distracting to see one's partner "either because it contradicts the oral message, because it makes no sense and adds nothing to it, or because it distracts the learner's attention," and that certain interlocutors prefer the audio channel, as "it appears that the use of a webcam image is more important in terms of its availability as a possible resource in case of need than as a favored type of communication." Guichon & Cohen [32] compared audio-synchronous and videoconferenced pedagogical interaction and found that the audiovisual condition enhanced neither the feeling of online presence, nor comprehension.

Most important, the webcam's field of view typically captures only the head and shoulders of each interlocutor, dramatically reducing the gesture space and leaving many gestures off screen and not visible [7]. As noted above, Tellier & Stam [6] found that when talking to non-natives, teacher trainees tended to make large gestures, thus widening the gesture space. In order for gestures to be visible onscreen, however, gestures must be made in a small space close to the face and shoulders, which is not natural, especially during exolingual conversation. Teacher trainees must adapt, as Develotte *et al.* [31] explain: "the video window can be compared to a theatre stage that the teacher trainees use to enact their role: they learn to adapt their gestures to the size of the stage." Because of this, we postulate that most gestures that are visible onscreen are allocentric [33], meaning that they were made in the center gesture space with a communicative purpose.

A teacher trainee is someone who is developing techniques to strategically monitor the online pedagogical interaction, and has three channels through which to deploy an array of strategies: the verbal channel, the gestural channel, and the textual channel. Verbal strategies used to facilitate comprehension may include use of synonyms, definitions, examples, translations, repetitions, reformulations, metalanguage, questions, and verifications of comprehension[1]. Foreign language teachers can consciously use gestures to inform, evaluate and animate [3]. The text chat is an invaluable tool for online language teachers because it can be used, among other things, to correct oral productions without interrupting the flow of the conversation [34], facilitate comprehension by repeating what is said orally [2] and allow learners to communicate in real time, modifying input and production, and responding to feedback all the while focusing on the form and structure of the language [14]. We shall see how teacher trainees divide their pedagogical moderation between each of these three channels during the interaction, especially during sequences in which there is manifest incomprehension.

3. Corpus

Our corpus is from the ISMAEL project[2] [35], which organizes webcammed multimodal interactions between future teachers of French at the University of Lyon 2 in France and business students who are learning French at Dublin City University in Ireland. Our corpus from the fall semester of 2013 is comprised of six weekly 45-minute interactions between groups of two (one teacher trainee, one learner) or three (one teacher trainee, two learners). For this study, we have chosen four interactions from the first week: two groups of two and two groups of three with durations of 41:19, 33:25, 43:53 and 42:14, respectively. The theme of this session was the French business world and its constituents (35-hour work weeks, paid vacation, coffee breaks, strikes, etc.). These interactions took place on an online language-learning platform called *Visu* (see figure 2) [34] and were saved using a screen recorder.



"(F1L1 Les Français et le monde de l'entreprise : les droits du travail" enregistrée le "15 Oct. 2013, à 15: 46"

Figure 2: A multimodal interaction on *Visu*

[1] Strategies gathered from our corpus.

[2] InteractionS et Multimodalité dans l'Apprentissage et l'Enseignement d'une Langue (in English: Interactions and multimodality in the learning and teaching of a language).

4. Research questions and hypotheses

The aim of this study is to find out whether or not teacher trainees use their gestures to repair incomprehension and whether or not they make more gestures during incomprehension sequences. Our hypotheses are that teacher trainees will make more gestures during incomprehension sequences than during normal sequences (i.e. no manifest incomprehension) and that the text chat will be used more during incomprehension sequences.

5. Methodology

All four video recordings were transcribed and annotated using ELAN, which is open-source software designed for annotating multimodal interactions [36]. The spoken verbal messages were transcribed using the ICOR transcription convention [37], and the textual messages sent by the teacher were copied without alteration. We then annotated the incomprehension sequences, which begin with the teacher trainee's response (see figure 1) and end when the teacher trainee chooses to end the sequence.

For gesture annotation, we classified all hand gestures visible on the screen into six categories, including the four dimensions proposed by McNeill [16], [38]: iconics, metaphorics, deictics and beats, and added emblems and non-identifiable gestures. To compare use of gesture with other channels used, we annotated the oral verbal strategies listed above and counted the number of messages sent in the text chat window. Indeed, the chat window can be used for verbal strategies such as synonyms and translations, but since the purpose of this study is to measure the use of each channel, we left a single category for the text chat messages and kept verbal strategies as oral only. To remove some subjectivity and test the viability of our transcription guide, we calculated an inter-annotator agreement percentage for 21 gestures and for incomprehension sequences during 10 minutes of video, achieving ample agreement scores of 67% and 78%, respectively.

6. Analysis

The gesture, verbal and text chat strategies were counted for each interaction, and the percentage of each was compared during "normal" sequences (NS) and during incomprehension sequences (IS), with the total number listed under "entire interaction" (EI). Table 1 shows the duration of each type of sequence in seconds, and Table 2 shows the repartition of strategies counted among the three channels.

	EI	NS	IS	%NS	%IS
Duration	10066	9252	814	91.9%	8.09%

Table 1: Duration of sequences types (in seconds)

	EI	NS	IS	%NS	%IS
Verbal	611	511	100	**69.0%**	**74.6%**
Text	43	29	14	**3.91%**	**10.4%**
Gesture	221	201	20	**27.1%**	**14.9%**

Table 2: Number of strategies counted per modality per type of sequence

We see that during normal sequences, the audio channel was used for nearly 70% of all strategies observed, with text accounting for only 3.91% and gestures 27.1%. During incomprehension sequences, use of the audio channel rose slightly while use of the text chat nearly tripled, validating our second hypothesis. However, contrary to our predictions, the gesture channel was used less during sequences in which there was manifest incomprehension, dropping from 27% to 15%.

Even though the gesture channel does not seem to be a favored method of repairing incomprehension, it was used in some cases. In order to illustrate the usefulness of this channel, we will show two examples[1], in which teacher trainees used gesture to help repair incomprehension. In the following example, Victor (teacher trainee) uses a gesture[2] (see figure 3) to help explain the meaning of a 35-hour workweek to his learner during an incomprehension sequence:

VIC so **thirty-five hours is the maximum allowed by the law**

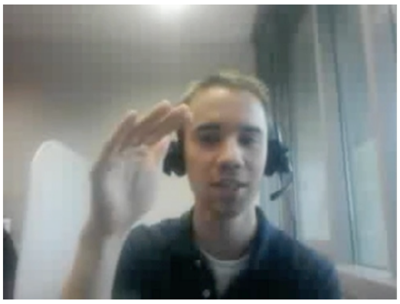


Figure 3

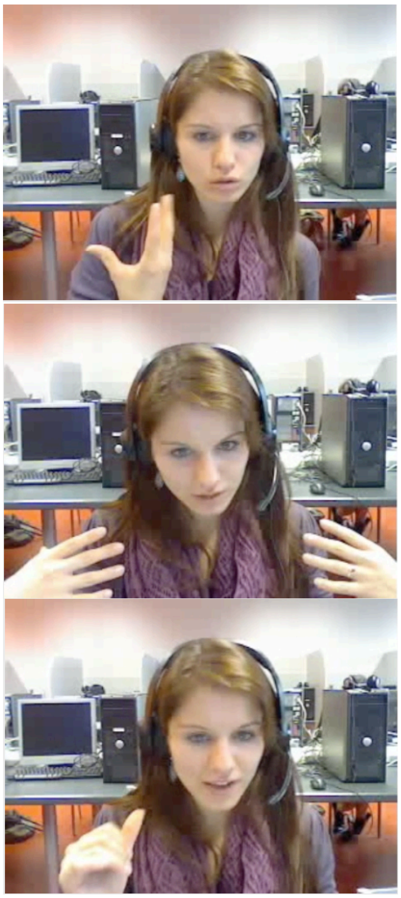
Victor's informational gesture [3], which is placed clearly within the webcam's field of view, complements his explanation of the 35-hour workweek law. The gestural representation of the concept of "maximum" may be easier for the learner to decode than the verbal explanation in a foreign language, and when used in conjunction with the word "maximum," which is the same in English as in French, helps to disambiguate its meaning.

In the following example, Melissa (teacher trainee) uses gestures (see figures 4, 5 and 6) to verify her own comprehension of what Alejandra (learner) is trying to say:[3]

ALE hum: in hum spain I hum: (0.9) lots of things/ (.) (inaud.) hum hot/ (0.3) (inaud.) (1.0) [°(inaud.)°]
MEL [°hot° yes/ **hot drinks/]**
ALE (yes) (.) (0.5) yes and (0.4) hum hum hum (0.6) hum the beer/ [hum]
MEL [ah] (0.7) ok it's very hot in spain (1.3) **it's [hot in] spain/**
ALE [hum:] (0.9) yes because I live in hum the canary islands\ (1.4) [th-]
MEL [and] so you (0.2) **you drink beers/ (.) to [cool** off a bit\]
ALE [no n-] norma[lly] no but ((laughs))
MEL [°no/°] (1.1) o[k\]

[1] Translated from French to English by us.
[2] The part of the sentence corresponding temporally with the gesture is in boldface text. The gestures are pictured in order of occurrence.
[3] : = lengthened word; (#.#) = duration of pause greater than 200 milliseconds; (.) = micro pause shorter than 200ms; / = rising intonation; \ = falling intonation; <((description)) production> = description of production; °° = whispered or spoken very softly; [] = overlapping turns; UPPERCASE = perceptual salience; x = inaudible syllable; (inaud.) = inaudible series of syllables

ALE [the] hum (inaud.) in the: (0.6) the break/ (.) for coffee/ and ((laughs)) x (0.5) [and drink] a glass
MEL [°ok\°] (1.6) ok



Figures 4, 5 and 6

Melissa uses gestures to give feedback and to show what she understands, so that Alejandra can adjust her output accordingly. Feedback given through the gesture channel may be easier to decode than verbal feedback in a foreign language. First, Melissa simultaneously verifies her own comprehension of what Alejandra is talking about and proposes the words “hot drinks” accompanied by a gesture that represents a drink (see figure 4). During Melissa’s following turn, she checks her own comprehension by asking if it’s hot in Spain, and then repeats the question while using her hands to show that it’s hot (see figure 5) to ensure comprehension. In figure 6, Melissa uses a French emblematic gesture of drinking to verify her own comprehension of drinking beer to cool off and to show Alejandra what was understood by her output in the foreign language. Alejandra’s negative response shows that Melissa’s question was understood, and due to Alejandra’s low level of comprehension overall during this conversation, it is reasonable to believe that Melissa’s gesture aided comprehension.

To further our understanding of gesture usage, we calculated the gesture rate (the number of gestures divided by the number of words spoken) for the entire interaction (EI), for normal sequences (NS), and for incomprehension sequences (IS) for each of the four teacher trainees (see table 5 on next page). Then, in table 3, we divided the gesture rate calculated for incomprehension sequences by the gesture rate calculated for normal sequences to see the difference. If the ratio is greater than one, then the gesture rate is higher during incomprehension sequences, and the inverse is true if the ratio is less than one. It is interesting to compare teacher trainees because each interlocutor has his or her own interactive and gestural profile which does not disappear during online interaction [7]. We found that teacher trainees 2 and 4 had a higher gesture rate during incomprehension sequences (ratio>1), whereas teacher trainees 1 and 3 had a lower gesture rate during incomprehension sequences (ratio<1), which supports [7]’s claim that each person uses gestures differently. The overall gesture rate for all four interactions (see table 4) shows that for our corpus, the gesture rate during incomprehension sequences was lower than the gesture rate during normal sequences.

Interaction	1	2	3	4
Incomp/norm	0.56	1.30	0.64	1.7

Table 3: Ratio of incomprehension sequence gesture rate/normal sequence gesture rate

	IS	EI	NS	IS/NS
# gestures	20	221	201	0.76
# words	1265	10974	9709	
Gesture rate	0.016	0.020	0.021	

Table 4: Overall gesture rate

7. Discussion/conclusion

This was a preliminary study to examine how teacher trainees use multimodality in a videoconferencing-based teaching setting. The fact that these teacher trainees, overall, did not produce more gestures during incomprehension sequences than during normal sequences invalidates our first hypothesis. Perhaps the presence of a screen, keyboard and webcam alters the exolingual behavior of online teachers. Since the keyboard is widely used during technical problems [2], [39]–[41], teacher trainees might resort to using it during any type of problem, not distinguishing technical problems from linguistic incomprehension or miscommunication. Since it is impossible to make hand gestures and type at the same time, it seems in retrospect that our hypotheses may have been mutually exclusive. It would be interesting therefore to repeat this study in a videoconferencing environment without the possibility of sending text chat messages. The fact that these teacher trainees made fewer gestures during incomprehension sequences does not void the possibility of these gestures having specific, novel purposes worthy of study. One goal of our future research is therefore to define the functions of body language during videoconferenced pedagogical interaction.

It is difficult to draw generalizable conclusions based on these four teacher trainees’ interactions. It is important to keep in mind that this was the first week of online interactions and that these trainees were not attuned to harnessing the affordances of the videoconferencing platform. Whereas the use of verbal strategies is known by most teachers and can be effectively transferred from face-to-face to online pedagogical interaction, making visible hand gestures in front of the webcam and using all three channels harmoniously is an entirely new skill that must be developed. To address this corpus it was necessary to identify strategies channel by channel, but it is clear that teacher trainees often use multiple channels simultaneously, two and sometimes three at a time, and thus the interaction between the channels must be studied to better understand the multimodal nature of these interactions. This exploratory research was necessary to become familiar with a pedagogical situation that has been seldom studied, and for our future research we aim to see whether or not the teacher trainees progressively develop new strategies and ways of exploiting and combining the affordances offered by the videoconferencing platform over the course of a semester.

8. Acknowledgements

The authors are grateful to the ASLAN project (ANR-10-LABX-0081) of the University of Lyon for its financial support within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) of the French government operated by the National Research Agency (ANR). Special thanks is given to Erika Dussire and Jorane Soubesty for annotating sections of the corpus to obtain our inter-annotator agreement score.

Interaction	1 IS	1 EI	1 S	2 IS	2 EI	2 NS	3 IS	3 EI	3 NS	4 IS	4 EI	4 NS
# gestures	4	43	39	5	27	22	5	113	108	6	38	32
# words	340	2193	1853	413	2776	2363	186	2745	2559	326	3260	2934
Gesture rate	0.01	0.02	0.02	0.01	0.01	0.01	0.03	0.04	0.04	0.02	0.01	0.01

Table 5: gesture rates by teacher trainee

9. References

[1] C. Develotte and F. Mangenot, "Discontinuités didactiques et langagières au sein d'un dispositif pédagogique en ligne," *Glottopol*, no. 10, pp. 127–144, 2007.

[2] C. Develotte, N. Guichon, and R. Kern, "'Allo Berkeley ? Ici Lyon... Vous nous voyez bien ?' Étude d'un dispositif de formation en ligne synchrone franco-américain à travers les discours de ses usagers," *Alsic*, vol. 11, no. 2, pp. 129–156, 2008.

[3] M. Tellier, "Dire avec des gestes," *Fr. Dans Monde Rech. Appl.*, no. 44, pp. 40–50, 2008.

[4] M. Tellier, "The effect of gestures on second language memorisation by young children," *Gesture*, vol. 8, no. 2, pp. 219–235, 2008.

[5] M. Tellier, "Usage pédagogique et perception de la multimodalité pour l'accès au sens en langue étrangère," in *La place des savoirs oraux dans le contexte scolaire d'aujourd'hui*, R. Bergeron, G. Plessis-Bélair, and L. Lafontaine, Eds. Montréal: Presses Universitaires du Québec, 2009, pp. 223–245.

[6] M. Tellier and G. Stam, "Stratégies verbales et gestuelles dans l'explication lexicale d'un verbe d'action," in *Spécificités et diversité des interactions didactiques*, V. Rivière, Ed. Paris: Riveneuve éditions, 2012, pp. 357–374.

[7] J. Cosnier and C. Develotte, "Le face à face en ligne, approche éthologique," in *Décrire la conversation en ligne, Le face à face distanciel*, C. Develotte, R. Kern, and M.-N. Lamy, Eds. Lyon: ENS Éditions, 2011, pp. 27–50.

[8] C. Cornaire and C. Germain, *La compréhension orale*. Baume-les-Dames: CLE Internationale, 1998.

[9] M. Long, "Native speaker/non-native speaker conversation and the negotiation of comprehensible input," *Appl. Linguist.*, vol. 4, no. 2, pp. 126–141, 1983.

[10] R. Porquier, "Communication exolingue et apprentissage des langues," in *Encrages : acquisition d'une langue étrangère*, 1984, pp. 17–47.

[11] R. Porquier and B. Py, *Apprentissage d'une langue étrangère : contextes et discours*. Le mesnil-sur-l'Estrée: Imprimerie Nouvelle Firmin Didot, 2008.

[12] C. Sarré, "Computer-mediated negotiated interactions: how is meaning negotiated in discussion boards, text chat and videoconferencing?," in *Second language teaching and learning with technology: views of emergent researchers*, S. Thouësny and L. Bradley, Eds. Dublin: Research-publishing.net, 2011, pp. 189–210.

[13] E. M. Varonis and S. Gass, "Non-native/Non-native Conversations: A Model for Negotiation of Meaning," *Appl. Linguist.*, vol. 6, no. 1, pp. 71–90, 1985.

[14] Í. Yanguas, "Oral computer-mediated interaction between L2 learners: it's about time!," *Lang. Learn. Technol.*, vol. 14, no. 3, pp. 72–93, 2010.

[15] M. Long, "Input, interaction, and second-language acquisition," *Native Lang. Foreign Lang. Acquis.*, vol. 379, pp. 259–278, 1981.

[16] D. McNeill, *Hand and Mind*. Chicago: The University of Chicago Press, 1992.

[17] D. McNeill, "So you think gestures are nonverbal?," *Psychol. Rev.*, vol. 92, no. 3, pp. 350–371, 1985.

[18] B. Azaoui, "Multimodalité des signes et enjeux énonciatifs en classe de FL1/FLS," in *Le corps et la voix de l'enseignant : théorie et pratique*, Paris: Ed. Maison des langues, 2014, pp. 115–126.

[19] D. Sime, "'Because of her gesture, it's easy to understand'. Learners' perception of teachers' gestures in the foreign language class.," in *Gesture: second language acquisition and classroom research*, Routledge., S. G. McCafferty and G. Stam, Eds. New York, 2009, pp. 259–280.

[20] S. Kellerman, "'I see what you mean': the role of kinesic behaviour in listening and implications for foreign and second language learning.," *Appl. Linguist.*, vol. 13, no. 3, pp. 239–258, 1992.

[21] L. Cadet and M. Tellier, "Le geste pédagogique dans la formation des enseignants de langue étrangère," *Cah. Théodile*, no. 7, pp. 67–80, 2007.

[22] M. Macedonia, K. Müller, and A. Friederici, "The impact of iconic gestures on foreign language word learning and its neural substrate.," *Hum. Brain Mapp.*, vol. 32, no. 6, pp. 982–998, 2011.

[23] W. C. So, C. S. Chen-Hui, and J. L. Wei-Shan, "Mnemonic effect of iconic gesture and beat gesture in adults and children: Is meaning in gesture important for memory recall?," *Lang. Cogn. Process.*, vol. 27, no. 5, pp. 665–681, Jun. 2012.

[24] M. Macedonia and K. von Kriegstein, "Gestures enhance foreign language learning," *Biolinguistics*, vol. 6, no. 3–4, pp. 393–416, 2012.

[25] M. L. Rowe, R. D. Silverman, and B. E. Mullan, "The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language," *Contemp. Educ. Psychol.*, vol. 38, no. 2, pp. 109–117, Apr. 2013.

[26] T. W. Adams, "Gesture in Foreigner Talk," University of Pennsylvania, 1998.

[27] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Commun.*, vol.

57, pp. 209–232, 2014.

[28] C. Dejean-Thircuir, N. Guichon, and V. Nicolaev, “Compétences interactionnelles des tuteurs dans des échanges vidéographiques synchrones,” *Distance Savoirs*, vol. 8, pp. 377–393, 2010.

[29] H. De Chanay, “La construction de l’éthos dans les conversations en ligne,” in *Décrire la conversation en ligne, Le face à face distanciel*, Lyon: ENS Éditions, 2011, pp. 145–172.

[30] Y. Wang, “Supporting synchronous distance language learning with desktop videoconferencing,” *Lang. Learn. Technol.*, vol. 8, no. 3, pp. 90–121, 2004.

[31] C. Develotte, N. Guichon, and C. Vincent, “The use of webcam for teaching a foreign language in a desktop videoconferencing environment,” *ReCALL*, vol. 22, no. 3, pp. 293–312, 2010.

[32] N. Guichon and C. Cohen, “The Impact of the Webcam on an Online L2 Interaction,” *Can. Mod. Lang. J.*, vol. 70, no. 3, pp. 331–354, 2014.

[33] J. Holler, M. Tutton, and K. Wilkin, “Co-speech gestures in the process of meaning coordination,” in *Proc. 2nd GESPIN—Gesture and Speech in Interaction Conf., Bielefeld, 5–7 September 2011*, 2011.

[34] N. Guichon, M. Bétrancourt, and Y. Prié, “Managing written and oral negative feedback in a synchronous online teaching situation,” *Comput.-Assist. Lang. Learn.*, vol. 25, no. 2, pp. 181–197, 2012.

[35] N. Guichon, F. Blin, C. Wigham, and S. Thouësny, “ISMAEL Learning and Teaching Corpus. Dublin, Ireland: Centre for Translation and Textual Studies & Lyon, France: Laboratoire Interactions, Corpus, Apprentissages & Représentations.” 2014.

[36] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, “Elan: a professional framework for multimodality research,” in *Proceedings of LREC*, 2006, vol. 2006, p. 5th.

[37] Groupe ICOR, *Convention ICOR*. Lyon: Université de Lyon. http://icar. univ-lyon2. fr/documents/ICAR_Conventions_ICOR_2007. doc, 2007.

[38] D. McNeill, *Gesture & Thought*. Chicago and London: The University of Chicago Press, 2005.

[39] C. Develotte, R. Kern, and M.-N. Lamy, Eds., *Décrire la conversation en ligne: la face à face distanciel*. Lyon: ENS, 2011.

[40] C. Kerbrat-Orecchioni, “Conversations en présentiel et conversations en ligne : bilan comparatif,” in *Décrire la conversation en ligne, Le face à face distanciel*, Lyon: ENS Éditions, 2011, pp. 173–195.

[41] A. Liddicoat, “Enacting participation : hybrid modalities in online video conversation,” in *Décrire la conversation en ligne, Le face à face distanciel*, Lyon: ENS Éditions, 2011, pp. 51–69.

Multimodal Counter-Argumentation in the Workplace: The Contribution of Gesture and Gaze to the Expression of Disagreement

Jérôme Jacquin [1, 2]

[1] Department of French, University of Lausanne, Switzerland
[2] Department of Language and Information Sciences, University of Lausanne, Switzerland

jerome.jacquin@unil.ch

Abstract

This paper examines argumentative talk-in-interaction in the workplace. It focuses on counter-argumentative references, which consist of the various resources that the opponent uses to refer to the origin/source of his/her opposition, namely the confronted position and the person who expressed it. Particular attention is paid to the relationship – in terms of sequential positioning and referential extension – between reported speech, polyphony, pointing gestures and shifts in gaze direction. Data are taken from workplace management meetings that have been recorded in New Zealand by the Language in the Workplace Project.

Index Terms: disagreement, argumentation, reported speech, polyphony, pointing gestures, gaze direction, talk-at-work.

1. Introduction

Argumentation can be defined as a specific way of managing disagreements and conflicts ([1]–[4]). This definition highlights the importance of the contact between a position that is defended in a discourse and its contestation in a counter discourse. This paper focuses on the way this contact is multimodally embedded and managed in workplace meetings held in New Zealand. While the expression of disagreement has been studied extensively (see [5] for references in a discourse-analytical perspective), there has been little research on the multimodal resources – reported speech, polyphony, pointing gestures and shifts in gaze direction – that the opponent combines in context within "multimodal Ensembles" ([6]) or "multimodal *Gestalts*" ([7]) to make reference to the origin of his/her opposition, namely the confronted position and the person who expressed it.

Following an overview of the theoretical framework (2.), I will briefly describe the data I use (3.). I will then analyze several extracts that show how counter-argumentative references are multimodally embedded in their sequential context and, how they subsequently contribute to the general organization of the argument (4.). In the discussion that follows, I will summarize the communicative effects of the phenomena observed (5.).

2. Theoretical framework[1]

2.1. Argumentation as verbal practice

Argumentation is often defined as the verbal activity of convincing people. This definition is highly problematic, as has been shown for example by Angenot ([8]). It has been suggested that the issue of analyzing argumentative practices can be more satisfactorily undertaken by approaching argumentation as a verbal way of managing disagreements and conflicts ([1], [2], [4]). In this sense, argumentation emerges when a difference of opinion not only arises, but "crystallizes" ([9]) through the construction and consolidation of opposing positions with respect to a controversial question (e.g. "How should we reduce social inequalities?", "What will be the name of the baby?", "Should we abolish the death penalty?")[2]. Such a definition implies argumentation is both a specific way of connecting utterances (i.e. the textual dimension of argumentation) and managing relationships with others (i.e. the interactional dimension of argumentation). It follows that an interdisciplinary approach to argumentation is needed ([10]), namely through the combination of notions and methods provided by Conversation Analysis ([11], [12]), Interactional Linguistics ([13]) and Text Linguistics ([14], [15]). Despite their differences, these subfields of Linguistics can be adopted as complementary approaches to examine the use of linguistic units in the construction and negotiation of social reality in talk-in-interaction. Particular attention will be paid here to previous studies that stress the importance of embodiment in social interaction ([16], [17]).

As previously stated, this paper focuses on counter-argumentative reference-making. This requires a description of the different resources that are available to the speaker for making reference to someone or something (2.2.), before considering the different ways these resources are articulated so as to work as multimodal references (2.3.).

2.2. Resources for making reference

2.2.1. Linguistic devices

Argumentation, as has been defined above, relies on dialogism in the bakhtinian sense ([18]). For B to oppose to A's position requires A's position to be "taken into account" – without being "taken in charge" ([19]). This can be achieved in two ways: through reported speech ([20], [21]) and polyphony ([22], [23, Ch. 6]).

Reported speech means using "talk to report talk" ([24, p. 1]), as well as the embedded opinion or point of view. There are different ways of reporting speech, from direct, "depicting" forms, such as quotations, to indirect, "describing" patterns where the reported talk is not syntactically isolated, but integrated in the reporting talk ([25]; see also [20], [26],

[1] An extended version of this section was published in French ([51]).

[2] As shown by Plantin ([1]) and Doury ([52]), the controversial situation can be *in praesentia* (i.e. the opposing positions are defended by two different participants interacting together) or *in absentia* (i.e. at least one speaker argues against a position that no other participant to the interaction defends).

[27]). For example, “She said: ‘I disagree’” *depicts* the discourse and the position, whereas “She said that she disagreed” or “She disagreed” *describes* her discourse and position, without quoting it (see also [28]). As evidenced by the use of “she”, the possibility of reported speech acting as a reference relies heavily on the presence of referential expressions, such as proper names (*Nathan, Mr. X*), descriptions (*the president*), deictics (*I, you*), or anaphora (*he, she*) ([29]–[31]). In the case where there is no referential expression, recipients tend to evaluate reported speech based on its degree of similarity and proximity to previous talk. This leads us to the second linguistic device used in argumentation.

Polyphony is a complex category. Contrary to reported speech, a polyphonic discourse does not contain or embed another discourse, but only the point of view associated with that discourse. As has been demonstrated by Ducrot ([32], see also [33]), a negative formulation such as “this wall is not white” conveys two points of view (POV) which disagree with each other: While POV1 is [this wall is white], POV2 – which is the one endorsed by the speaker saying “this wall is not white” – is [POV1 is false]. Similarly, the adversative discourse marker “but” is polyphonic, as it “give[s] instruction pointing to the presence of voices other than the author’s” ([23, p. 257]). The origin of theses “voices” (i.e. points of view) are implicit, but can be identified: a polyphonic discourse can work as an indirect reference to another discourse and to the speaker who expressed it if the two discourses are spatially or temporally close to one another, such as two columns in a web or newspaper page or two adjacent turns-at-talk in a debate sequence ([34]–[36]).

2.2.2. Gaze direction

As has been frequently noted since the first studies on gaze in social interaction (see [37] for a synthesis), gaze in Western culture is used to manage speakership and recipiency, by indexing who talks to whom. In other words, speakers and recipients tend to look at each other. But although gaze is a resource for the speaker to index the recipient(s) of their talk, continuous gaze is marked and may convey other information (e.g. seduction or aggression). Gaze is then frequently available, both to the speaker and the recipient, to build joint attention on a third party (a person, an object or a direction). However, as gaze’s “home position” ([38]) is the recipient, a gaze shift to another participant at a specific sequential position can be interpreted as a shift of recipiency, or even as a solicitation (e.g. [39]).

2.2.3. Pointing gestures

Gestures cover a larger scope of phenomena ([40]). Pointing gestures, which have been studied extensively, literally point to an element of the context by selecting it as the focus of joint attention [40, Ch. 11], [41], [42]). This deictic resource allows the speaker to make reference to somebody or something independently of what is happening at the verbal level and without having to shift gaze direction. Pointing gestures are then particularly relevant for the study of references in verbal interaction.

2.3. The coordination of the multimodal resources: referential extension and sequential positioning

As has been frequently highlighted in Conversation Analysis and Interactional Linguistics, the actual meaning of a resource – not only a linguistic unit, but also a shift in gaze direction or a pointing gesture – is in a mutually constitutive relationship with its sequential context ([37], [40], [42]), which refers both to the direct pragmatic environment (i.e. the previous and the next actions) and to the broader type of activity participants are performing (e.g. brainstorming, information giving, or decision making). In other words, the referential extension and the sequential positioning of each resource have to be considered dialectically, in order to examine (i) how the different resources are coordinated in such a way as to produce a multimodal reference, and (ii) to which extent discrepancies between the instructions respectively given through the different modes contribute to create meaning in their combination[1].

3. Data

The data that I will consider are taken from a corpus of six video-recorded management meetings, held from 2004 to 2006 at a production company in New Zealand. In these meetings, the 11 managers of the company discuss practical issues (human resources, security, schedule) as well as more long-term developments (business model, company philosophy). These data have been recorded by the Language in the Workplace Project (LWP) at Victoria University of Wellington, New Zealand (for a general overview of the project and data, see [44])[2].

Previous studies suggest that New Zealand English speakers tend to strongly mitigate or even avoid direct expression of disagreement ([45]–[47]). In other words, they display a strong preference for agreement, implicitness and softening strategies such as tag questions, hedges, hesitations and gambits [48]. Stadler notes, in her comparative study of the expression of disagreement in German and New Zealand English, that “New Zealanders’ non-verbal behavior in disagreements differs little from their behavior in neutral speech” ([48, p. ii]), which consists for example of looking at the recipient less directly than in other cultures (e.g. in Germany). However, Stadler’s analysis is purely quantitative and no attention is paid to the situated coordination of the verbal and non-verbal resources that have been identified above. The present paper provides some insights about the role of pointing gestures and shifts in gaze direction in making disagreement accountable and, therefore, in compensating for the New Zealand preference for verbal indirectness.

4. Analysis

The 2-hour management meeting I will focus on was recorded in early 2005. An important portion of it was spent deciding whether to hire a new operator (Sue), who had been separately interviewed by three of the managers a couple of days before the meeting. This meeting is then an occasion for Jeason (JH), the General Manager, Seamus (SB), the Managing Director and Ivo (IS), the Pre Press Manager, to gather and argue their respective views. Both Jeason and Seamus underline the urgency to hire someone and acknowledge Sue’s skills and expertise, although Jeason still expresses doubts about whether the personality will fit in with the company since they have not yet been provided with a reference. In contrast, as the analysis will show, Ivo positions himself as an opponent, by expressing doubts about the relevance of Sue’s specific skills

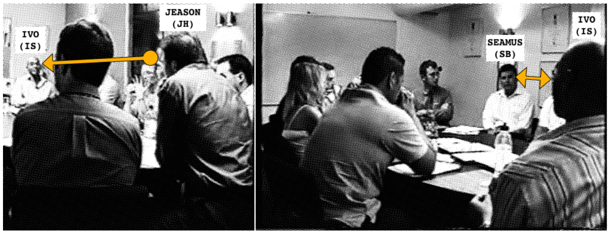
[1] Situations of pure redundancy between the meanings conveyed through the different modes are theoretically possible, but empirically unverified in the previous studies cited.

[2] Two cameras and one audio recorder were used; researchers were not present at the time of recording. All names are pseudonyms, and any identifying material has been removed.

in view of the evolution of the operational workflow the company will have to face in the near future. It should be noted that, before Extract 1, Ivo has been arguing that hiring Sue would be over-hasty, because the current production operators will soon be trained in the position that Sue would fill.

Extract 1a[1]

```
1  SB  %+¶the right time is not when the work arrives
       %looking at IS---------------------------->18
        +looking at IS--------------------------->22
         ¶looking at SB-------------------------->3
2  ??  #1 ((clears throat))
#im1
```



Following work by Ducrot and Nølke on negative formulations ([32], [33]), Seamus' assessment can be considered polyphonic as it combines two contesting points of view (POV). While Seamus takes POV1 [the right time is when the work arrives] into account, he endorses POV2 [POV1 is false]. The disagreement is mitigated, as POV1 is not explicitly attributed to Ivo. However, Ivo is identified as the origin (or at least as representing the contested POV) through Seamus and Jeason's continuous gaze in his direction (see image 1[2]).

Extract 1b

```
3  IS  no ¶ that's right (..) i mean that
       ---¶shift to table------------------------>6
4      (1.2)
5  IS  yeah (.) well it ((sighs)) u:m (..) my-
6      my thing is (...) is ¶ the (1.0) she is a
       ---------------------¶shift to SB--------->9
7      person who can put impost together quickly
8  JH  mhm
9  IS  but ¶ she won't be able to she won't (...)
       ----¶shift to JH-------------------------->
10     know ¶ where to put them (...) that's that's
       -----¶mult. shifts betw. SB and JH-------->39
11     where we [the bottleneck is XX]
```

Ivo stops looking at Seamus during the concession (lines 3-5), but shifts his gaze back again when returning to his argument at line 6 ("my thing is…"). After Jeason's agreement (8), Ivo starts looking at him while introducing his counter-argument through a negative formulation ("but she won't be able…"), before concluding with "that's where the bottleneck is". This contests Jeason's own identification of the "bottleneck" several minutes before and, by shifting his gaze during the negative formulation, Ivo seems to group Seamus and Jeason as people committed to the point of view contested by his negation.

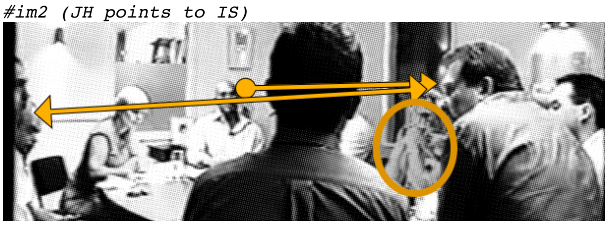
Extract 1c

```
12 SB            [we still need to her to do that
13     would we or i mean what a what sort of
14     position would you
15 IS  the that's where the m- metric side of things
16     comes in
17 JH  yeah
18     %(1.2)
       %shift to JH------------------------------->
19 IS  i mean that's where the % hold-ups are
       -----------------------%shift to IS------>25
20 JH  yeah
21     (...)
22 SB  okay + but we and we haven't got that
       -----+shift to SB------------------------>32
23     (..)
24 IS  not yet=
```

In his reply, Seamus appears to change his strategy. He is still looking at Ivo, but verbally he moves from an inclusive "we" in an assessment (12-13) to directly question Ivo, who is continuously shifting gaze direction very quickly between Seamus and Jeason. Both Ivo's first answer (15-16) and the recompletion (19) are followed by Jeason's agreement. While Seamus looks at Jeason after the first agreement (17-18), he continues gazing at Ivo not only after the recompletion (19) but also after Jeason's second agreement (20). By not looking at Jeason at that sequential position, Seamus seems to display the direct confrontation which is at stake between Ivo and him. When taking his turn at line 22, he begins with a concession ("okay") and continues with a negative formulation ("we haven't got that"), which contests Ivo's reformulated point of view ([we have got that]). Ivo answers "not yet", which concedes to Seamus' assessment without agreeing with it, as the situation he describes is presented as inevitable.

Extract 1d

```
25 JH  =so % what what #2 ivo's saying is that she's
       ----%shift to JH------------------------->28
#im2 (JH points to IS)
```



```
26     not that guru that we talked about
27     [she's not going
28 IS  [yeah she is not % she's not the
       ----------------%shift to IS------------->end
29     [(layout guru X)] she's the one that maybe&
30 ??  [XX XX XXX       ]
31 JH  [um XX XXX       ]
32 IS  &+creates the template really quickly
       -+shift to IS---------------------------->38
33 JH  mhm
34 IS  i mean and and i'm sure any any of the
35     operators like XX do the same
36 SB  yeah
37     (...)
```

In lines 25-26, Jeason reformulates Ivo's argument through indirect reported speech which he addresses to Seamus. However, just before mentioning Ivo, Jeason points to him. In this way, Jeason makes use of a variety of modes (gaze, speech, head orientation, and gesture) to not only show that he is taking Ivo's position into account, but also to mediate the direct confrontation between Seamus and Ivo, who is looking at Jeason. Ivo then takes this opportunity to support his position by providing new arguments (28-35).

Extract 1e

```
38 SB  I mean thi[s well she +
       ---------------------+shift to SB-------->48
39 IS            [she become X
40 SB  i ¶ guess the other thing that comes into
       --¶shift to SB--------------------------->end
41     play too even if the timing's not absolutely
42     perfect (...) u:m
43     (1.0)
44     these people don't grow on trees even with
45     her experience=
46 IS  =yeah
47     (...)
```

[1] See Appendix 6 for transcription conventions.

[2] The fact that Jeason's back is turned, at least in part, to the camera makes the identification of his gaze direction difficult, but by zooming in, it is apparent that his head is clearly turned to Ivo and not to the speaker, Seamus.

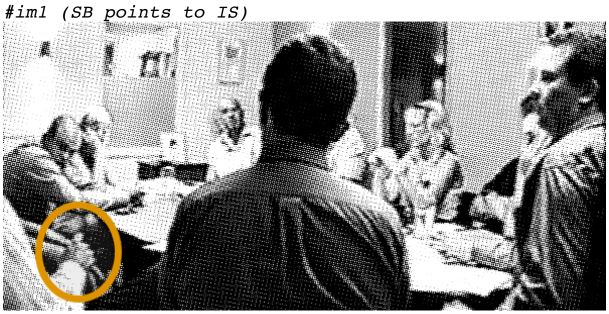
```
48 SB  +you know it's har- they're hard to find
       +shift to IS------------------------------->end
49     especially when you want them
```

Next, while looking at Ivo, Seamus provides a complex argument by articulating two negative formulations: “[P] even if the timing’s not absolutely perfect, [Q] these people don’t grow on trees”. While [P] is a way for Seamus to reformulate and concede the counter-argument that Ivo would, or may, formulate against Seamus’ argument, [Q], which is also polyphonic because of the negation, reformulates and contests Ivo’s position. In combination with the continuous gaze towards Ivo, these negations are used to attribute and make accountable the disagreement in the absence of explicit reference to the contested position.

After the exchange in Extract 1, Seamus continued to highlight the urgency of finding someone in order to deal with the upcoming rush. In Extract 2, although Ivo participates minimally (with the production of only two regulators), he stays at the center of the multimodal attention displayed in both Seamus’ and Jeason’s argumentation[1].

Extract 2a

```
1  SB  %um so (.)+ in the#1 term in the in the scheme
       %looking at IS---------------------------->15
                 +looking at SB------------------>12
#im1 (SB points to IS)
```



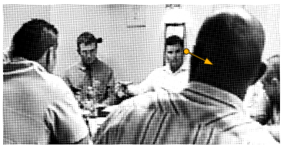
```
2      of things (..) when we're looking at risk
3      and do we d- employ this person or do we
4      put an extra person in (..) that is
5      infinitesimal compared with a (...) um (...)
6      punt that we've already taken
7  IS  mh
8  JH  yeah
9      (1.0)
10 SB  it's just a absolute no brainer to me (..) um
11     (3.5)
```

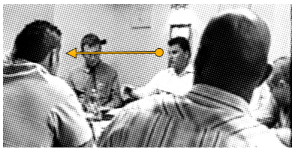
At line 1, Seamus points to Ivo. In the absence of a negative formulation that would convey Ivo’s point of view, this pointing gesture appears as a way to counteract the collegial “we” (lines 2, 3 and 6) and the impersonal formulation (“that is infinitesimal…”), which follows. As before, Seamus keeps looking at Ivo even when Jeason takes his turn to agree with him. Through the use of the directed pointing gesture, the attention is then exclusively focused on Ivo.

Extract 2b

```
12 SB  +and if timing dictates that you you make
       +looking at the table-------------------->18
13     the decision now and we wing it and we work
14     out how she fits or if your #2 if you've
15     decide that #3% she is
       --------------%shift to JH--------------->25
```

```
#im2                              #im3
```



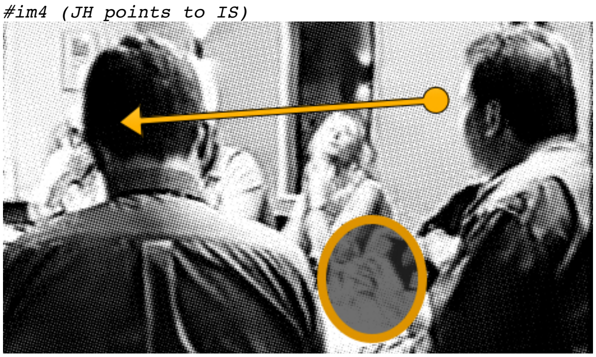


```
16 JH  yeah i mean my my decision is not do we
17     (..) do we need another person the decision
18     is + is she the right person that's (a)
       ---+shift to SB-------------------------->24
19     that's (.) you know
20 SB  yeah (.) is has she the right attitude and
21     (.)
22 JH  yep
23 SB  yeah=
```

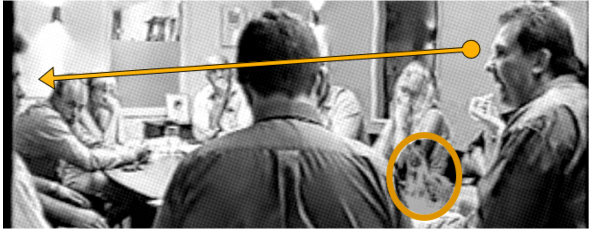
Just after the long pause following Seamus’ conclusion (“it’s just an absolute no brainer to me”, at line 10), Jeason shifts his gaze to the table, while Seamus takes another turn and proposes two possible decisions regarding the timing (12). Although Seamus says “if you have decide[d]” while looking at Ivo, he progressively orients himself to Jeason just before expressing the subordinate clause “that she is…” (see images 2 and 3). This shift seems to function as a repair of recipiency, which displays the participants’ orientation to Jeason’s leadership and professional role in the company. This hypothesis is also supported by Jeason himself, who not only takes his turn while Seamus’ clause is syntactically incomplete, but also begins it with “yeah I mean *my* decision…” (16; but see also the shift to a much more open and collective “*the* decision” at line 17). Collaborating with Seamus (19-23), Jeason reformulates the situation and his point of view on the situation.

Extract 2c

```
24 JH  =yep (.) + is she gonna fit with the team
       ---------+shift to IS-------------------->28
25     properly and % that sort of thing (...) um
       -------------%shift to IS------------------>
26     i'm % keen to #4 talk to darryl and um ivo
       ----%shift to JH------------------------->38
#im4 (JH points to IS)
```



```
27     about where exactly where she will fit
28     +#5 in that (...) as well that's w- that is
       +shift to SB----------------------------->30
#im5 (JH points to IS)
```



```
29     part of the decision because is she the
30     right person + does she have the right
       -------------+shift to table--------------->
31     skills (...) + um (.) um it's gonna be part
       -------------+shift to IS---------------->34
32     of it you know i think (...) on the surface
33     she seems to (of) to have the perfect
34     (set of)+ skills but(...) u:[m]+ le- let's&
       --------+shift to ?------------+shift to IS->
```

[1] During the entire extract, Ivo’s gaze is directed at the current speaker.

```
35 IS                           [um]
36 JH  &just#6 fit that into our workflow how it's
#im6 (JH points to IS)
```

```
37      gonna + work (...) make sure + we've made
        ------+shift to SB-----------+shift to IS--->
38      that  right decision ((long sighs))
```

Both Jeason and Seamus progressively re-orient themselves towards Ivo when Jeason reformulates again the pending questions (see shifts in gaze direction at lines 24 and 25). Referring to further discussions about the issues raised by Ivo, Jeason mentions him explicitly (26) just after pointing to him (image 4). In doing so, Jeason displays that he takes Ivo's point of view in consideration. Jeason maintains the pointing gesture but shifts his gaze to Seamus (28, image 5) as he mentions the issue they discussed together at the end of Extract 2b. Jeason then looks at the table during another reformulation of the pending questions, before shifting his gaze once again to Ivo.

Jeason continues his turn on the state of the decisions, by progressively inserting clauses that reflect the different points of view at stake. Jeason starts with Ivo (see also "you know", 32), referring to Sue's operating skills and how they will fit in the workflow. Also, by pointing to him just before mentioning the "workflow" (image 6), Jeason orients towards Ivo as the expert for this part of the decision-making process. While inserting a new clause ("how it's gonna work…", lines 36-37), Jeason quickly looks at Seamus, before returning his gaze to Ivo and concluding with the decision, which, surprisingly, seems to have already been made: "(let's) make sure we've made that right decision" (37-38). It is worth noting that here the "we" is unclear in regard to its referential extension; it could be an inclusive "we" (at least I and you) – Ivo is part of the decision – or an exclusive "we" (at least I but not you) – only Seamus and Jeason made the decision. This ambiguity can be considered a resource, as Jeason leaves the door open for Ivo to join the decision, even if he has not yet agreed to it.

This last extract is interesting as it shows how Jeason makes the state of the discussion accountable without having to explicitly mention the disagreement at stake.

5. Discussion

The above analysis has confirmed previous studies on typical features associated with the verbal expression of disagreement in New Zealand English: people tend to use only a few explicit resources such as "but" or reported speech quoting or reformulating the contested position; additionally, the frequent use of mitigation strategies – such as hedges ("I think"), gambits ("I mean", "you know"), concessions and hesitations – underlines the dispreference associated with disagreement.

However, the data also highlighted the high frequency and crucial importance of negative formulations for the accountability of disagreement, something which has not been taken into account in previous studies (e.g. in [48]). Negative formulations allow the speaker to uncover (and contest) a point of view without having to attribute it to other speakers in an explicit way. And as the analysis showed, speakers tend to use gaze direction in order to attribute the uncovered point of view and, in that way, make reference to someone as being its origin and as taking the responsibility for it. In other words, gaze direction compensates for the referential ambiguity of negative formulations.

The analysis also underlined the frequency and importance of the unit *we*, whose referential ambiguity can be strategically used in disagreements. When used and interpreted as inclusive (at least *I* and *you*), *we* can carry the idea of a community of values and interests, for example when mentioning the challenges faced by the company. When used and interpreted as exclusive (at least *I*, but without *you*), *we* splits people and creates coalitions ([49]). Most notably, in the data analyzed, some of the instances of *we* were used in combination with a pointing gesture (while in most of the others, instances of *we* were part of negative formulations). By using a pointing gesture in a specific sequential environment, the speaker organizes the referential extension of *we*, by recalling the relevance of a position that has been previously expressed without having to rephrase it and therefore having to provide accounts about it. However, these pointing gestures occur rapidly and infrequently, which seems to confirm their impoliteness for New Zealand English speakers ([48]).

More generally, this paper showed the crucial importance of gesture and gaze direction for the production and interpretation (i.e. the accountability) of disagreements in contexts where the verbal expression of disagreement is strongly mitigated.

6. Appendix: Transcript conventions

Data were transcribed according to ICOR conventions (v. 2013, *Groupe ICAR*):

/ \	Rising and falling intonations
:	Prolongation of a sound
-	Abrupt interruption in utterance
(.) (..) (...) (n)	Pauses (1/4, 1/2, 3/4 second; n = seconds)
MAIS	Emphasis
[YY YYYY]	Overlapping speech
&	Extension of the turn after an overlap
=	Latching
(it; eat)	Speech which is in doubt in the transcript
XX XXX	Speech which is unclear in the transcript
((laughs))	Annotation of non-verbal activity

Gaze has been transcribed with the following conventions, inspired by [50]:

+---+,*----*	Delimits gaze direction for each participant. The symbols +, ¶ and % refer respectively to JH, IS and SB's gaze.
------------>	The phenomenon continues across the subsequent line
------------>8	The phenomenon continues across line 8
#1 #im1	Picture 1, with comments on gestures

7. Acknowledgements

The author has been supported in 2014 and 2015 by the Swiss National Science Foundation [P2LAP1_155110]. The author would like to acknowledge the members of the Language in the Workplace Project – and in particular Janet Holmes, Meredith Marra and Bernadette Vine – for sharing their experience and data with him during his stay at the Victoria University of Wellington, New Zealand. He is most grateful to

Keely Kidner for revising his English text and to the two reviewers for their helpful comments.

8. References

[1] C. Plantin, "Le trilogue argumentatif. Présentation de modèle, analyse de cas," *Langue française*, no. 112, pp. 9–30, 1996.

[2] M. Doury, *Le débat immobile. L'argumentation dans le débat médiatique sur les parasciences*. Paris: Kimé, 1997.

[3] F. H. Van Eemeren and R. Grootendorst, *A Systematic Theory of Argumentation : the pragma-dialectical approach*. Cambridge: Cambridge University Press, 2004.

[4] J. Jacquin, *Débattre. L'argumentation et l'identité au coeur d'une pratique verbale*. Bruxelles: De Boeck Duculot, 2014.

[5] J. Angouri and M. A. Locher, Eds., *Theorising Disagreement / Journal of Pragmatics 44 (12)*. 2012.

[6] G. Kress and T. van Leeuwen, *Multimodal Discourse*, London: Arnold, 2001.

[7] L. Mondada, "Bodies in action: Multimodal analysis of walking and talking," Language and Dialogue, vol. 4, no. 3, pp. 357–403.

[8] M. Angenot, *Dialogues de sourds : traité de rhétorique antilogique*. Paris: Mille et une nuits, 2008.

[9] V. Traverso, "Cristallisation des désaccords et mise en place de négociations dans l'interaction : des variations situationnelles," in *La négociation au travail*, M. Grosjean and L. Mondada, Eds. Lyon: PUL, 2005, pp. 43–69.

[10] J. Jacquin and R. Micheli, "Entre texte et interaction : propositions méthodologiques pour une approche discursive de l'argumentation en sciences du langage," in *Actes du CMLF 2012 - 3ème Congrès Mondial de Linguistique Française*, F. Neveu et al., Eds. Lyon: EDP Sciences (www.linguistiquefrancaise.org), 2012, pp. 599–611.

[11] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language (Baltim).*, vol. 50, no. 4, pp. 696–735, 1974.

[12] J. Sidnell and T. Stivers, Eds., *The Handbook of Conversation Analysis*. Chichester: Wiley-Blackwell, 2013.

[13] M. Selting and E. Couper-Kuhlen, Eds., *Studies in Interactional Linguistics*. Amsterdam: Benjamins, 2001.

[14] M. A. K. Halliday and R. Hasan, *Cohesion in English*. London: Longman, 1976.

[15] J.-M. Adam, *La linguistique textuelle : introduction à l'analyse textuelle des discours*. Paris: Armand Colin, 2008.

[16] J. Sidnell and T. Stivers, Eds., *Multimodal Interaction / Semiotica N°156*. 2005.

[17] J. Streeck, C. Goodwin, and C. LeBaron, Eds., *Embodied Interaction: Language and Body in the Material World*. Cambridge: Cambridge University Press, 2011.

[18] M. M. Bakhtin, *Problems of Dostoevsky ́s Poetics*. Minneapolis: University of Minnesota Press, 1984.

[19] D. Coltier, P. Dendale, and P. De Brabanter, "La notion de prise en charge : mise en perspective," *Langue française*, vol. 2, no. 162, pp. 3–27, 2009.

[20] J. M. L. Munoz, S. Marnette, and L. Rosier, Eds., *Le discours rapporté dans tous ses états*. Paris: L'Harmattan, 2004.

[21] E. Holt and R. Clift, Eds., *Reporting Talk. Reported Speech in Interaction*. Cambridge: Cambridge University Press, 2007.

[22] J. Bres, P. P. Haillet, S. Mellet, H. Nølke, and L. Rosier, Eds., *Dialogisme et polyphonie : approches linguistiques*. Bruxelles: De Boeck Université, 2005.

[23] K. Fløttum, T. Dahl, and T. Kinn, *Academic Voices: Across Languages and Disciplines*. John Benjamins Publishing, 2006.

[24] R. Clift and E. Holt, "Introduction," in *Reporting Talk. Reported Speech in Interaction*, E. Holt and R. Clift, Eds. Cambridge: Cambridge University Press, 2007, pp. 1–15.

[25] H. H. Clark and R. J. Gerrig, "Quotations as Demonstrations," *Language (Baltim).*, vol. 66, no. 4, pp. 764–805, 1990.

[26] L. Rosier, *Le discours rapporté : histoire, théories, pratiques*. Bruxelles: De Boeck Université, 1999.

[27] A. Bangerter, E. Mayor, and S. Pekarek Doehler, "Reported Speech in Conversational Storytelling During Nursing Shift Handover Meetings," *Discourse Processes*, vol. 48, no. 3, pp. 183–214, 2011.

[28] M. Johansen, "Agency and responsibility in reported speech," *Journal of Pragmatics*, vol. 43, pp. 2845–2860, 2011.

[29] M. Charolles, *La référence et les expressions référentielles en français*. Paris: Ophrys, 2002.

[30] S. C. Levinson, "Deixis," in *The Handbook of Pragmatics*, L. R. Horn and G. Ward, Eds. Malden, MA & Oxford: Blackwell Publishing, 2006.

[31] T. Stivers and N. J. Enfield, Eds., *Person Reference in Interaction: Linguistic, Cultural and Social Perspectives*. Cambridge University Press, Cambridge, 2007.

[32] O. Ducrot, *Dire et ne pas dire*. Paris: Hermann, 1972.

[33] H. Nølke, "Ne... pas : négation descriptive ou polémique? Contraintes formelles sur son interprétation," *Langue française*, no. 94, pp. 48–67, 1992.

[34] I. Hutchby, *Confrontation Talk : arguments, asymmetries, and power on talk radio*. Mahwah, N.J.: Lawrence Erlbaum Associates, 1996.

[35] M. Burger, "Argumentative and hierarchical Dimension of a Debate Sequence : a Micro Analysis," *Studies in Communication Sciences*, special issue, pp. 249–264, 2005.

[36] J. Jacquin, "Orientation interactionnelle et multimodale vers l'activité de 'débattre' : analyse d'un extrait de débat public," *Mots*, no. 96, pp. 45–62, 2011.

[37] F. Rossano, "Gaze in conversation," in *The Handbook of Conversation Analysis*, J. Sidnell and T. Stivers, Eds. Chichester: Wiley-Blackwell, 2013, pp. 308–329.

[38] H. Sacks and E. A. Schegloff, "Home position," *Gesture*, vol. 2, no. 2, pp. 133–146, 2002.

[39] T. Stivers and F. Rossano, "Mobilizing Response," *Research on Language & Social Interaction*, vol. 43, no. 1, pp. 3–31, 2010.

[40] A. Kendon, *Gesture : Visible Action as Utterance*. Cambridge: Cambridge University Press, 2004.

[41] J. Hindmarsh and C. Heath, "Embodied reference: A study of deixis in workplace interaction," *Journal of Pragmatics*, vol. 32, no. 12, pp. 1855–1878, Nov. 2000.

[42] C. Goodwin, "Pointing as Situated Practice," in *Pointing: Where Language, Culture and Cognition Meet*, Mahwah, NJ: Lawrence Erlbaum Associates, 2003, pp. 217–241.

[43] N. J. Enfield, S. Kita, and J. P. de Ruiter, "Primary and secondary pragmatic functions of pointing gestures," *Journal of Pragmatics*, vol.39, no.10, pp. 1722–1741, 2007.

[44] J. Holmes, M. Marra, and B. Vine, *Leadership, Discourse and Ethnicity*. Oxford: Oxford University Press, 2011.

[45] J. Holmes, *Women, Men and Politeness*. London: Longman, 2013 [1995].

[46] J. Holmes and M. Marra, "Leadership and Managing Conflict in Meetings," *Pragmatics*, vol. 14, pp. 439–462, 2004.

[47] M. Marra, "Disagreeing without being disagreeable: Negotiating workplace communities as an outsider," *Journal of Pragmatics*, vol.44, no.12, pp. 1580–1590, 2012.

[48] S. A. Stadler, "Multimodal (im)politeness: the verbal, prosodic and non-verbal realization of disagreement in German and New Zealand English," University of Auckland & Universität Hamburg, 2006.

[49] S. Bruxelles and C. Kerbrat-Orecchioni, "Coalitions in polylogues," *Journal of Pragmatics.*, vol. 36, no. 1, pp. 75–113, 2004.

[50] L. Mondada, "Multimodal resources for turn-taking: pointing and the emergence of possible next speakers," *Discourse Studies*, vol. 9, no. 2, pp. 194–225, 2007.

[51] J. Jacquin, "S'opposer à autrui en situation de co-présence : la multimodalité de la désignation contre-argumentative," *Semen*, no. 39, pp. 19–38, 2015.

[52] M. Doury, "Preaching to the Converted. Why Argue When Everyone Agrees?" *Argumentation*, vol. 26, no. 1, pp. 99–114, 2015.

Gesture annotation scheme development and application for entrainment analysis in task-oriented dialogues in diverse cultures

Maciej Karpiński[1,2], *Ewa Jarmołowicz-Nowikow*[1,2], *Agnieszka Czoska*[1]

[1] Institute of Linguistics, Adam Mickiewicz University, [2] Center for Speech and Language Processing, Adam Mickiewicz University

maciej.karpinski@amu.edu.pl, ewa@jarmolowicz.art.pl, agaczoska@gmail.com

Abstract

In the present text, we propose a gesture annotation scheme and procedure designed for the analysis of gestural and gestural-prosodic entrainment in multilingual and multicultural corpora. The annotation scheme is based on the traditional gesture phrase structure but focused on more detailed characteristics of the stroke. In addition, it contains head movement, body position and gaze direction tags. The scheme is implemented as a hierarchical, multi-tier template for ELAN, with pop-up lists of available tags and predefined inter-tier dependencies. It can be used with any other annotation software that offers time-aligned multi-tier interval-based annotation. We also propose how annotation data based on our scheme can be applied to design and calculate measures of entrainment.

Index Terms: gesture annotation, multicultural corpus, dialogue, entrainment

1. Gesture: Cross-cultural differences and entrainment in dialogue

Even though there is still much need for systematic comparative studies on intercultural aspects of communication, cross-cultural differences in the realisation and understanding of gestures are widely acknowledged [1]. Kita [2] identifies four main factors that diversify the use of gestures across cultures: culture-specific convention between the form of gesture and its meaning, spatial cognition, linguistic differences and gestural pragmatics.

Emblematic gestures are probably the most spectacular examples of gestural cultural differences as sometimes only a specific form of a gesture may enable its interpretation or even recognition. Most of the studies that examine emblematic gestures are focused on cultural variation of their form, however not much attention is devoted to their origin and cultural determinants influencing their structure [e.g., 3]. Much more complex and less obvious in interpretations are intercultural studies that explore how different languages influence the use of gestures that are tightly linked to the co-occurring speech. Efforts are made to determine whether various structures of languages and their lexical resources are reflected in the use of gesture. Comparative cross-linguistic studies were carried out among speakers of Japanese, Turkish, English, Spanish and Chinese [4, 5, 6] in order to find how certain aspects of motion events, like manner or path, are expressed and described. Research questions touch upon also some cultural determinants of gesture use. Pointing gestures have been probably the ones most often described and compared across different languages and cultures in terms of their form, function and politeness [7, 8, 9, and many others]. It has also been noticed that gestures performed by speakers coming from different cultures varied with regard to the place of their realisation [10, 11, 12]. Unfortunately, most of those studies were not designed as comparative ones but rather served to explore the gestures within a given culture.

Cross-cultural comparative analysis of gestures should be based on the culture-independent system of annotation. The fundamental question is whether it is possible to design such a scheme. PAGE GAS (Gesture Annotation Scheme) is intended to describe the physical aspects of the movement, gesture morphology, posture of the body, head movements and gaze. Limiting the description to the physical aspects of the body motion seems to be a relatively safe solution as it allows for abstracting from culture-grounded categorisations and interpretations of gestures. Potentially culture-dependent interpretation is avoided and the data are still adequate for tracking and analysing many forms of gestural co-ordination. In further steps, it may also allow to explore higher levels of communicative alignment.

For many cultures there is hardly any scientific description of their gesture systems and even if such is available, its application may put extremely high requirements on the skills and training of non-native annotators. It is possible that sometimes they would find themselves in the position similar to those who try to transcribe foreign speech but are not certain whether the variation in what they can hear is phonologically relevant. For example, lip pointing typical of certain cultures [13], may not be even noticed by annotators who use only their hands for this purpose.

Gestures have been shown to take part in the processes of inter-speaker entrainment [14]. In the recent decade, they have been extensively explored as a factor in this process, mostly within the paradigm proposed or inspired by Pickering and Garrod [15]. Entrainment and related phenomena have been shown to help to predict success or failure of interpersonal interaction [e.g., 16; 17]. On the other hand, as they range from physical to the level of mental representation, they may also significantly influence complex cognitive processes [e.g., speech perception; 8]. Due to their complexity, the phenomena in question still remain largely understudied. Little is known on how they are influenced by culture- and language-related factors.

In the PAGE project (Prosodic and Gestural Entrainment in Conversational Interaction Across Diverse Languages), techniques and methods for entrainment measurement in dialogues in various cultures and in various languages. Task-oriented “tangram” dialogues (see Figure 1) were designed and recorded in various parts of the world (including Europe, North America and Oceania) by the project team in order to gather culturally diverse material. In the initial stage, project involved the design and implementation of annotation schemes for both verbal and non-verbal components of dialogue interactions. In the present paper, a gesture annotation scheme designed and used in the PAGE project is described along with some issues that arose during its preparation and implementation. Moreover, data analysis methods for tracking intercultural differences and inter-speaker entrainment are proposed.

2. PAGE multimodal corpus: Rich and culturally diverse data

The selection of features and forms of movement to be annotated is normally determined by the aim of the study but it may be also influenced by the culture. This means that some aspects of gestures that are essential from communicative point of view in some cultures may simply be neglected by the authors of the annotation scheme as these features of the gestures play no communicative role in their own culture or any culture they have explored. We tried to include in the PAGE GAS all the movements of the body that may potentially be important for entrainment in the dialogue situations we planned to explore. Therefore, for example, the movements of the lips and eyebrows that are regarded as essential for Papuan communication [13], however not playing an important function in any European culture, were added as optional subtier to the HeadMovement tier. Another example is the fact that in the PAGE GAS we also include handedness, even though we are aware that lateral (left-right) axis is not conceptually contrastive for all cultures [2].

It is possible to ensure a high degree of scheme flexibility in order to open it for new forms of gestures or to use relatively wide categories of motion characteristics. Still, the system may require continuous upgrades and corrections. This may be considered as a natural and desired process unless it leads to necessary corrections of earlier annotations. However, it is extremely difficult to train “intercultural annotators”. Cross-cultural differences in spatial cognition [2], gestural patterns typical for a culture may strongly influence the perception of certain features of the movement.

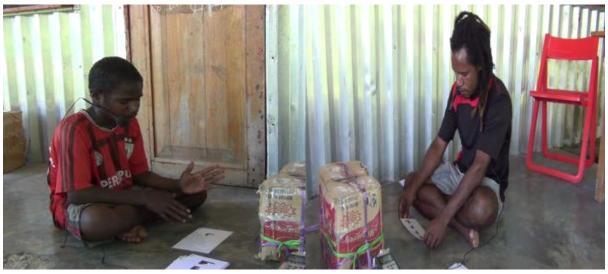


Figure 1: Speakers of Yali taking part in the “tangram” dialogue task (courtesy of Sonja Riesberg and Nikolaus Himmelmann)

3. Annotation scheme

3.1. Factors in the design of annotation scheme

Linguists dealing with spoken language often stress the fact that transcription always involves interpretation [18]. It is not different in the case of gesture annotation. Adopting any of numerous views to gesture structure and dynamics results in different coding and different symbolic representations. Popular gesture annotation schemes range from operating on a relatively limited number of tiers and categories [CoGest: 19, Mumin: 20, DiaGest: 21; 22] to ones with fine-grained categorisation of features and numerous, hierarchical layers [23]. They differ in the way they refer to gesture types or categories (as annotation most often involves categorisation) as well as to their peculiar features. Most of them, even if intended for more universal applications, grew out of specific projects and their requirements. They result from mediation between conceptual work and technical or practical limitations. According to Kendon [24] various gesture classifications that have been proposed can not be established as a universal and useful for all researchers as so many different dimensions of comparison are possible to distinguish for certain purposes. Most of the gestures annotation schemes, even if intended for more universal applications, grew out of specific projects and their requirements. As a consequence, the number of gesture annotation schemes is already large but new ones still are required and proposed for specific purposes or because of new, emerging theoretical perspectives.

Some of the most common questions involved in the challenge of gesture annotation scheme development are:

- Which body parts should be described for their movements?
- Should any kind of movement be annotated or rather some pre-selected categories of movement (e.g., gestures)?
- If gestures are annotated then how to reflect their morphology? Should they be divided into any smaller units (e.g., gesture phases) or be parts of some greater entities (e.g., gesture units)?
- How detailed (fine-grained) categories – if any – of gestures should be used?
- How much of the physical characteristics of gestures should be preserved in annotation or how much of the semantic or pragmatic meaning should be identified and ascribed by annotators?

Most of these and other issues go far beyond technicalities and may be related to very fundamental questions of multimodal communication studies, e.g., what a gesture is or what categories of gestures one can or should distinguish. Another important factors are economy and efficiency. Extremely detailed annotation can rarely be applied to large corpora. Moreover, with growing complexity, there are more demands on the skills of annotators and higher chances for discrepancies among them.

Among the goals of PAGE project was to propose annotation schemes that would be relatively easy to learn (since a few different teams had to be able to work with them) and simple to use (with a minimum number of categories and layers). However, such a scheme should still allow for coding and extracting the data crucial for entrainment analysis. Therefore it was decided to not only mark the number of gestures but also a selection of gesture features that seemed relevant. Since the “tangram” task may limit possible hand shapes (because participants may be prone to show various geometrical figures), other features (like representation technique or gesture space) were added. In order to control whether in a given dialogue similarity of gesture features between participants was related to entrainment or accidental, gaze annotation was added. Head movements were also annotated as important for feedback analysis.

3.2. PAGE Gesture Annotation Scheme (PAGE GAS)

There are only few studies showing which features of gestures are more or less relevant in the process of entrainment. In our scheme, we followed Bergmann’s and Kopp’s [14] as well as suggestions found in Kimbara [25] and Mol et al. [26] and our own findings from previous projects in order to define a preliminary set of features that may be more relevant for entrainment analysis. The inventory is certainly not exhaustive but it may be extended on the basis of exploration of our data. GAS is designed to include further categories or features without the necessity of reorganising the entire system.

Gesture annotation is hierarchical which is reflected by a hierarchical structure of tiers:

- *Gesture Unit* (obligatory);
- *Gesture Phrase* (obligatory);

- *Gesture Phase* (only marking the Stroke phase is obligatory but labels are available for all the phases of the basic gesture phrase model: Preparation, PreStrokeHold, Stroke, PostStrokeHold, Retraction).

Further tiers are bound to the Gesture Phase tier as subordinate and intended mostly as descriptors of the stroke. They are not directly tied to the time axis. Gesture Phase tier as well as all its subordinate tiers were provided with lists of possible tags available as pop-up menus. Many of these sets are based on or inspired by Bressem [22].

- *Hand shape*. While some schemes [e.g., HamNoSys: 27] tend to use the entire set of sign language handshapes as a reference, PAGE GAS is limited to four basic handshape categories: OpenPalm, Fist, OneFinger, ManyFingers.
- *Palm orientation*. Refers to palm relative orientation and offers the following labels: Inside, Outside, Up, Down, InsideUp, OutsideUp, InsideDown, OutsideDown.
- *Movement direction*: Left, Right, Up, Down, LeftUp, RightUp, LeftDown, RightDown, AwayFromSelf (in front), TowardsSelf (in front).
- *Trajectory*. Refers to the trajectory of the stroke using the following labels: Straight, Arch, Circle, Spiral, S-line, OtherShape.
- *Kinematics*. An optional tier, refers to the dynamics of gesture realisation: QuickNarrow (quick gesture, narrow range), QuickLarge (quick gesture, wide range), SlowNarrow (slow gesture, narrow range), SlowWide (slow gesture, wide range).
- *Representation technique*. This is a higher level category, not a simple descriptor of the physical properties of movement which may prove difficult to apply without in-depth cultural competence and knowledge on gesture usage. Still, it is offered as an optional component (wherever it is relevant) as it was pointed at by Bergmann and Kopp [14]. The choice of categories is based mostly on Kendon's approach [28] as well as Parrill's [29] and Lis' [30] recent works. The following labels are available: Acting, Depicting, Embodying, Indexing, Other.
- *Usage of hands in gesturing*. Handedness of each speaker should be learned from interview. This information will be partially duplicated by what is coded in the tiers confessed to the hands in order to simplify further analyses. The following labels are available: LeftHandUsed, RightHandUsed, BothHandsUsed Symmetrically, BothHandsUsedIdentically, BothHands UsedDifferently
- *Location in the Gesture Space*. Only the location of the peak of the stroke is taken into account. Although information on where each of gesture stages is performed might be useful for analysis, it would require much more resources and would introduce more discrepancies). Available values are: CenterCenter, CenterUp, CenterDown, PeripheryLeft, PeripheryRight, PeripheryUp, PeripheryDown, PeripheryLeftUp, PeripheryRightUp, PeripheryLeftDown, PeripheryLeft Down, ExtremePeripheryLeft, ExtremePeripheryRight, ExtremePeripheryUp, ExtremePeripheryDown, Extreme PeripheryLeftUp, ExtremePeripheryRightUp, Extreme PeripheryLeftDown, ExtremePeripheryLeftDown.

There are separate tiers for head and body movement annotation as well as gaze direction annotation.

- *Head movement*. Low-level, categorical head movement tags based on the work by Kousidis et al. [31] includes the following labels: Nod, Pitch, Jerk, Tilt, Shake, Yaw, Pro, Retro, Turn, Bobble, Slide, Shift, Waggle. This tier includes optional sub-tiers for lips and eye-brows shape annotation.
- *Body Posture***.** Optionally tagged as relative to the partner; even if there's no movement that can be labelled here. Metadata should contain some info on the initial position/posture of the participants – e.g., sitting, standing, directly facing each other. Available values are: LeaningTowardsPartner, LeaningAwayFrom ThePartner, TurningAwayFromThePartner, Bending/ Bowing (with no tendency to reach the partner).
- *Body Posture Kinematics*. Another optional tier, based on a tagset similar to Gesture Kinematics: Quick, Slow, QuickNarrow (quick gesture, narrow range), QuickLarge (quick gesture, wide range), SlowNarrow (slow gesture, narrow range), SlowWide (slow gesture, wide range).
- *Gaze* is (whenever possible) annotated manually using a set of two categories: Partner, OffPartner.

The scheme is implemented as an ELAN [32] template (see Figure 2) in which Gesture Unit, Gesture Phrase and Gesture Phase tiers are attached to the time axis but mutually bound by temporal relations (e.g., Gesture Phrase must be a part of Gesture Units, and a Gesture Phase must be a part of Gesture Phrase). Further tiers contain descriptors of the gesture phase (in our project, used solely for the stroke). They are attached to the annotations on the Gesture Phase tier using "symbolic association" in the definition of respective linguistic type in ELAN. Lists of available tags for these tiers are defined as "controlled vocabularies" and available as pop-up menus. For each hand, separate set of tiers is available. They are independent from the tiers for Body Posture and Kinematics, Head Movement and Gaze Direction.

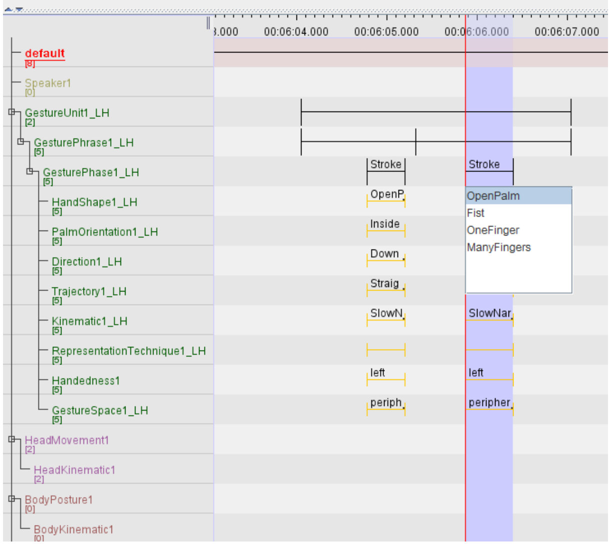


Figure 2: Tier structure for one hand and annotation gesture using PAGE GAS and ELAN template

3.3. Annotation and validation procedure

Annotation should start from finding Gesture Units, which is a relatively simple task even for inexperienced annotators. Only potentially communicative movements should be taken into consideration. Annotation does not include adaptative gestures, playing with items placed around the speaker. Then, Gesture Phrases should be found and tagged. Next, the Stroke phase should be found and tagged in each Gesture Phrase. For each Stroke, its features should be defined as tags on subordinate tiers. All aspects of the movements taken into consideration are described separately for both hands.

In certain circumstances it may be preferred that annotators do not hear speech signal from the recordings as it may influence and direct attention to co-occurring movements.

Reliability of coding was achieved using a double-check procedure. Annotations were checked for technical mistakes, scrutinised by a gesture expert, and returned to the annotators for corrections with detailed instructions. This procedure was repeated until acceptance by the expert. Independently, selected samples of recordings are currently annotated by different persons in order to calculate *kappa* tests. Special attention is paid to annotation training. This involves not only extending basic knowledge on gesture and its structure but also increasing annotators' sensitivity to various aspects of movement and its possible meaning.

4. Exploring annotations: Data analysis

In order to explore entrainment-related phenomena, we selected some preliminary methods of analysis. We assume that entrainment may find its expression in measures related to sequential and temporal arrangement of gestures as well as in the similarity of their forms as they are performed by the two speakers. We will also analyse cross-modal interaction between gestural and prosodic behaviour of the speakers (see section 5).

ELAN offers several options for data analysis (see e.g., section 4.2.). It provides basic annotation statistics and inter-annotator reliability measures but also limited *n*-gram analyses. For more advanced analyses, data-mining and statistical methods using external software tools will be applied. For preliminary data exploration, machine learning algorithms provided by Weka [33] will be used. Annotation Pro [34] will be employed for moving time-window and further overlap analyses. This software already provides a plug-in for time-window data extraction [34] and new plug-ins for overlap and simultaneity analysis are being created. Standard statistical software (e.g., IBM SPSS) will be also used to analyse data extracted from annotations as time series.

4.1. Data exploration with machine learning algorithms

Weka offers a number of machine learning algorithms from regression analysis or classification algorithms to artificial neural networks. They can be used for data exploration and preliminary testing of hypotheses.

Weka provides a choice of regression algorithms: simple, multiple and logistic regression. The methods may be used not only for testing theoretically-motivated hypotheses but also for searching for new ones. The algorithms implemented in Weka search for single independent variable crucial for predicting the values of dependent variables or a set of variables significantly influencing the dependent one.

As Weka offers a wide choice of classification and clustering algorithms (some of which are also available in IBM SPSS), it allows for testing differences between cultures, languages or dialects as well as dialogue settings or stages of interaction. A great virtue of using decision trees or Bayes algorithms for data exploration is the possibility of using both numerical and categorical (as most of our tags) data in one analysis as well as comparing different methods. For example, the possibility of comparing the outcomes of supervised and unsupervised methods is also a way of generating new hypotheses for further testing.

4.2. *N*-grams

Here *n*-grams are understood as sets of n subsequent tags. Exploration of their occurrence and frequency may help to reveal typical sequences of actions in communicative behaviour, for example more or less frequently used sequences of gestures. ELAN offers a simple tool for multiple file *n*-gram analysis that extracts *n*-grams from a set of files. However, in our analyses we go a step further and propose to analyse cross-tier n-grams, i.e. *n*-grams in which sequences may contain tags from various tiers – from one of the speakers of from both of them. This can be technically achieved using advanced search options or data export options in ELAN. The former approach is applicable mostly when one intends to look for certain hypothetical *n*-grams that are to be expected in the analysed material. The latter method requires an additional software tool that is able to look for *n*-grams in temporally ordered lists of tags [e.g., AntConc: 35]. Cross-tier and cross-speaker *n*-grams may provide much information on typical dialogue sequences in communicative interaction.

4.3. Time series analysis

Weka offers algorithms for time series analysis and forecasting. They can be applied not only to the raw data but to more direct measures of alignment like the ratios of events (e.g., the proportion of the number gestures or syllables by each speaker) or measures of similarity at a given stage of interaction (e.g., a number of same handshapes). Moreover, a powerful sequential analysis method can be applied to our data in THEME as an option of exporting data from ELAN into THEME has recently occurred.

All the above methods may be used to interpret variability of certain parameters on the time axis, especially to trace the changes in the strength or structure of alignment in time.

4.4. Time windows and overlaps

Moving time window approach is based on analysing average values of parameters or occurrences of events within a fixed-size time window that is moving along the time axis by a fixed step [e.g., 36, 37]. Subsequent time windows are usually overlapping by 25% to 75%. Various window sizes and overlap proportions may help to show various types of phenomena in the material under study. While ELAN itself does not provide the option of automatic time-window based analyses, we use ELAN export options and Annotation Pro software plug-ins [38] for this purpose.

Annotation Pro plug-ins for overlap analysis extract numerical or categorical data from all or a selected subset of events on a given annotation tier. Both types of plug-ins are able not only to analyse simultaneity of communicational behaviour but also provide an easy way of extracting data from a given stage of interaction. Therefore, it is possible to find whether, for example, entrainment or co-ordination measured as the proportion between the number of events in the two speakers, differs between the initial and final stage of interaction or its selected sub-stages. Using joint data from various tiers may allow for further analyses, reaching the alignment on the mental representation level.

Each dialogue was also divided into stages during which participants discussed one figure, from the beginning till the agreement over the task question. Such stages may be also regarded as time windows. A preliminary data analysis based on a subset of annotated movies shows, for example, that the number of one speaker's gestures in the n^{th} stage is a better predictor of the number of the second speaker's gestures in the $(n+1)^{\text{th}}$ stage (correlation coefficient = 0.496; see Figure 3) than in the same stage (correlation coefficient = 0.299).

Time window analysis will be also used to measure cross-modal correlations both within one speaker and between speakers (entrainment). Thus, it will be possible to test whether for example one speaker's speech is the source of entrainment or her gestures or both using multiple regression and other data-mining methods [39].

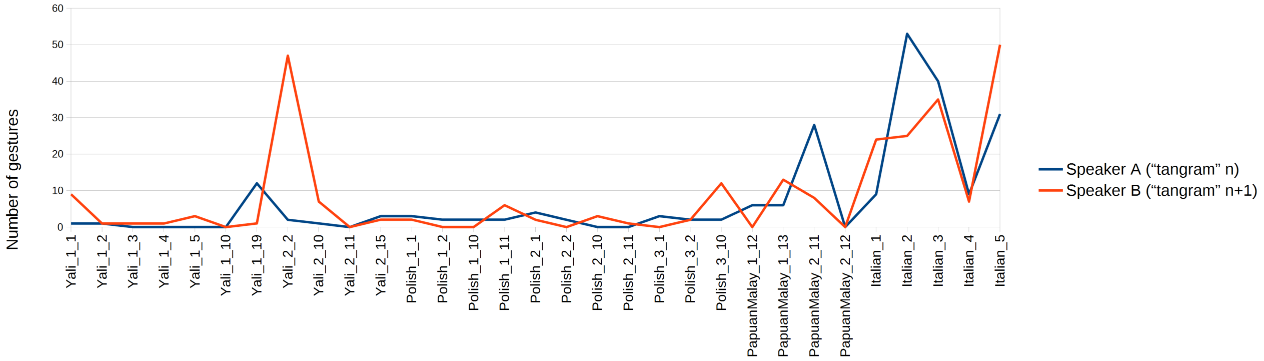


Figure 3: Correlation between the number of Speaker A gestures in the n^{th} stage (tangram) and the number of Speaker B gestures in $(n+1)^{th}$ stage.

4.5. Other options for statistical analyses

Multiple regression analysis allows to deal with multiple variables (here – the data on speech, gestures, gaze and posture shifts) in measuring co-ordination between speakers' behaviour. Multiple regression analysis provides information on temporal behaviour co-ordination between speakers and indicates whether it is intra-modal (e.g., only or mainly speech rhythm is correlated [38]) or cross-modal (e.g., one speaker's speech rhythm highly correlates with another speaker's gesturing rhythm).

In order to find whether the variables crucial for alignment are the ones related to culture specificity, interaction stage, task characteristics or speakers' familiarity, mediation and moderation analysis will also be performed. Eventually structural equation modelling may also be applied to the data.

Intercultural differences in gestures will also be analysed. Preliminary analysis (prior to annotation) shows that there are differences in the usage of representational and non-representational gestures between speakers of Papuan Malay and all the other languages annotated so far (see Figure 4).

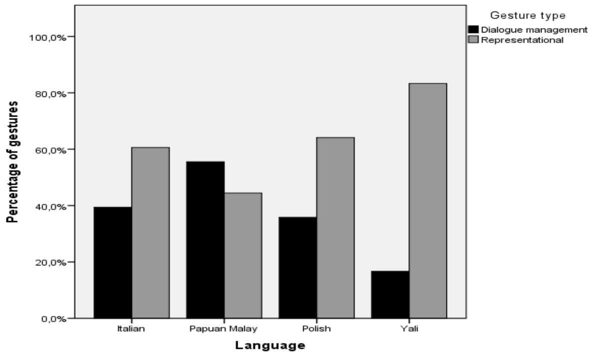


Figure 4: Intercultural differences in the proportion of representational (gray) and dialogue-management gestures (black) (*Chi-square*=16.2; *p*=0.001).

5. Conclusion

One can observe a growing interest in cultural determinants of gesture usage. However most of available studies describe the context and specificity of a single culture while only contrastive cross-cultural studies allow to distinguish these aspects of gestural entrainment that are in fact culturally specific and these which are universal. While there are a few cross-cultural studies focusing on gestures, none of them touches upon gestural entrainment.

In manual gesture annotation it is not possible to completely abstract from culture- and language-related factors. This refers both to the design of annotation scheme and the limitations on what and how can be annotated, and may result in omitting phenomena that are not perceived as gestures by non-native annotators. Although it seems that automatic motion capture and machine gesture categorisation would bring a significant dose of "objectivity", available systems are still difficult to set-up and apply, while their abilities to categorise gestures according to traditional systems are still very limited. In such circumstances, we must use some provisional solutions that let us further explore our data and help to proceed with better models of phenomena under study. We propose to annotate a limited set of physical properties of gestures. In order to avoid problems related to detailed segmentation of gestural phrases, we suggest to focus on the stroke. Further, our annotators are instructed by a specialist in gesture studies and intercultural communication, and the training is aimed at increasing their sensitivity to gesture-related phenomena as well as their potential shape in various cultural contexts. In order to test whether an annotation scheme is culture independent it would be in fact necessary to compare how annotators from different cultures describe the same set of video material after instructions obtained from the same source.

PAGE project involves both prosody and gestures analysis. These two phenomena are similar in many respects, including their functions in communication [e.g., 40]. They tend to harmoniously co-exist in utterances, coming from and synchronised by a hypothetical common source [41]. Therefore, drawing from works on prosodic entrainment (e.g., [42, 43]) may prove profitable in gesture analysis both in terms of methods and results. In further steps we intend to apply moving-time window and overlap analysis methods to measure correlations between gesture and prosody.

In spite of its obvious limitations, PAGE GAS is a flexible platform that can provide rich annotation data relevant for entrainment analysis. Its hierarchical and categorical design as well as the ELAN-dedicated template allow for easy data exchange and extraction. This, in turn, makes possible the application of a number of available data analysis methods that provide new insights into the structure and dynamics of entrainment in dialogue.

(ELAN template is available from the web page of the project:
http://www.page.home.amu.edu.pl)

6. Acknowledgement

This research was funded by **VW-Stiftung** as a part of the project Prosodic and Gestural Entrainment in Conversational Interaction across Diverse Languages.

7. References

[1] Kendon, A. "Some topics in Gesture Studies.", in Anna Esposito, Maja Bratanic, Eric Keller, Maria Marinaro, [Ed], The fundamentals of verbal and non verbal communication and the

biometrical issues. pp. 3-19, Amsterdam: IOS Press BV for NATO Series Publishing, 2007.

[2] Kita, S., "Cross-cultural variation of speech-accompanying gesture: A review." Language and Cognitive Processes, 24(2), 145-167, 2009.

[3] Matsumoto, D. and Hwang, H. S., "Cultural similarities and differences in emblematic gestures." Journal of Nonverbal Behavior, 37(1), pp. 1-27, 2013.

[4] Duncan, S., "Co-expressivity of speech and gesture: Manner of motion in Spanish, English, and Chinese.", in Proceedings of the 27th Berkeley Linguistic Society Annual Meeting, pp. 353-370. Berkeley, CA: Berkeley University Press, 2006.

[5] Kita, S. and Özyürek, A., "What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking." Journal of Memory and Language, 48, pp. 16-32, 2003.

[6] Özyürek, A., Kita, S., Allen, S., Brown, A., Furman, R., and Ishizuka, T., "Development of cross-linguistic variation in speech and gesture: Motion events in English and Turkish." Developmental Psychology, 44(4), pp. 1040-1054, 2008.

[7] Mueller, C., "Zur Unhoflichkeit von Zeigegesten" in: Osnabrucker Beitrage zur Sprachtheorie 52, pp. 196-222, 1996.

[8] Kita, S., [Ed], "Pointing: where language, culture, and cognition meet.". Mahwah, NJ: Lawrence Erlbaum. 2003.

[9] Jarmołowicz-Nowikow, E., "Pointing by hand; Types of reference and their influence on gestural form." in: Müller, C., Cienki, A., Fricke, E., Ladewig, S. H., McNeill, D. and Bressem, J., [Ed], Body – Language – Communication: An international Handbook on Multimodality in Human interaction. (Handbooks of Linguistics and Communication Science 38.2), pp. 1824-1833. Berlin/Boston: De Gruyter Mouton, 2014.

[10] Efron, D. Gesture, race and culture. The Hague. Mouton, 1972.

[11] Kendon, A. Contrasts in gesticulation: A Neapolitan and a British speaker compared. In C. Mueller and R. Posner, eds. The Semantics and Pragmatics of Everyday Gesture, pp. 173-193, Berlin: Weidler Buchverlag, 2004.

[12] Müller, Cornelia. Gesture-space and culture. In: Christian Cavé, Isabelle Guaitella and Serge Santi [Eds.], Oralité et Gestualité: Interactions et comportements multimodaux dans la communication, pp. 565–571, Montréal/ Paris: L'Harmattan, 2011.

[13] Wilkins, D., "Why pointing with the index finger is not a universal (in sociocultural and semiotic terms)." in Kita, S. [Ed.], Pointing: Where Language Culture and Cognition Meet., pp. 171–215. Hillsdale, N.J.: Lawrence Erlbaumaum, 2008.

[14] Bergmann, K. and Kopp, S., "Gestural Alignment in Natural Dialogue.", in Peebles, D. Miyake N. & Cooper R.P. [Ed.], Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci 2012), Austin, TX: Cognitive Science Society, pp. 1326–1331, 2012.

[15] Pickering, M. J., and Garrod, S., "Toward a mechanistic psychology of dialogue.", Behavioral and Brain Sciences, 27, pp. 169–225, 2004.

[16] Reitter, D. and Moore, J. D., "Predicting success in dialogue," Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 808–815, 2007.

[17] Ramseyer, F. and Tschaecher, W., "Nonverbal synchrony or random coincidence? How to tell the difference.", in Esposito A. [Ed], COST 2102 international Training School, pp. 182–196, Heidelberg: Springer Verlag, 2009.

[18] Hesselwood, B., "Phonetic Transcription in Theory and Practise." Edinburgh: EUP, 2013.

[19] Gut, U., Looks, K, Thies, A., Trippel, T. and Gibbon, D., "CoGesT – Conversational Gesture Transcription System- Version 1.0, Bielefeld http://coral.lili.unibielefeld.de/modelex/publication/techdoc/cogest/CoGesT_1.pdf, 2002.

[20] Allwood, J., Cerrato, L., Dybkær, L., Jokinen, K., Navarretta, C. and Paggio, P., "The MUMIN multimodal coding scheme." Technical report available at www.cst.dk/mumin/stockholmws.html. 2004.

[21] Karpiński, M., Jarmołowicz-Nowikow, E., Malisz, Z., Szczyszek, M. & Juszczyk, K., „Recording and transcription of speech and gesture in the narration of Polish adults and children" Investigationes Linguisticae , vol. 16, pp. 83-98, 2008.

[22] Bressem, J., "Transcription systems for gestures, speech, prosody, postures, gaze", in: Müller, C., Cienki, A., Fricke, E., Ladewig, S. H., McNeill, D. and Teßendorf, S., [Eds.], Body-Language-Communication: An international Handbook on Multimodality in Human interaction., Berlin, Boston: De Gruyter: Mouton, 2013.

[23] Lausberg, H., "NEUROGES - A coding system for the empirical analysis of hand movement behaviour as a reflection of cognitive, emotional, and interactive processes." in: Müller, C., Cienki, A., Fricke, E., Ladewig, S. H., McNeill, D. and Tessendorf, S. [Eds.], Body – Language – Communication: An international Handbook on Multimodality in Human interaction. Volume 1. Handbooks of Linguistics and Comunication Science, pp, 1022 – 1037 Berlin, New York: De Gruyter Mouton, 2013.

[24] Kendon. A. "Gesture: Visible action as utterance." New York: Cambridge University Press, 2005.

[25] Kimbara, I., "Gesture form convergence in joint description." Journal of Nonverbal Behavior, 32, pp. 123-131, 2008.

[26] Mol, L., Krahmer, E., Maes, A. and Swerts, M., "Adaptation in gesture: Converging hands or converging minds?" Journal of Memory and Language, 66, pp. 249-264, 2012.

[27] Prillwitz, S., Leven, R., Zienert, H., Hanke, T. and Henning, J. , "Hamburg Notation System for Sign Languages–An introductory Guide.", International Studies on Sign Language and the Communication of the Deaf, 5, 1989.

[28] Kendon, A., "Gesture – Visible Action as Utterance.", Cambridge University Press. 2004.

[29] Parrill, F., "Viewpoint in speech-gesture integration: Linguistic structure, discourse structure, and event structure.", Language and Cognitive Processes, 25(5), 650-668, 2010.

[30] Lis, M., "Multimodal representation of entities: A corpus-based investigation of co-speech hand gesture.", Copenhagen: Faculty of Humanities, University of Copenhagen, 2014.

[31] Kousidis, S., Malisz, Z., Wagner, P. and Schlangen, D., "Exploring annotation of head gesture forms in spontaneous human interaction." Proceedings of the Tilburg Gesture Meeting (TiGeR 2013), Tilburg, The Netherlands. 2013.

[32] Sloetjes & Wittenburg 2008. Annotation by category: ELAN and ISO-DCR. Proceedings of LREC 2008.

[33] Hall, M. Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H., "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, Issue 1, 2009.

[34] Klessa, K., Karpiński, M. and Wagner, A., „Annotation Pro – A new tool for annotation of linguistic and paralinguistic features.", Proceedings of the Tools and Resources for the Analysis of Speech Prosody Workshop (TRASP), Aix-en-Provence, 2013.

[35] Anthony, L., AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/, 2014.

[36] Kousidis, S., Dorran, D., Wang, Y., B. Vaughan, Cullen, C., Campbell, D., McDonnell, C. and Coyle E. , "Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues", Proceedings of Interspeech, pp. 1692-1695, 2008.

[37] De Looze, C., Scherer, S., Vaughan, B., & Campbell, N., "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction." Speech Communication, 58, pp.11-34, 2014.

[38] Karpiński, M., Klessa, K. and Czoska, A., "Local and global convergence in the temporal domain in Polish task-oriented dialogue", Speech Prosody 7, Dublin, 2014.

[39] Czoska, A., Klessa, K., Karpiński, M., Jarmołowicz-Nowikow, E., "Prosody and gesture in dialogue: Cross-modal interactions", Gesture and Speech in Interaction 4, Nantes 2015.

[40] Gibbon, D., "Modelling Gesture as Speech.", Poznań Studies in Contemporary Linguistics, 47(3), pp. 447-, 2011.

[41] McNeill, D. and Duncan, S., "Growth points in thinking for speaking.", in McNeill, D., [Ed], Language and gesture., pp. 141-161, Cambridge: CUP, 2000.

[42] Truong, K. P. and Heylen, D., "Measuring prosodic alignment in cooperative task-based conversations", Proceedings of Interspeech 2012, 9-13 September 2012, Portland, OR, USA, 2012.

[43] Wagner, P., Malisz, Z., Inden, B. and Wachsmuth, I., "Interaction phonology – a temporal co-ordination component enabling representational alignment within a model of communication", in: Wachsmuth, I., de Ruiter, J., Jaecks, P. and Kopp, S. [Eds.], Alignment in Communication: Towards a New Theory of Communication. Amsterdam: Benjamins, 2013.

Not so dependent after all: functional perception of speakers' gestures with and without speech

Kasper Kok [1], *Kirsten Bergmann*[2], *Stefan Kopp* [2]

[1] Faculty of Humanities, VU University, Amsterdam, the Netherlands
[2] CITEC, Faculty of Technology, Bielefeld University, Bielefeld, Germany
k.i.kok@vu.nl, kirsten.bergmann@uni-bielefeld.de, skopp@techfak.uni-bielefeld.de

Abstract

Speakers' spontaneous gestures are traditionally thought of as a *dependent* system, their meaning relying heavily on what is expressed verbally. Nonetheless, studies using muted videos as stimuli consistently report that gestures have some degree of communicative import in the absence of speech. Here, we argue that the dependence-autonomy question can be advanced by adopting a functional (linguistic) perspective. We ask whether access to speech is necessary to understand what kind of semantic or pragmatic function a gesture performs. Based on a large-scale web-based perception study, we report that when the notion of meaning is operationalized on a functional level, the dependence of gesture on speech largely diminishes.

Index Terms: gesture, multimodality, functional linguistics

1. Introduction

Whereas gestures undoubtedly have communicative value [1, 2], it is commonly assumed that their meaning is underspecified and dependent on the accompanying speech. As noted by McNeill [3:41], the same object or situation can be referred to by gestures with different formal characteristics in different contexts. This context-dependence, according to McNeill, "contrasts with the stability of lexical forms, [which] present modifications of the core meaning in all contexts but do not give up their basic form in this process. If no such basic and context-independent forms exist in gestural expression, it can moreover be expected that the meaning of a gesture cannot be determined when the access to the accompanied speech is denied. This hypothesis has been tested in a variety of studies where participants' performance on comprehension tasks was compared in conditions with and without access to speech [4-7].

Feyereisen et al. [4] tested whether participants were able to qualify the gestures of speakers in naturalistic discourse as either iconic or batonic when video recordings were presented with or without sound. The authors reported that raters were substantially more accurate when the auditory channel was available, but performance was still above chance level when access to the speech was denied. In a second experiment, it was found that when video clips were presented without access to the audio channel, subjects were generally unable to guess the original utterance from a fixed list of options. However, a high degree of consistency was found in participants' responses to this task. According to the authors, this hints at the existence of intrinsically meaningful qualities of the gestures in their stimulus set.

A similar paradigm was used by Krauss et al. [5], who presented participants with audio-muted video clips and asked them to guess which of two words was the gesture's 'lexical affiliate' (a term used after Schegloff [8], referring to "the word or words deemed to correspond most closely to a gesture in meaning"). Raters' performance on this task was far from perfect, yet significantly above chance level. In a follow-up experiment, participants were instructed to assign semantic categories (action, location, object name, or description) to a set of gestures in conditions with and without sound. It was found that when the speech was available, the classifications very closely reflected the semantic content of the accompanied speech. In the absence of the verbal channel, the judgments were not random either: gestures were often assigned the same semantic category as their lexical affiliate, to which participants had no access. Given that the presence or absence of speech was consistently found to have some degree of predictive value, the authors conclude that "although gestures can convey some information, they are not richly informative, and the information they convey is largely redundant with speech" [5:743].

In accordance with this conclusion, Hadar and Pinchas-Zamir [6] argue that the meaning of gestures is best understood in terms of different levels of 'semantic specificity': some gestures have a specific, lexeme-like meaning, whereas others convey meaning in a "vague and tentative" fashion. Taking this notion as their point of departure, two experiments were carried out where participants had to select a word from a list that related most closely to a given gesture. Participants less often chose the word that had been coded as the lexical affiliate of the gesture when the speech was muted than when the speech or a transcription of it was present. Among the 'erroneous' responses, moreover, visually or semantically related distractor words were chosen more often than unrelated distractors. Based on this graded effect, the authors claim that gestures are on a cline of degrees of semantic specificity.

A recent study by Kibrik & Molchanova [7] used a more contextualized task to investigate the relative dependence of speech, gesture and prosody. Participants watched segments of movies or videotaped conversations in conditions with audio only, video only, prosody only, or combinations of these. They then answered a set of multiple choice questions about the content of the movie clips (e.g. "What does Tamara offer Masha before the beginning of the conversation?"). In line with previous findings, it was found that although participants were more accurate when speech was available, a substantial subset of the questions was answered correctly in the absence of speech as well. On the basis of this result, the authors argue that whereas speech might be the 'leading channel' of information conveyance, gestures (and prosody) do carry some degree of independent semantic load.

All four of these papers report a rather ambivalent relation between speech and gesture: on the one hand, gestures alone are not as informative as speech-gesture combinations. On the other hand, gestures by themselves provide sufficient cues for participants to score well above chance level on various types of comprehension tasks. The authors roughly agree that this ambivalence reflects the fact that gestures are semantically 'underspecified': they carry some intrinsic meaning, but only on a rather schematic level (cf. [9, 10]). In accord with this claim, it has been demonstrated that the degree of semantic congruence between speech and gesture influences processing latency [11]. It remains a relatively open question, however, how the level of schematicity present in co-speech gestures is best characterized. What types of meaning are associated with gestural forms irrespective of speech, and on what level of abstraction?

Getting a grip on this question requires a comprehension task of a less specific character than those used in the research discussed above. Previous studies commonly assessed utterance comprehension using very concrete and detailed questions (e.g., 'what is the lexical affiliate of this gesture?' or 'what did X say to Y?'). Thus, the notion of meaning is often simply conceived as *reference* to some specific object or situation – a view that has been under heavy dispute in contemporary cognitive-functional linguistics. An alternative perspective is to characterize meaning in terms of the *functional contribution* of an expressive unit to the ongoing discourse. According to recent work, this view provides an appropriate level of abstraction for studying semantic and pragmatic interfaces between speech and gesture [12-14]. In terms of experimental design, this entails that instead of asking participants to identify what object or event a gesture refers to exactly, one needs to ask whether access to speech is needed to infer *what kind of semantic or pragmatic function* a given gesture performs.

In this paper, we examine the question of speech-gesture dependence, adopting and implementing a functionally inspired perspective. In particular, we examine four prominent functions of gestures: object reference, attribution of a static property (e.g. depicting the shape or size of an object), attribution of a dynamic property (e.g. depicting a process of movement), and meta-communicative signalling (e.g. indicating difficulty in speech production). The perception of these functions by naïve observers is investigated in an online perception study with the presence or absence of speech as the critical manipulation.

2. Methods

The research procedures are akin to those described in [15], supplemented with an audio-only condition. In a web-based study, participants watched video fragments of direction-giving discourse recorded in a relatively natural setting. They were asked to indicate their interpretation of the gestures in these videos by filling in a controlled survey. This survey consisted of a number of statements (e.g. *The hands of the speaker refer to an object or person*) to which participants assigned an agreement score on a 7-point Likert-scale. In a between-subject design, the same videos were presented with the sound on (Sound-On condition) and with the sound muted (Sound-Off condition). The substantial data set yielded by this design (449 gestures, 16 raters per gesture) allows for a detailed assessment of the degree to which the functional interpretation of speakers' gestures depends on observers' access to the co-expressed speech.

2.1. Stimuli

The video snippets that served as stimuli were fragments of the Bielefeld Speech and Gesture Alignment corpus (SaGA, [16]). This corpus consists of German-spoken dialogues where one participant gives directions to another participant after having taken a tour through a virtual town. The stimulus set used for this study contained 173 videos and a total of 449 individually marked gesture strokes. To promote ecological validity, *all* of the gestures of the route-giver were treated as potential stimuli. That is, we did not filter out any gestures based on a priori interest. The video fragments that were used as stimuli were taken from five different dialogues. The start and end points of these clips were moments where the hands of the speaker were in rest position, and/or the speech was paused. Thus, the stimuli comprised relatively isolated discourse units.

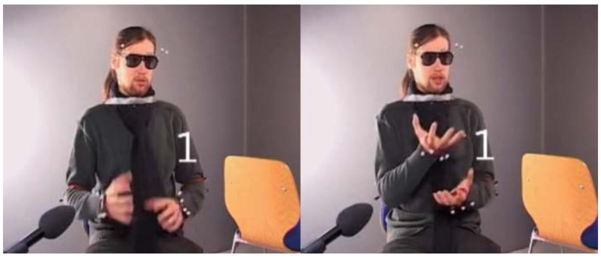


Figure 1: *Video stills of example stimulus. Numbers were edited into the video in concurrence with the gesture units.*

Most of the stimuli contained more than one gesture stroke (isolated, single-stroke gestures are rare in naturalistic conversation). Therefore, numbers were edited into the video clips in coincidence with the individual stroke phases, so that participants could be instructed to pay attention only to one specific stroke at a time (see Figure 1). In order to prevent the video fragments from being too long, gesture sequences with more than six consecutive strokes were discarded from the stimulus set. For all videos, two versions were created: one was simply a fragment of the original corpus, and one was a version of this fragment where the sound was muted.

2.2. Participants

Participants were recruited and reimbursed via the online research platform Crowdflower[1]. In order to assure reliability of the data, the following performance thresholds were implemented. Participants' responses were included only if they (1) had passed 80% or more of the test questions that served to assess their attention (see 2.3); (2) had taken more than seven minutes to complete the survey (mean completion time was 21.52 minutes); (3) had sufficient variance in their data (those participants who consistently gave only one or two unique answers to all questions were excluded). After having applied these procedures, a total of 347 participants remained, resulting in 16 judgers for all stimuli. The participants were aged 16 to 69 (M=37.3, σ=12.1) and 129 were female. All were present in Germany at the time of participation (according to their IP-address) and reported to have full proficiency of the German language.

2.3. Procedure

Participants were assigned either to the Sound-on or the Sound-off version of the experiment. Before the survey was presented, their availability over well-functioning video and audio equipment was verified by means of a digital 'captcha', where visually and auditorily presented words had to be typed in, in order to proceed. Participants who were able to do so were presented with the following instructions:

> In this survey you will be asked to answer questions on short video segments. The questions concern the relationship between the speaker's gestures and what he/she says. Every list of questions concerns *one single gesture*. When multiple gestures occur in a video segment, they will be marked with numbers.
> Please read the questions carefully and answer on the basis of your own perception. Be aware that for some videos none of the options is applicable, while for other videos, multiple options may apply. In the video segments, someone explains a route through a city to his/her interlocutor. You can play the video as often as you want.

Hereafter, a page was displayed with an embedded video on top and a list of statements below. These statements were inspired by the levels of semantic and pragmatic analysis recognized in functional linguistics (cf. [12] for details), but formulated in a simplified way to be accessible to naïve participants. In the current study, we focus on a subset of four possible functions of gestures. Three of these correspond roughly to three prevalent semantic functions of linguistic constituents: gestures' ability to refer to entities in the world (as might also be realized in speech by noun phrases like "the cup"); gestures' ability to attribute a static feature to some entity (as might be realized in speech by an adjective like "big"); and gestures' ability to attribute a dynamic property to some entity (as might be realized in speech by a verbal phrase like "is spinning"). The fourth question concerns what one might call a 'meta-communicative' function of gesture: the signalling of difficulty with word retrieval. Note that although these functions are examined as individual variables in the current paper, they may be intercorrelated; a single gesture may for instance perform referential as well as attributive functions [15]. Table 1 lists the questions/statements corresponding to the four functions of interest, as well as the labels that will henceforth be used for referring to them.

Table 1. *Survey questions*

Question label	English translation
Refer-to-Object	The hands of the speaker refer to a specific *object or person.*
Depict-Shape	The hands of the speaker depict the *size or shape* of an object or person.
Depict-Movement	The hands of the speaker depict the *movement* of an object or person.
Word-Search	The hands of the speaker show that he/she has *difficulty finding the right words.*

[1] http://www.crowdflower.com

Participants' task was simply to indicate whether they thought each of the statements applied to the gesture of interest (referred to via the number that appeared in the screen.) These judgments were given on a 7-point Likert scale, ranging from 'certainly does not apply' to 'certainly applies'.

The experiment was preceded by a practice trial, so that participants could get accustomed to the task. Subsequently, twenty video clips were presented, one at a time, which were randomly sampled from the total stimulus set. Thus, every participant was exposed with a different subset of video snippets. Because all analyses of interest are item-based, we allowed participants to take part multiple times, with a maximum of five. All cases where the same video had coincidentally been assigned to a participant twice were automatically filtered out from the data set. For each video stimulus, the order of the survey questions was randomized. Test questions that served as diagnostic for participants' attention were added to the list of survey question (in an unmarked, randomly determined location) for one out of every four questions. These asked to simply tick one of the boxes in the Likert-scale (e.g., the third one from the left).

3. Results

Two types of analysis were conducted to gain insights into the role of speech with respect to the functional interpretation of the gestures. To find out whether participants' perception of the gesture's functions varied with the presence or absence of speech *across the board*, we compared the frequency distributions of the mean ratings on all videos across conditions. Second, we conducted correlational analysis to investigate whether the video stimuli were assigned similar ratings when presented with and without sound. Results of these analyses are discussed in the light of qualitative inspection of the data.

3.1. The object reference function

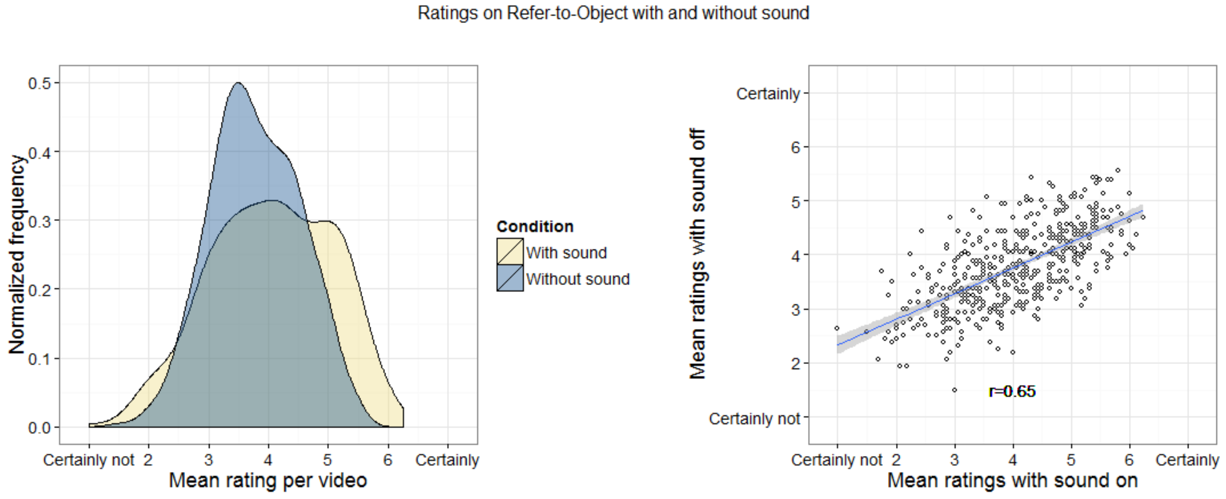
As apparent from Figure 2, scores on the Refer-to-Object question range throughout the entire spectrum of certainty ratings. The mean score over all stimuli is 4.06 (s=1.02) for the Sound-On condition and 3.78 (s=.75) for the Sound-Off condition. From a paired samples t-test, this difference appears significant ($t_{[448]}$ =18.0, p<.001, d=.31). However, note that the effect size is very small: the difference in means comprises less than a third of the standard deviation. Thus, although gestures are more often judged to refer to an object when perceived in the context of speech, their referential function is almost equally easily recognized without access to the verbal channel. This is further corroborated by the correlational analysis. A Pearson test reveals that the ratings in the Sound-On and Sound-Off conditions were strongly correlated ($r_{[447]}$=.65 , p<.001).

As apparent from Figure 2, there is a difference between the conditions in terms of the variance in the data. The frequency distribution of scores in the Sound-Off condition is much steeper than the one corresponding to the Sound-On condition (s with sound: 1.02; s without sound: 0.75). From this observation, it appears that the presence of speech does not directly contribute to the qualitative interpretation of the gesture, but rather *decreases uncertainty* among observers. The high rate of 'cohesive gestures' [17] in the stimulus set plausibly relates to this finding. Many video fragments contain gestures that refer to an object by referring to a region of space that has previously been associated with a certain landmark object. In the absence of the verbal channel, such gestures may have the appearance of non-representational movements of the hands (e.g. beat gestures; [4]) or, more generally, may not be easily recognized as carrying a referential function.

For some gestures in the stimulus set, the referential status is not clear-cut even in the presence of speech. With regard to trajectory-tracing gestures that co-occur with sentences like *over there you should go to the left*, there is generally low consensus whether or not the hand refers to an object or person. Though such gestures can be thought of as having a referential component, where the hand 'embodies' the interlocutor (cf. [18]), they were often judged to be merely attributive by nature, tracing the movement of some entity without actually referring to it. Thus, for many of our stimuli, responses to the Refer-to-Object question were not entirely homogeneous across participants, regardless of the presence or absence of speech.

3.2. The shape representation function

Mean ratings on the question Depict-Shape do not differ to a statistically significant degree across conditions ($t_{[448]}$=-1.8, p=.075, d=.052). In fact, correlational analysis suggests that gestures were attributed approximately equal ratings on this question when presented with and without sound ($r_{[447]}$=.84, p<.001). Figure 3 presents these data graphically. We see that in the Sound-On condition, the distribution of mean scores on the stimuli has two peaks: one close to the 'certainly not' pole, and one corresponding to more certain ratings. Thus, for a large portion of the gestures in the stimulus set, participants were rather certain whether or not the shape or size of some object was represented. In the Sound-Off condition, the histogram's shape more closely approaches a normal distribution. Here we see that when the sound was muted, the overall uncertainty rate with respect to the Depict-Movement question was substantially higher. A likely explanation is that the SaGA corpus, from which the stimuli were extracted, contains many tracing gestures, i.e. gestures where the hand traces a line through the air as if holding a drawing utensil. Such gestures can be semantically ambiguous: tracing handshapes can be used either to draw the shape or outline of some object, or to depict the movement of some object in space [18, 19]. In the context of a verbal utterance such as *you will see a curved river*, a horizontal tracing gestures will be given a different interpretation than in the context of the utterance *you have to turn left and then right like this* (static-



(a) (b)

Figure 2: *Ratings on the question Refer-to-Object with and without sound*

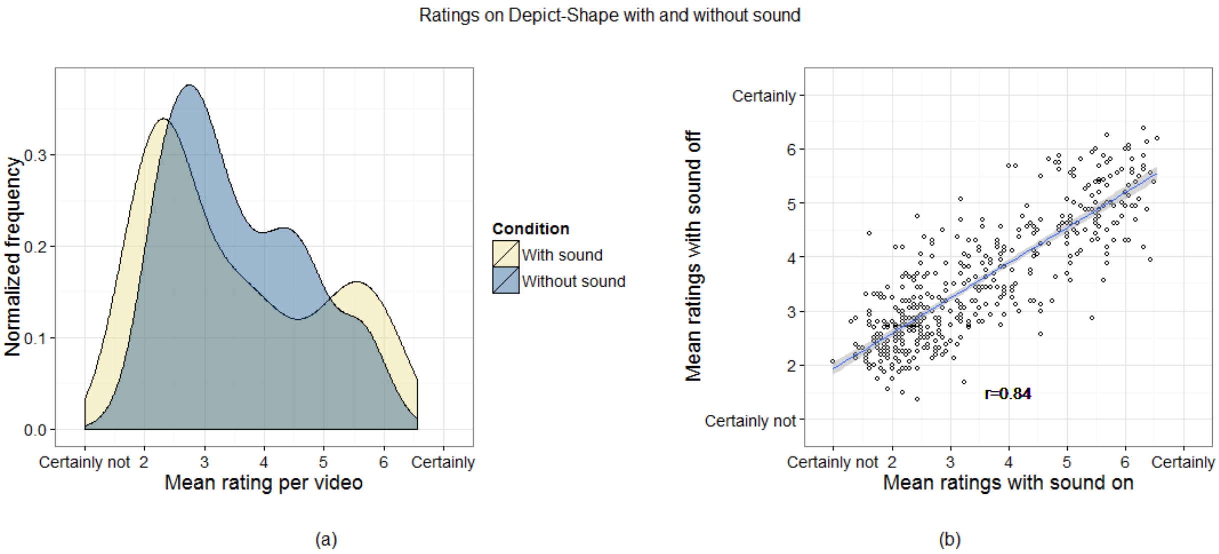


Figure 3: *Ratings on the question Depict-Shape with and without sound*

attributive in the first case, processual in the second). As reflected by the relatively high degree of gesture stimuli in the middle region of Figure 3a, many tracing gestures therefore remain underspecified in the Sound-Off condition.

3.3. The movement representation function

Scores on the Depict-Movement question diverge to some extent between conditions, but in a different direction than we have seen thus far: the mean rating across all stimuli is higher in the Sound-Off condition than in the Sound-On condition (see Figure 4). This difference is statistically significant ($t_{[448]}$=-15.9, p<.001,), but the effect size is small (d=.28). Ratings on the stimuli correlate significantly across conditions ($r_{[447]}$ =.55, p<.001).

According to these data, not many gestures in the corpus clearly depict the movement of an object or person. Strikingly, direction-tracing gestures such as those discussed earlier (e.g. those co-occurring with *go to the left*) are not consistently judged as depicting movement. Moreover, likely due to the ambiguity of tracing gestures discussed in the previous section, the Sound-Off condition yielded a high rate of uncertain responses. Many tracing gestures were judged to be potentially depicting some movement when presented in the absence of speech, but were clearly perceived as part of the act of reference to an object (i.e., non-processual in their semantics) when presented with the audio on. Thus, in some cases, the speech is needed to decide whether a tracing gesture has a referential or attributive function. Overall however, we see that the scores in the two conditions correlate strongly, suggesting that this dependence is more of an exception than a rule.

3.4. The meta-communicative function

The three functions discussed so far relate to the domain of semantics: they involve gestures' ability to refer to entities and situations and attribute properties to them. The fourth and final function taken into consideration here has a rather different nature. We here look at gesture's ability to signal that the speaker experiences difficulty in formulating his utterance, a type of action sometimes referred to as 'own communication management' [20]. This section inquires whether access to speech is necessary to understand that a gesture is meta-communicative by character. As clear from Figure 5, the pattern in the data we find here is not substantially different from what we have seen above. Mean scores are differently distributed in the two conditions ($t_{[448]=}$4.2, p<.001), with higher scores in the Sound-On conditions, but the effect size is marginal (d=.20). The scores in the Sound-On and Sound-Off

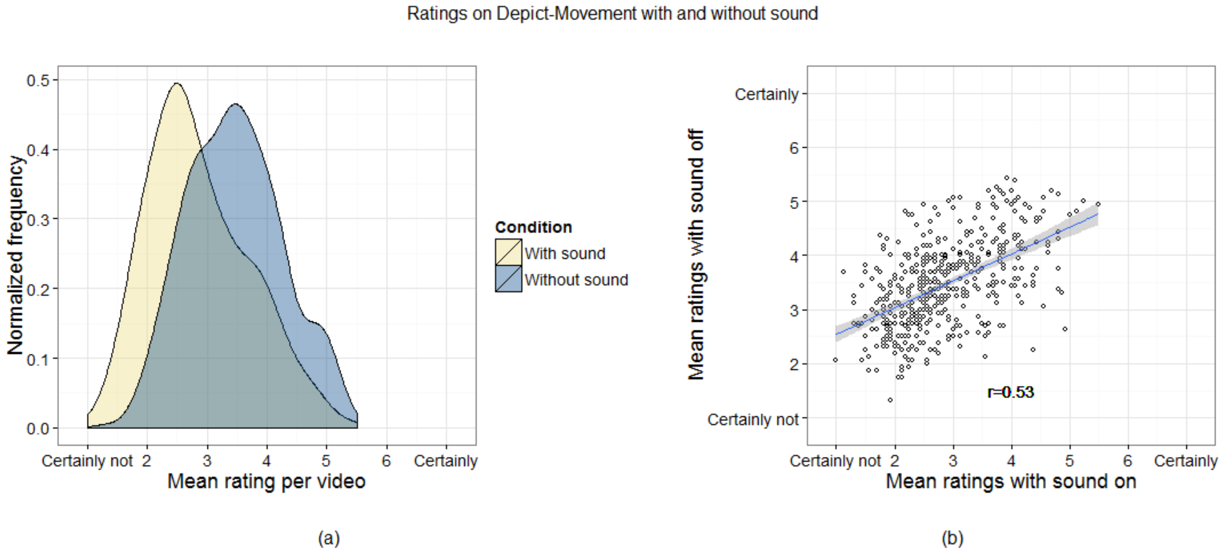


Figure 4: *Ratings on the question Depict-Movement with and without sound*

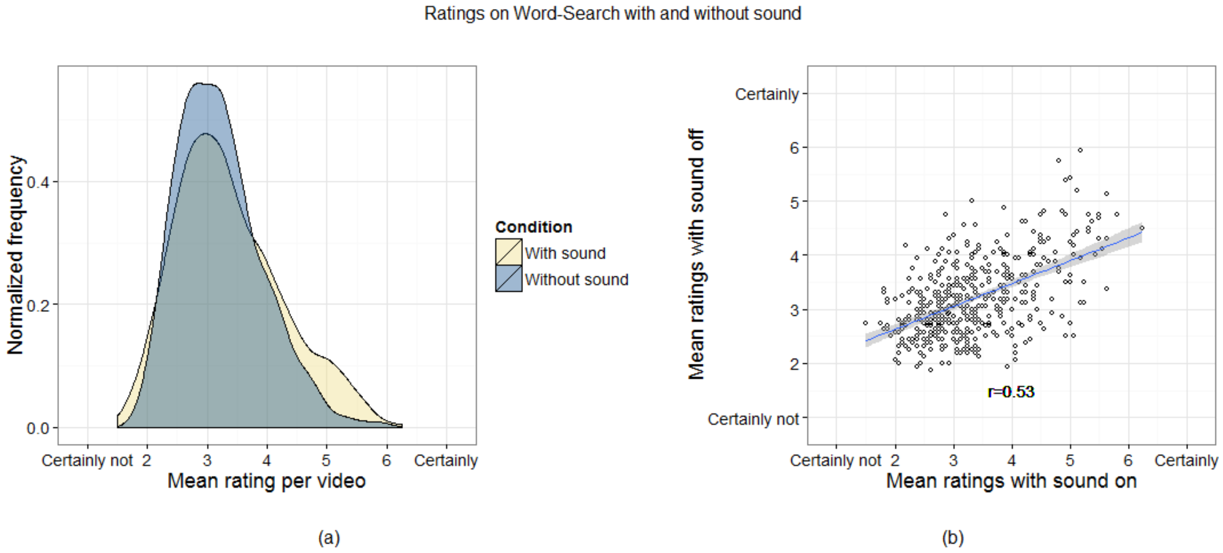


Figure 5: *Ratings on the question Word-Search with and without sound*

conditions are again strongly correlated ($r_{[447]}$=.53, p<.001).

These data are largely in line with the hypothesis that own communication management is associated with particular formational patterns (finger snapping or rapid cyclic movement of the hands, cf. [20]): the presence or absence of speech may play a role, but is not a crucial factor in deciding whether a given gesture signals own communication management. Nonetheless, the shape of a gesture alone is not always enough to recognize the gesture as having this function. Cyclic gestures, for instance, are not only linked to word search, but also to different contexts, such as encouragement of the interlocutor to continue speaking [21]. In accord with this underspecificity, we see that the right-most region of the distribution in Figure 5A, which represents the most certain/positive responses on the Word-Search question, is virtually empty in the Sound-Off condition (6 stimuli with mean scores above 5), but does have a substantial number of responses in the Sound-On condition (25 stimuli with mean scores above 5). This indicates that even relatively fixed patterns in gesticulation, such as the cyclic gesture and other conventional ways of meta-communicative signaling, are to some degree dependent on verbal context. Nonetheless, a substantial overlap exists in the response profiles across conditions. It appears, as we have seen before, that the judgments in the presence of speech are qualitatively similar, but more assured, than those in the video-only condition. In other words, the influence of the verbal channel on the perception of the gestures is largely limited to the reduction of uncertainty.

4. Discussion and conclusion

Despite ample attention to the question of speech-gesture dependence, it has to date remained relatively unclear how the meaningful qualities that are inherently present in speakers' gestures are best characterized. Here, we have argued that a functional (linguistic) view on gestural meaning could provide a fruitful level of abstraction to point out which aspects of gestural meaning can be understood without speech. In an internet-based perception study, we have compared people's perception of gestures with and without access to speech, with respect to four specific functions: object reference, static attribution, dynamic attribution and signaling of own communication management.

Our results suggest that access to the verbal channel does not have strong influence on the perceived functions. All effect sizes were small (between one third and one fifth of a standard deviation) and response profiles show a great deal of overlap across conditions for all functions investigated. Note that the reported p-values, of which three out of four were statistically significant, can be misleading. The substantial size of our stimulus set (449 stimuli) yields very high statistical power and an increased risk of false positives. Therefore, in line with current trends in statistics, our conclusions are primarily based on the reported effect sizes and correlations.

A second finding is that the variance in the responses between conditions diverged strongly – it was consistently found that participants were more hesitant and inconsistent in their responses on the video-only stimuli than on those with video and audio. Thus, rather than having a strong influence on the 'direction' of the functional interpretation of the gestures, the presence of speech appears to play a role in reducing uncertainty among raters.

These findings raise theoretical as well as methodological considerations. First, they suggest that the question to what degree gestures constitute an independent (sub)system of communication hinges strongly on the level of abstraction one adopts when characterizing their meaning. Using somewhat schematic, functionally oriented questions, we find that the interpretation of gestures is not as dependent on speech as previously supposed on the basis of experiments which ask very concrete questions (e.g., *what is the lexical affiliate of this gesture?*). This is an important insight, because it informs the level of description appropriate for understanding how speech and gesture intersect: rather than seeking to understand gesture as a system that functions akin to speech (with a full repertoire of very specific form-referent mappings), it is probably better to ask to what extent speech and gesture may have the same kind of *communicative role* in a given discourse setting, and what formational patterns these roles correlate with in either modality. Thus, the existence of some form of 'inherent meaning' of gestures (i.e., speech-independent form-meaning associations) cannot be dispensed with, but should be sought for on a higher level of abstraction than in the case of verbal meanings.

Apart from these theoretical considerations, the results presented here may have methodological implications for the longstanding debate whether annotation practices should be performed with or without audio. Whereas most traditional coding schemes require access to the verbal channel in order to be applicable [3, 22], some contemporary approaches pursue the view that the recognition of meaningful patterns in gestural expression entails at least one round of annotation with the audio turned off ([23, 24]). The high degree of overlap between the Sound-On and Sound-Off conditions in our study supports the latter approach: according to the results reported, it can be fruitful to annotate basic-level functional categories of gestures on the basis of video only. This could benefit the objectivity of the research, because gesture interpretation will be less biased by the co-occurring speech. When dealing with more detailed research questions, however, additional annotation remains necessary. Access to the verbal channel can help the annotator to disambiguate – or specify the meaning of – the gestures of interest. The potential presence of intrinsic high-level meanings in the gestures investigated in this paper, hence, does not suggest that speech and gesture are to be approached as independent systems. Rather, the divergence between our and earlier findings seems to point out that the further we go along the cline from schematic to lower-level layers of meaning, the more important it is to annotate, and think of, speech and gesture in the context of one another.

5. Acknowledgements

The first author is grateful to support from the Netherlands Scientific Organization (NWO; grant PGW-12-39) and the German Academic Exchange Service (DAAD; scholarship 91526618 - 50015537). Moreover, this research received support by the German Research Foundation (DFG) in the Collaborative Research Center 673 “Alignment in Communication” and the Center of Excellence 277 “Cognitive Interaction Technology” (CITEC).

6. References

[1] Kendon, A., "Do gestures communicate? A review", Research on Language and Social Interaction, 27:175-200, 1994.

[2] Hostetter, A. B., "When do gestures communicate? A meta-analysis", Psychological bulletin, 137:297, 2011.

[3] McNeill, D., *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press, 1992.

[4] Feyereisen, P., Van de Wiele, M., and Dubois, F., "The meaning of gestures: What can be understood without speech?", Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 8:3-25, 1988.

[5] Krauss, R. M., Morrel-Samuels, P., and Colasante, C., "Do conversational hand gestures communicate?", Journal of Personality and Social Psychology, 61:743-754, 1991.

[6] Hadar, U. and Pinchas-Zamir, L., "The Semantic Specificity of Gesture Implications for Gesture Classification and Function", Journal of Language and Social Psychology, 23:204-214, 2004.

[7] Kibrik, A. A. and Molchanova, N. B., "Channels of multimodal communication: Relative contributions to discourse understanding," in Proceedings of the 35th Annual Conference of the Cognitive Science Society, Berlin, Germany, 2013.

[8] Schegloff, E. A., "On Some Gestures’ Relation to Talk," in J. M. Atkinson and J. Heritage [Ed.^Eds], Structures of Social Action, 266-298, Cambridge: Cambridge University Press, 1984.

[9] Kopp, S., "The spatial specificity of iconic gestures," in Proceedings of KogWis05 (the 7th Conference of the German Cognitive Science Society), 112-117, Basel: Schwabe, 2005.

[10] Cienki, A., "Image schemas and gesture," in B. Hampe and J. E. Grady [Ed.^Eds], From perception to meaning: Image schemas in cognitive linguistics, 421-441, Berlin: Mouton de Gruyter, 2005.

[11] Kelly, S. D., Özyürek, A., and Maris, E., "Two sides of the same coin: speech and gesture mutually interact to enhance comprehension", Psychological Science, 21:260-267, 2010.

[12] Kok, K., "The grammatical potential of co-speech gesture: a Functional Discourse Grammar perspective", Functions of Language, in press.

[13] Martinec, R., "Gestures that co-occur with speech as a systematic resource: the realization of experiential meanings in indexes", Social Semiotics, 14:193-213, 2004.

[14] Muntigl, P., "Modelling multiple semiotic systems: The case of gesture and speech," in E. Ventola, C. Charles, and M. Kaltenbacher [Ed.^Eds], Perspectives on multimodality, 31-49, 2004.

[15] Kok, K., Bergmann, K., Cienki, A., and Kopp, S., "Mapping out the multifunctionality of speakers' gestures", Gesture, 15, in press.

[16] Lücking, A., Bergman, K., Hahn, F., Kopp, S., and Rieser, H., "Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications", Journal on Multimodal User Interfaces, 7:5-18, 2013.

[17] McNeill, D., "Catchments and contexts: Non-modular factors in speech and gesture production," in D. McNeill [Ed.^Eds], Language and Gesture, 312-328, Cambridge: Cambridge University Press, 2000.

[18] Müller, C., "Iconicity and gesture," in S. Santi, I. Guaïtella, C. Cavé, and G. Konopczynski [Ed.^Eds], Oralité et gestualité, communication multimodale, interaction, 321-328, Paris: L'Harmattan, 1998.

[19] Bressem, J., "Repetitions in gesture: Structures, functions, and cognitive aspects," Viadrina European University PhD thesis, Frankfurt (Oder), Germany, 2012.

[20] Allwood, J., Ahlsén, E., Lund, J., and Sundqvist, J., "Multimodality in own communication management." vol. 92, in J. Allwood, B. Dorriots, and S. Nicholson [Ed.^Eds], Papers from the Second Nordic Conference on Multimodal Communication, 43-62, Göteborg, 2006.

[21] Ladewig, S., "Putting the cyclic gesture on a cognitive basis", CogniTextes. Revue de l’Association française de linguistique cognitive, 6:http://cognitextes.revues.org/406, 2011.

[22] Wundt, W., *The language of gestures*. The Hague: Mouton, 1973.

[23] Lausberg, H. and Sloetjes, H., "Coding gestural behavior with the NEUROGES-ELAN system", Behavior Research Methods, 41:841-849, 2009.

[24] Bressem, J., Ladewig, S. H., and Müller, C., "A linguistic annotation system for gestures. ," in C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and S. Teßendorf [Ed.^Eds], Body-Language-Communication: An International Handbook on Multimodality in Human Interaction, Berlin, Boston: De Gruyter Mouton, 2013.

Aspects of Digital Gesture Lexicography

Ulrike Lynn

Faculty of Humanities, Technische Universität Chemnitz, Germany

ulrike.lynn@phil.tu-chemnitz.de

Abstract

The multi-modality of speech requires an extension or revision of the applicable lexicography. This article will examine points to consider when designing or planning a gesture dictionary and will attempt to identify features that should be present regardless of the focus of the gesture collection. Where appropriate, these features are compared with similar structures in spoken language lexicography. We will discuss what advantages are to be gained by conceiving of the dictionary in digital from right from the outset and what problems may arise that are unique to a digital edition.

Index Terms: gesture, digital, dictionary, lexicography

1. Introduction

The computer technology that we have available to us today invites new and creative thinking about how a dictionary can be designed and realized. Ideas that were once unfeasible or unrealistic are suddenly possible. We will attempt to incorporate these possibilities into a logically structured conception of gesture lexicography. We will especially be interested in a non-linear presentation with flexible reference and search functions that could ideally be created on the fly to cater to a specific user's needs, and the inclusion of multimedia content to enhance the multi-modal aspects of gesture.

The aspects of lexicography we discuss here are based on the ‘Dictionary of Contemporary Physical Contact Gestures in the Mid-Atlantic region of the United States’ [1], currently the only dictionary of emblematic gestures, which we will use as a foundation for this article. The goal of that dictionary was to develop a structure for presenting, analyzing and organizing a specific type of gesture. We will take a closer look at the results and see if it is possible to extract and generalize features that could be used in a gesture dictionary regardless of focus.

In the course of compiling the dictionary, the author identified areas of future expansion that, due to constraints of time and space, did not find a place in that work. The possibility of a digital version sometime in the future was mentioned. It hardly seems relevant anymore, even just a few years later, to consider doing a print version of a gesture dictionary being newly planned. As we discuss the aspects of lexicography extracted from the aforementioned work, we will present ideas for how these elements can benefit from digitalization.

We intentionally use the term 'digitalization' as distinct from 'digitization'. This reflects the view that what we seek is more than the translation of a work designed for print into some digital form. This would be the digitization of an existing work into a pdf or even into html form. The work is not designed nor structured to take advantage of digital tools. We can think of digitization as a *static digital* form. Digitalization, on the other hand, would be planning the work from the outset using tools and techniques that the digital realm provides. Creating, in other words, a *dynamic digital* version that will change, adapt and grow through continued input from both author(s) and users.

Zimmer [9] makes a similar distinction between e-texts and hypertexts. E-texts, in this conception, are published online but do not take advantage of any of the expanded presentation possibilities that the internet allows. They can as well be printed out without any loss of information or usability. Hypertexts, on the other hand, can not exist on paper. They incorporate other media that can be combined and linked in any number of ways. They allow communication on the visual and auditory levels simultaneously and can only be accessed through a digital device.

The contrast between digitization and digitalization takes this a step further. It is not only the incorporation of multimedia elements and the ability to connect information with single-click access that we hope to bring to digital lexicography. We would also hope to add a fluid non-hierarchical structure that can respond to a user's needs as well as the ability to collaborate or incorporate feedback to create a dynamic work. This also avoids the potentially loaded term 'hypertext' which still tends to conjure an image of simple text that is linked to something else.

We will proceed logically through the foundation work to see what may be adapted or generalized to suit a broad range of desired outcomes. Another opportunity for expanding the ‘Dictionary of Contemporary Physical Contact Gestures’ [1] mentioned by the author is to investigate the historical development of the included gestures. This we will refer to as *etymology* even if it is so only in a broadened sense. With this in mind, we will use this as an example when we need one to discuss applying the generalized principles. We will then examine how each of these structural elements might possibly be improved through the use of a dynamic digital format rather than printed or otherwise static form. In conclusion we will address some questions that arise solely within the context of digitalization and will attempt to answer them.

2. Inclusion Criteria

Any gesture dictionary must be in some sense limited in scope. There will always be a concentration or a specific goal. The gestures selected for inclusion must necessarily be a finite set taken from a much larger pool of potential gestures. This is readily apparent in the context of the physical contact gesture dictionary which deals with a specific type of gesture, those involving physical contact, practiced within a specific geographical region, the Mid-Atlantic United States.

Gestures can be selected for inclusion for any of a number of reasons. They may, for example, fit a certain research concentration or share a specific attribute, such as being solely speech-accompanying, or they may be drawn from a specific culture or region. The decision of what to include may seem too obvious to deserve mention, but it is essential when creating a new dictionary. The selection criteria must be explicitly and exactly defined to prevent the work from expanding to unmanageable proportions, both for the author and for the user. It also makes clear, again for both author and user, what the main objective of the dictionary is and exactly what audience it it designed to serve. Inclusion criteria can significantly simplify the data collection process, though it is important to note that they can also be a form of preconception and should be treated as such. They have the potential to cause you to see the specific attribute everywhere or perhaps miss it in borderline cases that may subconsciously have already been decided.

A significant advantage to a dictionary in digital form is that a series of dictionaries, each with a different set of inclusion criteria, could be connected as modular units within a larger whole. Different groups of researchers could compile their dictionaries then use a centralized framework to publish their work. This allows a level of integration that would not be possible with individual books nor even with individual digital versions. An example helps illustrate this: research group A publishes a dictionary of gestures in cultural space *x* using a shared framework. Research group B then publishes a dictionary tracing the etymology of the gestures collected by group A. This information then becomes available as supplemental links within the original work. Should group C then publish a dictionary of gestures drawn from cultural space *y* within the framework, the possibility for comparative study is greatly expanded.

3. Data Collection

The particular details of this stage are so unique to each project that it is hardly possible to generalize an approach. That it is arguably the most important stage in a work should be apparent to anyone deciding to undertake a gesture dictionary. This is, however, an area that can be radically changed in the context of a digital edition. Gestures are primarily observed directly, whether through live interaction or through recorded visual media. It is rare to find written descriptions that are sufficiently detailed for lexicographic purposes. Modern techniques of data mining that may be suitable for the written word are not yet advanced enough to parse an equivalent amount of information that must be visually extracted. Primary source material, once collected, is then condensed into what becomes the dictionary entries but then usually disappears into the background. Once freed from traditional space constraints of a book format, however, some portion of this supporting material could be accessed, if needed, from the main dictionary entry. This has a parallel in the use of example sentences used to demonstrate use of a word in a spoken-language dictionary.

As we will see later, the context in which a gesture can be found plays a role in the organization of the dictionary. The explanation and presentation of this context information could be well established and presented by multimedia source material. Even something so simple as the frequency of an observed gesture as documented in source materials could help provide a general clue as to the pervasiveness of a particular gesture.

Data collection can also be greatly expanded once a dictionary is planned in digital form. The idea of crowd sourcing information is an interesting one and similar to wiktionary-type trends in spoken language dictionaries. In addition to the same questions of organization, quality control and authorship that arise in that context, a gesture dictionary also presents unique hurdles. The necessarily large amount of photo and video information makes distributed collaboration more difficult than simple editing of text within a browser. Using distributed collaboration could, however, greatly enhance the resolution and accuracy of collecting gestures for inclusion out of a given gesture space. We will come back to this idea of collaboration later as we discuss potential pitfalls in the digitalization of gesture lexicography.

4. Presentation

Once the inclusion criteria have been defined and a set of gestures collected through suitable means, they must be presented in some logical form. Here we are not talking about the organization of all the entries, this we will discuss in terms of navigation, but rather the presentation of each specific gesture entry. Here we already digress from speech dictionaries in that photo/video material or illustration is more or less required. While a written description of the gesture may or may not accompany the entry, it is almost unthinkable to leave out a visual representation.

When using still photos it is important to capture the essence, Kendon's *stroke* [6], of the gesture. The ‘Dictionary of Contemporary Physical Contact Gestures’ [1] provides a sequence of still images where the preparation or post-gesture movement sequences are important to understanding the gesture. This problem of selecting the 'right' moment of a gesture goes away in the context of a dynamic digital edition. With only minimally more production overhead, videos can be produced and embedded that depict the entire gesture sequence; from pre-gesture state of rest to post-gesture state of rest. In this way, no decision has to be arbitrarily made as regards the essence of the gesture.

This brings us to the question of where such photo or video material should come from. In our source dictionary, one finds the gestures collected recreated for the purposes of the photos in order to provide a unity of presentation. This is a practical as well as an aesthetic choice. It provides a level of control so that extraneous elements can be eliminated from the scene. It also provides a uniform experience for the user when cross-referencing or comparing gestures. This centralized production of multimedia material, as we will discuss later, is one of the main obstacles to group collaboration. During the data collection phase, however, much additional video and photo material was collected and there is no reason this cannot provide background or additional information as needed. It is nearly unthinkable to plan an etymological gesture dictionary without multimedia depictions of various stages of the gesture development. The changes over time may be too subtle to capture in words or could be made dramatically clear when presented sequentially in the course of a single video.

Depending on the goal of the dictionary, it may also be desirable to shift focus away from or otherwise completely remove certain elements from the multimedia depiction. In the ‘Dictionary of Contemporary Physical Contact Gestures’ [1], a decision was made to blur the facial expressions of the subjects performing the gestures for the photographic material as this detracted from the analysis of the communicative intent of the gesture and in some cases collapsed the polysemy of the gesture meaning. In deciding to recreate gestures and produce original material it is important for extraneous elements to fade into the background as much as possible so that the foreground is reserved solely for the gesture.

Comprehensive video material may also render written descriptions of a gesture extraneous. However, it must be said that a written description can subtly focus the viewer on the key elements of the gesture that are then reinforced by the visual depiction.

Another possibility provided by digitalization is the simple use of sound in cases of accompanying speech or colloquial spoken renderings of gestures. These could easily be embedded as audio files to enhance the user experience and also to avoid confusion when dealing with the sometimes creative spelling of interjections or sounds uttered while performing a gesture.

4.1. Naming

Naming the entries should not be taken lightly. We will discuss this in more detail in the context of navigation, but it deserves mention here. The name is the first point of contact for the user with a gesture and establishes it as a formal entity. It should be easy to understand what gesture is meant without going so far as to provide initial analysis. Names can be a brief description of the movement but it is important that they remain on the morphological level without being burdened with meaning. As an example; 'Placing a hand on someone's back' is preferable to 'Demonstrating Support', which would already provide information at the level of meaning.

5. Definition

5.1. Analytical Tools

For a collection of gestures to truly be considered a dictionary, it is important to provide some framework of analysis that allows a discussion of meaning. What exactly this framework is could be unique to each project. For the 'Dictionary of Physical Contact Gestures' [1], the main focus was on the communicative intent and content of the gestures. They were analyzed, therefore, in terms of Austin/Searle's [8] speech-act theory. This provided analytical tools that allowed classification and structured interpretation of the communication. It may be, however, that a specific dictionary is not concerned primarily with communication. By way of example, a dictionary concerned with tracing the development of use-movement into gesture would have a different focus and, therefore, need a different set of analytical tools, perhaps those of philology or semantic shift. In our view, it is desirable to see what can be borrowed from the world of spoken communications. These theories provide a generally more robust context for analysis as they will be better tested than anything developed uniquely for a gesture dictionary. They also strengthen the connection between spoken and non-verbal communication and can provide an element of familiarity. The key, of course, is that a system of analysis be defined and applied consistently. This is ultimately what supplies any dictionary with its usefulness and will play a deciding role in judging its quality.

5.2. Context

It may be important, again depending on the focus of the dictionary, to stipulate certain contexts in which the gesture may or may not be found. This, in our opinion, should be part of the analytical framework in order to have a clear taxonomy of contexts. One option would be the register labels of the Oxford Dictionary to provide predefined descriptions of the tonality of a gesture to help to contextualize them.

It is also advisable to indicate when a gesture belongs to the passive gesture vocabulary of a particular region or group. These would be gestures that are recognized but not actively used. This can be important for completeness when documenting the gestures of a given time and place.

6. Variations

6.1. Expression

Gestures are fluid and their use is constantly developing. It is a given that during the data collection, many gestures will be observed with slight variations at the morphological level. It is best to identify a stereotype for the gesture. That can be decided through frequency of use or even through simplicity; the unadorned version of the gesture becomes the dictionary entry and any other versions are treated as a variation of that. The decision as to what constitutes a variation must be made by the lexicographer. Taking a handshake as an example; we would hardly consider turning the hand 45 degrees from the vertical plane a variation, though it was witnessed many times. A speech dictionary also does not attempt to document every possible regional difference in pronunciation. Adding a second hand to clasp the shaking hand, though, is different enough to deserve mention but is not, in and of itself, a unique gesture. The list of variations at the level of expression should be comprehensive enough to cover most use cases without being so detailed as to overburden the user. These variations, like the main entry, can benefit from embedding in video form

6.2. Meaning and Use

As seen above, once a gesture has been classified in some way using the analytical tools of choice, it is possible to compare or contrast variations at the level of meaning. In the case of a dictionary primarily concerned with communicative content, one gesture generally has several meaning variations; a range of communicative content intended by use of the gesture. In this case, a single entry will have a number of 'definitions'.

6.3. Combining

Another pertinent fact can be the combinability of gestures. It may be, rather than being an extension or variation of a given gesture, that a gesture with its own individual entry may be combined with another independent gesture in the course of one communicative act. We are speaking here of simultaneous performance of the gestures and not successive performance as would be found in gesture dialogue. The scope of possibilities of successive gesture chains is simply too broad to be a part of a gesture dictionary unless, of course, this is the primary focus of the work. One could imagine, however, that in the context of a digital version complex chains *n* gestures long might be generated algorithmically then verified through human input. Gestures that are frequently combined should be indicated as such to facilitate cross-reference and to underscore their combined use in a gesture vocabulary. Some gestures are not combinable but it would only be in a special case that it would make sense to indicate non-combinability.

7. Navigation

Non-linear organization differentiates dynamic digital editions from both linear audio and video documents as well as documents designed for print. Here we will see the true advantages of this mode of presentation once a dictionary is freed from the limitations of a print-based edition.

A crucial element of lexicography is the design of the user interaction. Creating a navigation structure demands understanding the target audience and thinking through a variety of use cases. Each project will have different demands, again based on the focus of the specific dictionary. It will likely prove beneficial to have gestures be locatable by using either the expression or the meaning levels. That means if a user has seen a gesture but doesn't know what it means, they can search based on the physical action. This can be through the titles which, we have seen, are a brief description of the movement involved.

Should a user, however, want to collect all gestures that communicate something specific, they could consult an index which categorizes the gestures according to general communicative intent. The following categories are to be found in the 'Contemporary Physical Contact Gestures' [1]: Greeting Gestures, Attention Gestures, Confirmation Gestures, Institution Gestures, Consolation Gestures, Encouragement Gestures, Affection Gestures, Attraction Gestures, Sexual Relation Gestures, Assistance Gestures, Aggression Gestures, Playful Power Gestures and Indirect Contact Gestures.

Whether or not these categories are useful to a specific project is irrelevant. The main point is that there should be some sort of organization based on the level of meaning that allows gestures with similar communicative intent to be easily discovered and grouped.

7.1. Linking

Clearly one of the main advantages to digitalization is to be found in dynamic linking. Instead of a table of contents or an index full of page numbers, each entry is simply accessible as a link within the current view. This also obviates the need for a numbering system which can have the impression of giving the entries a hierarchy that is not intended. Additionally, a print edition can only have a limited number of navigation methods. Otherwise the problem becomes how to notate them all for a user and still retain some sense of order. If one gesture links to ten others because of their similarity of form and another twenty due to their similarity of meaning, there is hardly room for any other navigation criteria in the jungle of page numbers and cross-references. This is, naturally, not a problem in digital form. As each gesture is only one click away from any other, there is no longer reason to avoid bringing any two in relation when the user no longer needs to flip endlessly through a book. It is easy to imagine in addition to two main navigation schemes - at the expression and meaning levels - there being any number of navigation matrices that are customized to a user's needs. As they would only be treating the dictionary entries as a dataset, they could be added and removed after the fact without altering the work itself. Each type of navigation could remain in place and be dynamically populated with the content appropriate to the gesture currently being viewed.

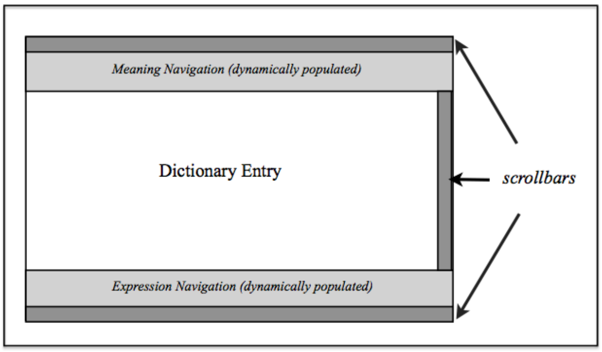

Meaning Navigation (dynamically populated)
Dictionary Entry
scrollbars
Expression Navigation (dynamically populated)

Figure 1: Example of a possible navigation scheme.

7.2. Performance Navigation

Yet another possibility in the digital realm, though perhaps not immediately realizable, is navigation of a gesture dictionary simply by performing the gesture in question. This is, naturally, very much dependent on the final delivery medium of a digital gesture dictionary. Assuming, however, that it is designed to be viewed on a camera-equipped device, it is not hard to imagine gesture performance-based navigation. With sinking costs of such technology and commercial products such as the Kinect or the Myo easily available, it will not be long before this is feasible.

8. Discussion

We have seen along the way that there are many advantages to be gained by conceiving of gesture dictionaries in digital form. There are, however, some issues that need to be addressed. In many ways, these are the same problems that are currently the focus of so much debate in the digital humanities in general. Any new plan to create a gesture dictionary, however, should at least consider ways to mitigate these issues even if they can not be solved completely.

8.1. Medium

The first question is in what format the dictionary should be made available. It is one thing to say 'a digital version rather than print' but that implies that there is only one possible digital version. Generally, saying this means wanting to build a website around the content. But even this is not as straightforward as would be hoped. First of all it is important that the dictionary remain accessible and compatible for as long as possible. This may not be that difficult in the case of purely text as most speech dictionaries are, but even just talking about embedding video complicates the issue. A simple web version has to support the wide (and ever increasing) variety of devices with which users will access it. It should, in the interest of longevity and compatibility, adhere to web standards. All multimedia content needs to be unencumbered by intellectual property issues and should, ideally, be in a format that is also free to use, and that will remain so. An active site will need server space and bandwidth, and that for the entire planned life of the work. It will also require maintenance, both routine and proactive to avoid obsolescence.

Which brings us to:

8.2. Competence

In planning a publication team, it is important to have people with hard digital skills. In concrete terms, this means programmers, possibly interface designers and someone who can 'publish' and maintain the dictionary. If this means interdisciplinary collaboration, that is all the better. It does, however, need to be considered when budgeting and planning.

It should also be taken into account that if the gesture researcher(s) themselves are not doing the programming that there is a possibly detrimental time-lag between idea and realization that should also be taken into account.

8.3. Collaboration

As touched upon briefly above, digitalization provides robust and powerful tools for collaboration. These tools, however, can not simply be blindly implemented. There is the primary question 'who has access to these tools?' That is to say, with whom do we want to collaborate? Naturally if one is designing a dictionary focusing on a specific culture or region, it would be helpful to have input from its members as regards to the performance and interpretation of meaning of gestures. But when the net is cast too wide, the quality of the information suffers. Not only that, as we saw in section 2, a very specific definition of the target gestures is essential to the success of a project which will also potentially limit the group of potential collaborators.

Perhaps contribution to the dictionary is strictly limited to users; scholars, researchers or academics. Who then decides who has access and what are the restrictions in place to prevent abuse or vandalism? In many ways, these questions have been addressed by many wiki-based projects. But as already noted, what happens when editing is not simply a matter of typing text into an input field? How would crowdsourced multimedia content be produced? Would it need to be centrally produced to the specifications of the 'crowd'?

This brings up the problem of authorship. For better or worse, the academic field is a meritocracy. If a group of authors sets a work loose for digital collaboration, how much credit do they get? Will an institution support a project with running maintenance costs that is so widely distributed as to bring no immediate prestige gain? A new proposal by the creator of the wiki concept for so-called *federated wikis* attempts to alleviate some of this confusion by preserving articles intact but allowing them to be changed when incorporated into another's work. (https://medium.com/backchannel/the-failed-promise-of-deep-links-aa307b3abaa5 and http://fed.wiki.org/)

One can imagine a mechanism by which users can easily submit feedback regarding the content or usability of a digital dictionary. This would then need to be curated and, as needed, implemented by the authors. Or, as in section 7.1 - Linking, there could be an interface allowing custom interaction with the dictionary as dataset. Inventive methods of accessing or cross-referencing the data could then potentially be integrated into the dictionary itself. Another option would be to have the possibility to preserve a user's *link trail* - the path they chose on the fly to navigate the dictionary - as an option that can be shared with future visitors. This would be a way of sharing not just a link but access to a specific interaction with the information. Yet another variation would be to have the data collection phase open for collaboration with as few barriers to users as possible. Once the phase is deemed completed, the collected input is collated and published by the authors in a much more restricted way.

A dictionary can, of course, be authored by a core group and the collaboration limited in scope so as not to cause any confusion as to authorship. Our only point is that the level and mechanism of collaboration should be discussed right from the start.

9. Conclusions

We have proposed a series of modular characteristics that should be present in any type of gesture dictionary and discussed ideas for how to enrich a dictionary through the use of currently available digital tools. We made a distinction between dynamic and static digital editions in which the dynamic fully embraces the possibilities for interaction, multimedia content, collaboration and fluidity of presentation that technology allows. A static digital edition, on the other hand, is rooted in print-based thinking and is, in our opinion, no longer relevant in our increasingly interconnected lives. Some of the issues that arise in planning and implementing such a dictionary have been presented with the understanding that, as the field matures and more works are published in this manner, some of these problems will disappear while others, inconceivable at the moment, may take their place. We have the possibility now to realize ideas in lexicography that until now were impractical or even outlandish. The nascent field of gesture lexicography is, unburdened as it is with the residue of a long history, is the perfect place to stretch the limits of what is considered possible in creating a reference work.

10. References

[1] Author, “Keep in Touch – A Dictionary of Contemporary Physical Contact Gestures in the Mid-Atlantic region of the United States”, Phd Dissertation, Technische Universität Berlin: Digital Repository of Technische Universität Berlin. Online: https://opus4.kobv.de/opus4tuberlin/frontdoor/index/index/docId/3484, 2011

[2] Author, Gestures in dictionaries: Physical contact gestures. In: C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill and J. Bressem (eds*.), Body – Language – Communication. An International Handbook on Multimodality in Human Interaction* (HSK 38.2). De Gruyter Mouton, 2014. 1502-1511

[3] Author, Levels of abstraction. In: C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill and J. Bressem (eds*.), Body – Language – Communication. An International Handbook on Multimodality in Human Interaction* (HSK 38.2). De Gruyter Mouton, 2014. 1702-1712

[4] Fricke, E., J. Bressem and C. Müller, Gesture families and gestural fields. In: C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill and J. Bressem (eds*.), Body – Language – Communication. An International Handbook on Multimodality in Human Interaction* (HSK 38.2). De Gruyter Mouton, 2014. 1630–1640.

[5] Kendon, A., Die wechselseitige Einbettung von Rede und Geste. In: C. Schmauser, T. Noll (Hrsg.), *Körperbewegungen und ihre Bedeutungen*. Arno Spitz, 1998.

[6] Kendon, A., *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.

[7] Noll, T., Wie blättert man in einem Gestikon? In: C. Schmauser, T. Noll (Hrsg.), *Körperbewegungen und ihre Bedeutungen*. Arno Spitz, 1998.

[8] Searle, J., Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, 1969.

[9] Zimmer, E. D., *Die Bibliothek der Zukunft*. Hoffmann und Campe, 2000.

Coming of age in gesture: A comparative study of gesturing and pantomiming in older children and adults

Ingrid Masson-Carro, Martijn Goudbeek, Emiel Krahmer

Tilburg center for Cognition and Communication (TiCC), Tilburg University, The Netherlands

i.massoncarro@uvt.nl, m.b.goudbeek@uvt.nl, e.j.krahmer@uvt.nl

Abstract

Research on the co-development of gestures and speech mainly focuses on children in early phases of language acquisition. This study investigates how children in later development use gestures to communicate, and whether the strategies they use are similar to adults'. Using a referential paradigm, we compared pantomimes and gestures produced by children (M=9) and adults, and found both groups to use gestures similarly when pantomiming, but differently in spontaneously-produced gestures (in terms of frequency of gesturing, and of the representation techniques chosen to depict the objects). This suggests that older children have the necessary tools for full gestural expressivity, but when speech is available they rely less on gestures than adults, indicating both streams aren't fully integrated yet.

Index Terms: Gesture, pantomime, representation techniques, development, older children, adults.

1. Introduction

The co-development of speech and gestures has been thoroughly studied, especially regarding the first years of life. This is not surprising, given that gestures are an invaluable source of information when it comes to studying cognitive development [1]. First, looking at the development of gestures helps researchers understand how both modalities (gestures and speech) become an integrated system in spontaneous talk [2], [3], [4]. Second, looking at the type of gestures produced by children provides researchers with a "window" into the development of conceptual representations, and (arguably) the shift towards symbolic thinking (e.g., [5], [6], [7]). Typically, these studies have been conducted as children learn their first words and transition into the two-word stage (approximately between 17 and 23 months of age [3]). However, few studies have addressed gesture development after this phase [4], and even less studies have looked at how older school-aged children (e.g., after the age of 6) produce gestures –despite the fact that the gestural system is thought to keep developing until adolescence [2].

In the present paper we look at how older children use gestures in referential communication, and we compare their performance to that of adults, with the aim to find out not only whether children and adults accompany their descriptions with gestures to a similar extent, but also whether their gestures exhibit similar patterns, in terms of the representation techniques [8] used to represent objects. Analyzing these techniques provides valuable information about the iconic strategies used by children to translate mental representations into gestures, and often the choice of technique can be seen as an indicator of cognitive development (e.g., see [6]).

Furthermore, we investigate gestures in two communicative modalities, namely speech and gesture, and gesture-only (or *pantomime* [2]), to assess how both age groups use representation techniques to express meaning when gestures play a primary, or a secondary role in communication.

1.1. Becoming a "mature" gesturer

Several studies have helped define a series of stages in the co-development of gesture and speech. It is well-documented that children start producing gestures before they start producing their first words [2], [9], [3], [4]. At this stage, children combine vocalizations with deictic gestures, produced to direct their caregiver's attention towards objects in the environment. Around the age of twelve months, children start producing their first words, and their first iconic gestures [9], [3]. Importantly, at this age there is little integration of the gestural and speech modalities, with children referring to objects by either producing a gesture or uttering a word, but generally not both at the same time. The first gesture-word combinations emerge around 14 months, preceding (and perhaps facilitating) the onset of two-word combinations, provided that these are not simply gesture-word co-occurrences, but that they together convey idea units [3], [10]. Not many studies have systematically analyzed how gestures and speech continue to co-develop after the two-word stage, with a few exceptions. For instance, Mayberry and Nicoladis [11] conducted a longitudinal study following 5 boys between the ages of 2 and 3;6 years old, and showed that already at 2 years old children used gestures in combination with speech, but there were differences regarding the type of gestures children produced, in comparison with adults. For instance, these gestures remained deictic in their majority (in contrast to adults, who produce deictic gestures to a fairly low extent), with iconic and beat gestures increasing with age as language constructions became more complex. After this stage, the production of iconic and beat gestures continues to develop throughout the third, fourth and fifth years of life [2].

Stefanini, Bello, Caselli, Iverson and Volterra [12] looked at how children aged between 2 and 7 years represented objects and actions in gesture during a naming task. Their findings suggested that the production of spontaneous gestures decreased with age, but did not disappear (even at the age when children had sufficient vocabulary to simply name the objects). This decrease in gesture production was particularly pronounced for deictic gestures, which were the most produced gesture type in all age categories. This indicates that there is a progression, as children age, towards producing less deictic and more iconic gestures. But, how does their gesturing compare to adults'? In a narrative context, Alibali and colleagues [13] examined how children aged 5 to 10 gestured while retelling a cartoon, as compared with college students. They found no significant differences in the amount of gesturing between adults and children, but they did find differences in terms of how "redundant" gestures were in relation to speech, with children producing less redundant speech-gesture combinations than adults. In sum, these studies suggest that the relationship between speech and gesture is not stable throughout childhood, and keeps on changing during later developmental stages, possibly until adolescence [2].

1.2. Representation techniques and symbolic thinking

So far, we looked at the amount and type of gestures produced by children during early linguistic development. But what about the type of information these gestures convey? Speakers are known to combine different techniques when they depict referents in gesture (e.g., [8], [14]) and in pantomime (e.g., [15]). For instance, in describing a clock, a speaker may draw a circle in the air, or tilt an extended finger to the left and to the right, pretending the finger to be the clock hand. In her work, Müller [8] recognizes four basic representation modes, often employed by speakers in spontaneous gesturing. The hands *imitate*, when they pretend to use an imaginary object; they *portray*, when they pretend to be an object or character; they *draw*, when they trace a silhouette in the air; and they *mold* when they pretend to "sculpt" shapes. These techniques reveal information about how speakers conceptualize objects. Previous work has addressed the question of how specific object characteristics influence the choice of representation technique seen in speakers' gestures. For instance, Masson-Carro and colleagues [14] found that speakers used mostly imitating gestures when describing manipulable objects, than when describing non-manipulable objects, where their exhibited a tendency towards shape gestures. In this study, we expand this line of research by looking the influence of age on the choice of representation technique.

A few studies have addressed the use of representation techniques by children at different stages of language acquisition. Overton and Jackson [5] asked children aged 3, 4, 6, and 8 years old to pantomime the typical use of a series of common objects. Their study was one of the first to reveal a representational shift from "body part as object" gestures (using the index finger as a toothbrush –also called portraying gestures) to "symbolic" gestures (e.g., hand grabs imaginary toothbrush by handle, and pretends to brush teeth –also called imitating gestures). Thus, in about 80% of all the gestures observed in 3 year olds, children used their own body to represent objects, and this decreased the older children got, in favor of gestures where children pretended to use an object directly. By age 8, symbolic gestures constituted nearly 70% of the gestures produced by children, and body part as object gestures only the remaining 30%. Several studies have replicated this finding. For instance, Boyatzis and Watson [6] asked 3, 4 and 5 year olds to pretend to use 8 common objects, and also found a preference for body part as object gestures in 3 year olds (80%), but a preference for imaginary object use at age 5 (69%). In a second experiment, they explored the ability of these children to imitate a series of gestures executed by the experimenter, and found that younger children had trouble to reproduce imaginary-object gestures, in comparison with older children. A study by O'Reilly [7] showed that, at age 3, not only do children have trouble producing these imaginary-object gestures, but they also have trouble with comprehending these representations. In a narrative context, McNeill [2] describes a similar phenomenon. He examined cartoon retellings in children aged 2, 5, and 8, and compared their gestures with those produced by adults in the retelling of the same cartoons, to find that older children (aged 8) exhibited a mix of mature and immature gestural features, with a tendency to produce "enacting" gestures that was not found in adults. In conclusion, these studies show that during the first years of life, a cognitive shift takes place, as children begin to understand and produce (iconic) gestures, not purely as actions, but as communicative *symbols*. In this respect, gestures act as indicators of the transition from action to abstraction, from physical to conceptual knowledge; and this transition can be seen as a milestone in the development of symbolic thought.

1.3. The present study

In the present paper we ask the following question: Do children in late developmental stages use gestures entirely similarly to adults? We aim to find out by examining the gestural strategies employed by older children (mean age 9) in referential communication about objects, and comparing them with those employed by adults (mean age 25). Given that previous studies have shown that children and adults gesture differently about manipulable, and non-manipulable objects (e.g., [16], [14]), we will control for object-type by including "manipulability" as a variable in our design[1].

We examine gestural behavior in two communication modalities that differ in the extent to which they are tied to speech, namely speech and gesture (henceforth, speech), and gesture-only (henceforth, pantomime). Pantomimes, in the context of this study, are defined as gestures that occur in the absence of speech [2]. Like co-speech gestures, pantomimes are not conventionalized; however, unlike co-speech gestures, pantomimes must be sufficiently informative to be interpreted on their own. Thus, this allows us to make a first exploration of the techniques used by speakers in gesturing, not only at different developmental points, but also in different modalities, allowing us to gain insight into several aspects of gesture production, for instance, about the extent to which the choice of a representation technique is dependent on speech production.

2. Method

2.1. Participants

20 adults and 20 children participated in this study. The adults were students of Tilburg University (Mean age = 25.5 years, 7 male), and participated in exchange for partial course credit. The children taking part in this study (Mean age = 9 years, 10 male) were members of the Scouting Rambonnetgroep in Naaldwijk (The Netherlands) and participated voluntarily, after receiving written consent from their legal tutors. All participants were native speakers of Dutch.

2.2. Stimuli

The stimuli was composed by eleven manipulable objects (book, eraser, pencil, ruler, sharpener, stapler, scotch tape, scissors, calculator, brush and shovel), and eleven non-manipulable objects (tree, slide, sandpit, blackboard, table, school, chair, treehouse, clock, shelves, seesaw). Manipulable objects were defined as "objects operated with the hands, whose operation may induce a change in the physical world". For instance, the use of a pair of scissors typically results into the division of a sheet of paper into smaller units. All items were compiled into two presentation documents (one for manipulable, and one for non-manipulable objects), plus two counterbalanced versions. The stimuli were shown to the speakers by the experimenter on a 10'' Ipad, where the items

[1] While we acknowledge the effect of *manipulability* is interesting in itself, its discussion falls beyond the scope of the present paper and thus we mainly focus on the influence of age and modality on gesturing.

were displayed full-screen. A digital video camera was placed behind the addressee, to record the speaker’s speech and gestures.

2.3. Procedure

The experiment was carried out in pairs. Each pair was assigned to a condition, namely speech, or pantomime, in turns (e.g. A-B-A-B). The task was introduced to the participants as a guessing game, such like *Taboo* (in the speech and gesture condition), or *Charades* (in the pantomime condition). The procedure was as follows: Participant A described eleven objects (either manipulable or non-manipulable) to participant B, one by one. Participant B had to guess the name of the object being described, and say it out loud. Once the first eleven objects were described, roles were reversed, and participant B described the remaining eleven objects to participant A –for instance, non-manipulable objects, if participant A had described manipulable objects.

2.4. Data analysis

We annotated all the gestures produced by the speakers using the multimodal annotation tool Elan [17]. We classified gestures according to four main gesture types, namely iconic gestures [2], pointing gestures, interactive gestures directed at the addressee [18] and other (e.g., emblems, beats). All gestures were coded from preparation to retraction.

Next, we annotated all iconic gestures for representation technique. We annotated six representation techniques: *portraying*, *molding*, and *tracing* (based on Müller [8]), *enacting*, *object use*, and *object use + portray* (dual) (subdivision of Müller’s *imitating* gestures). We added a seventh category to account for gestures that did not fit any other type, coded as *other*. Definitions and examples are provided in Table 1.

Table 1. Coding scheme for representation techniques

Representation Technique	**Description**
Object use	The actor simulates the performance of an object-directed action. Example: pretend to hold a pencil, and write
Portraying	The hand is used to portray an object, as if it had become the object itself. Example: the hand portrays a pair of scissors, with index and middle fingers stretched out, and simulates the action of cutting through paper.
Use & Portraying	One hand portrays an object, while the other performs an object-directed action. Example: one hand portrays a book, with a flat palm facing up, while the other hand pretends to turn the pages of the book.
Enacting	The actor simulates the performance of an intransitive action. Example: the whole arms swing back and forth in alternated movements, simulating the motion of the upper body while running.
Molding	The hand molds or sculpts the shape of an object. Example: a flat hand with the palm facing down moves along the horizontal axis, representing the “flatness” of an object’s surface.
Tracing	The hand draws a shape in the air with a stretched index finger. Example: tracing a big square with the tip of the finger to represent a quadratic object such as a window.
Other	Gestures that do not fit other categories (e.g., using the fingers to count)

2.5. Design and statistical analyses

The effects of manipulability (manipulable, non-manipulable), age (children, adults), and modality (speech, pantomime) on our dependent variables (gesture rate, and representation technique) were assessed using linear mixed models for continuous variables (i.e., gesture rates), and logit mixed models for categorical variables (i.e., representation techniques) (see [19]). In all of the analyses, participants and items were included as random factors. Due to space limitations, our results section will only report test values for significant results.

3. Results

The communication task generated 420 descriptions, containing a total of 1497 gestures. Iconic gestures accounted for 74% (1098) of the gestures annotated, the remaining 26% consisting of other gesture types (deictics 6%, interactive gestures 12%, and other gestures 8%). With the exception of iconic gestures (discussed below), the type of gestures produced by speakers was not influenced by age, manipulability, or modality. The remainder of this section focuses on the iconic gestures produced by speakers.

3.1. Analysis of iconic gesture rate

We analyzed the effects of our independent variables on the mean number of iconic gestures produced per description. Not surprisingly, we found a main effect of modality ($\beta = -3.6205$, $SE = 0.48$, $p < .001$), indicating that speakers who accomplished the task in the pantomime condition (no speech) produced more gestures ($M = 4.34$, $SD = 3.67$) than speakers who could both speak and gesture ($M = 1$, $SD = 1.97$). We found no main effects of age on the production of iconic gestures, but a significant interaction between age and modality ($\beta = -1.64$, $SE = .77$, $p < .001$), showing that children produced more gestures than adults in the pantomime condition, but less gestures than adults in the speech condition (see Figure 1). In contrast to [14], we found no evidence that children or adults gestured more frequently about manipulable than about non-manipulable objects

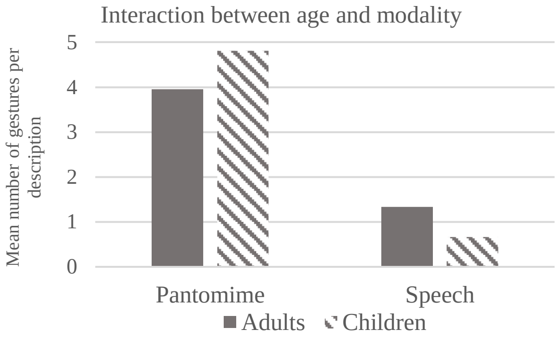


Figure 1: *Mean of gestures per description produced by older children and adults, in the pantomime and speech conditions. (Interaction significant at p < 0.001)*

3.2. Analysis of representation techniques

Our analyses of the representation techniques yield interesting insights. First of all, molding and object use were the most preferred techniques used to represent objects gesturally, together accounting for 60% of all gestures produced. Both age (Figure 2) and manipulability influenced the use of several techniques to represent objects. Age was found to influence the preference for object use gestures, whereby the speaker pretends to carry out an object-directed action (β = .83, *SE* = .37, *p* < .05), with children exhibiting more object use gestures (*M* = .45, *SD* = .49) than adults (*M* = .4, *SD* = .49). Similarly, children also used more object use gestures in combination with portraying gestures (*M* = .11, *SD* = .18) than adults (*M* = .03, *SD* = .31) (β = 1.7, *SE* = .37, *p* < .001). In contrast, adults exhibited more molding gestures (*M* = .24, *SD* = .42) than children (*M* = .18, *SD* = .38) (β = -.84, *SE* = .4, *p* < .05).

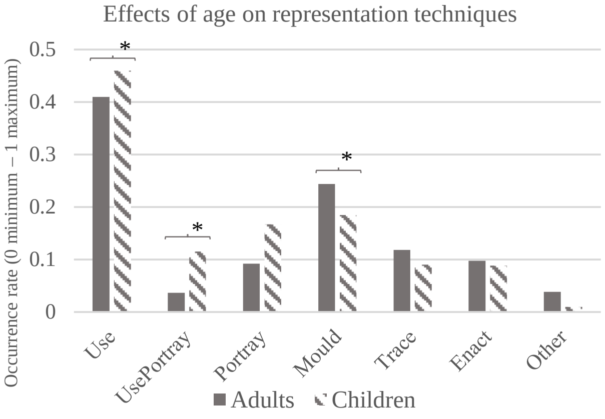


Figure 2: *Effect of age on the representation techniques. *Significant at p < 0.05, **significant at p < 0.001.*

There was no main effect of modality, but several interactions were found in the data between age and modality (Figure 3). It is interesting to note that all three interactions occur at the level of speech. While no differences are observed for pantomime, in speech we find that children produced more enacting (β = -2.46, SE = .88, p < .01) and (marginally) more object use gestures (β = -1.1, SE = .58, p = .059) than adults. The opposite pattern is found for molding gestures, where adults produced more gestures than children (β = .99, SE = .46, p < .05).

As expected, manipulability affected the choice of representation technique. Object use gestures accompanied more often manipulable (*M* = .69, *SD* = .45) than non-manipulable objects (*M* = .18, *SD* = .39) (β = -3.41, *SE* = .77, *p* < .001), and the same was found for gestures where object use was combined with portraying gestures (β = -3.14, *SE* = 1.42, *p* < .05 [manipulable *M* = .13, *SD* = .34; non-manipulable *M* = .01, *SD* = .12]). In contrast, non-manipulable objects were more often gestured by using molding (β = 1.87, *SE* = .46, *p* < .001 [manipulable *M* = .09, *SD* = .29; non-manipulable *M* = .32, *SD* = .46]), tracing (β = 1.87, *SE* = .46, *p* < .001 [manipulable *M* = 0.6, *SD* = .24; non-manipulable *M* = .14, *SD* = .35]), and enacting gestures (β = 2.71, *SE* = 1.18, *p* < .05 [manipulable *M* = .007, *SD* = .08; non-manipulable *M* = .17, *SD* = .37]). Lastly, manipulability did not interact with age, or modality, showing that its effects on the representation techniques used are independent.

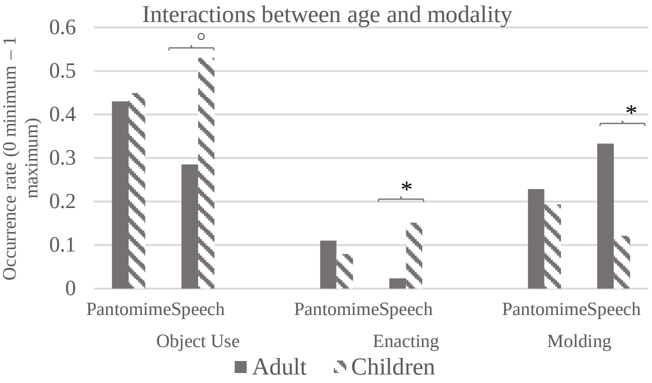


Figure 3: *Interactions between age and modality regarding the use of representation techniques. *Significant at p < 0.05, **significant at p < 0.001, ° <.1.*

4. Discussion

In the present study, children and adults were asked to either pantomime (only gesture) or verbally describe (speech and gesture) a series of items to a peer. We measured the occurrence of iconic and non-iconic gesture types, and annotated the representation techniques that speakers used to convey meaning. In addition, we also manipulated the type of objects that had to be described, by including manipulable, and non-manipulable objects.

We first looked at the type of gestures produced by speakers. We first analyzed the type of gesture used, and found that the vast majority of the gestures produced by both children and adults were iconic. This means that other gesture types, such as deictics (which constituted a 6% of all the gestures annotated), were performed to the same extent both by children and adults, regardless of the communication modality and object type. These results extend the findings by previous studies, for instance Stefanini et al. [12], who showed a decrease in pointing gestures between the ages of three and four, with already low deictic rates by the age of 6;4. Thus, it appears that around the age of 9 the use of pointing gestures has decreased to adult-like levels, at least for referential tasks in which pointing is not required. Both in the speech and pantomime conditions, for instance, children still used pointing to directly refer to the location of referents outside the room (e.g., to indicate the trees outside, or the blackboard downstairs). The same was observed in adults, but adults displayed a type of pointing that children did not, namely they pointed directly at their own gestures to highlight or clarify what the referent is. For instance, in portraying gestures where the hand pretended to be an object, speakers often used the other hand to point at the portraying hand, to indicate that it is not the action but the object what was relevant.

The remainder of the discussion section, we focus on iconic gestures. We found that participants produced more iconic gestures in the pantomime than in the gesture condition. This did not come as a surprise, as in the pantomime condition the use of gestures was obligatory, whereas in the speech condition no instructions were given concerning the use of gestures, so gestures are assumed to have arisen spontaneously. We found an interaction between age and modality, meaning that children gestured more than adults in the pantomime condition, but less than adults in the gesture condition. The differences in the amount of pantomimes produced by children and adults can be seen as a reflection of task difficulty. It took children more gestures to be understood by their addressees, and some children explicitly reported that they found the pantomime task hard. It may be the case that children still need to learn to fully exploit the expressivity offered by the manual modality, and are unsure in selecting the features of the target referent that will be easiest to represent with the hands, and also best understood by their addressees. Concerning the production of speech-accompanying gestures, our study showed that older children gesture at lower rates than adults. This is consistent with previous research. For instance, in the context of a narrative task, Mayberry, Jaques and DeDe [20] compared the amount of words accompanied by gestures produced by stuttering and non-stuttering older children (mean age = 11) and adults. The results for the non-stuttering control group showed that adults accompanied their speech with gestures almost three times as much as children did. Using a narrative task, Alibali and colleagues [13] also found children to gesture less than adults, although these differences did not prove statistically significant. Altogether, it seems that older children gesture to a lesser extent than younger children [21] and also than adults, as shown by the present study and suggested by previous research [20]. This U-shaped pattern in gesture production may be an indicator of the ever-changing relationship between speech and gesture, with gesture production oscillating between higher and lower peaks until the relationship between the two modalities becomes fully consolidated, possibly in adolescence. In sum, we conjecture that, while younger children may use gestures as anchors to coordinate their representations, in late childhood gestures have become optional, and children do not fully regard their gestures yet as communicative devices that can be relied upon in order to communicate more efficiently.

Lastly, it is interesting to note that manipulability did not influence the amount of iconic gestures produced by speakers. Previous studies have shown that objects that are manipulable are more often gestured about than objects which are not, and this was found to be the case both for children [16] and for adults [14]. The explanation for this phenomenon is that manipulable objects may evoke action simulation, which could in turn prime gesturing in speakers (see [22]). In this study, we could have expected both groups to gesture more about manipulable objects in the speech condition. However, that is not what we found. One tentative explanation for this finding is that non-manipulable objects may have been harder to describe than manipulable objects, which could have increased the gesture rates for non-manipulable objects, to facilitate their description.

4.1. Representation techniques

Our study revealed different patterns regarding how older children and adults used gestural techniques to represent objects. Perhaps the most striking finding is that these differences were only found for spontaneous gesture production (recall Figure 3), indicating that in pantomime both adults and children represented objects similarly. Pantomimes, unlike other gesture types such as emblems (e.g., the thumbs-up sign), are not given by convention. However, a recent study by Van Nispen and colleagues [15] found regularities in the use of pantomimes (by adults) in the communication about objects, suggesting that speakers share to a certain extent similar mental representations. Our study extends these findings by showing that adults and children use representation techniques similarly when pantomiming. Furthermore, we observed the occurrence of combinatorial patterns in both groups. For instance, in depicting a sandpit, gesturers would typically produce a shape gesture (e.g., tracing the shape of the sandbox) followed by an action gesture (e.g., pretending to use a shovel). These examples highlight how, in the absence of speech, pantomimes begin to adopt consistent combinatorial patterns (e.g., first shape, then action), as suggested by [2], and also [23].

With respect to the age differences in spontaneous gesture production, we found that older children had a tendency towards producing more action gestures (whether transitive –object use- or intransitive –enactment-) than adults, who produced more shape gestures (in this case, molding gestures) than children. Furthermore, if we zoom into the techniques used by children (recall Figure 2), we can see that children produced twice as many object use gestures than portraying gestures, and in general twice as many action gestures than shape (molding or tracing) gestures. In adults, these differences were less pronounced. Therefore, although older children have left behind the phase where they represent objects and tools by using their own body as a cognitive anchor (as evidenced by younger children's preference for body-part-as-object gestures [5], [6], our results indicate that older children still have a preference for action-based [24] forms of iconicity, in contrast to perceptually-based [24] forms of iconicity, more present in the gestures produced by adults. This is interesting, if we consider how different gestural techniques vary in terms of their schematic complexity [24]. For instance, action gestures are closer to daily sensorimotor experience and seem relatively less schematic than perceptually-based shape gestures, which undergo a greater process of abstraction. Our results suggest that children are able to use more abstract representation techniques when gestures are consciously and deliberately produced (as is the case in pantomime) but perhaps they find action-based gestures easier to produce and therefore rely more on these when speech is the main communicative modality, and gesture production is optional.

The fact that there were no effects of modality on the representation techniques used is remarkable. Few studies have provided a systematic overview of the key differences between pantomiming and gesturing. Our study shows that whereas both forms of gesturing are non-conventionalized, speakers typically converge in the ways they gesture about objects, and come to use similar techniques. This could mean that both pantomimes and gestures, although constrained by language to different extents, emerge from the same representations. Thus, it could be the case that speakers (at least speakers who share the same language) have a natural tendency to converge in the iconic strategies they use to encode concepts into representational hand gestures, and that this process is free of the influence of (concurrent or not) speech.

4.2. Future research

The current study has a number of limitations. For instance, the scope of our analyses. While we were interested primarily in the gestural techniques that are used to convey meaning, there are other aspects of gesture production that are susceptible to the effects of age, and modality. For instance, we did not examine whether children used more whole-body gestures than adults (as suggested by McNeill [2]) or whether they tended to repeat the same gestures within one description, instead of combining different forms. Future studies could address these issues, to get a more complete picture of the development of the gestural system. Ultimately, the question should be asked whether these differences have an impact on how addressees interpret the meaning of utterances.

As for modality, we did not compare whether pantomimes were larger, or more precise, than gestures produced alongside speech, as one could expect if we take into account that, in pantomime, gestures are the sole vehicle for meaning expression, and their form may be enhanced for communicative purposes. Instead, we studied the techniques that speakers used to express information gesturally, and thus we can say something about the type of information that gestures conveyed, but we did not examine whether gestures and pantomimes really depicted the same, or different, features of objects. For instance, both a pantomime and a co-speech gesture might have outlined a shape for one particular object, but perhaps the shape depicted corresponds in each case to a different salient feature of the object. In future studies we plan to expand our dataset and look into these aspects.

5. Conclusions

In conclusion, this study showed a number of differences regarding how older children (aged 9) and adults use gestures in referential communication. When speaking was forbidden, and children could only rely on their hands to describe objects, they needed more gestures than their adult counterparts to complete the task. However, their gestures exhibited the same range of representation techniques to express meaning as adult gestures did. In contrast, when speaking was allowed, children relied less on gesturing than adults, and exhibited a bias towards producing action gestures, such as enactments, or imaginary object use gestures. Adults, in contrast, exhibited a wider range of techniques to help meaning come across, and relied on object use and shape gestures to a similar extent. This suggests that older children may already have all the tools needed for full gestural expressivity (as observed in the pantomime condition), but do not use them as smoothly as adults when speech and gestures are co-produced, indicating that both modalities haven't become fully integrated yet.

In addition to this, our study confirmed that, despite playing different (primary or secondary) communicative roles, co-speech gestures and pantomimes reflect similar aspects of the speakers' mental representations, and rely on the same techniques to encode information.

6. Acknowledgements

We thank Mirthe Treurniet for the help in collecting and annotating the children dataset. We received support for this research through a grant awarded by The Netherlands Organisation for Scientific Research (NWO) [grant number 322-89-010], which we gratefully acknowledge.

7. References

[1] Goldin-Meadow, S., "Widening the lens: what the manual modality reveals about language, learning and cognition", Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1651), 2014.

[2] McNeill, D., "Hand and mind: What gestures reveal about thought", Chicago: University of Chicago Press, 1992.

[3] Iverson, J. M., and Goldin-Meadow, S., "Gesture paves the way for language development", Psychological science, 16(5): 367-371, 2005.

[4] Gullberg, M., De Bot, K., and Volterra, V., "Gestures and some key issues in the study of language development." Gesture, 8(2):149-179, 2008.

[5] Overton, W. F., and Jackson, J. P., "The representation of imagined objects in action sequences: A developmental study", Child development, 44(2):309-314, 1973.

[6] Boyatzis, C. J., and Watson, M. W., "Preschool children's symbolic representation of objects through gestures", Child development, 64(3):729-735, 1993.

[7] O'Reilly, A. W., "Using representations: Comprehension and production of actions with imagined objects", Child development, 66(4):999-1010, 1995.

[8] Müller, C., "Iconicity and Gesture", in S. Santi, I. Guatiella, C. Cave and G. Konopczyncki [Eds.], Oralité et Gestualité. Montreal, Paris: L'Harmattan, 1998.

[9] Capirci, O., Contaldo, A., Caselli, M. C., and Volterra, V., "From action to language through gesture: A longitudinal perspective", Gesture, 5(1-2):155-177, 2005.

[10] McNeill, D., "Gesture and thought", Chicago: University of Chicago Press, 2008.

[11] Mayberry, R. I., and Nicoladis, E., "Gesture reflects language development evidence from bilingual children", Current Directions in Psychological Science, 9(6):192-196, 2000.

[12] Stefanini, S., Bello, A., Caselli, M. C., Iverson, J. M., and Volterra, V., "Co-speech gestures in a naming task: Developmental data", Language and cognitive processes, 24(2):168-189, 2009.

[13] Alibali, M. W., Evans, J. L., Hostetter, A. B., Ryan, K., and Mainela-Arnold, E., "Gesture–speech integration in narrative: Are children less redundant than adults?", Gesture, 9(3):290-311, 2009.

[14] Masson-Carro, I., Goudbeek, M. B., and Krahmer, E. J., "Can you handle this? The impact of object affordances on how co-speech gestures are produced", under review.

[15] Van Nispen, K., van de Sandt-Koenderman, M., Mol, L., and Krahmer, E., "Pantomime Strategies: On regularities in how people translate mental representations into the gesture modality", in P. Bello, M. Guarini, M. McShane, and B. Scassellati (Eds.), Proceedings of the 36th Annual Conference of the Cognitive Science Society, Austin, TX: Cognitive Science Society, 976-981, 2014.

[16] Huttunen, K. H., Pine, K. J., Thurnham, A. J., and Khan, C., "The Changing Role of Gesture in Linguistic Development: A Developmental Trajectory and a Cross-Cultural Comparison Between British and Finnish Children", Journal of psycholinguistic research 42(1):81-101, 2013.

[17] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H., "ELAN: a professional framework for multimodality research", in Proceedings of LREC, Fifth International Conference on Language Resources and Evaluation. Paris: ELRA, 2006.

[18] Bavelas, J. B., Chovil, N., Lawrie, D. A., and Wade, A., "Interactive gestures." Discourse processes, 15(4):469-489, 1992.

[19] Jaeger, T. F. "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models", Journal of memory and language, 59(4):434-446, 2008.

[20] Mayberry, R. I., Jaques, J., and DeDe, G., "What stuttering reveals about the development of the gesture-speech relationship", New Directions for Child and Adolescent Development, 79:77-88, 1998.

[21] Bello, A., Capirci, O., and Volterra, V., "Lexical production in children with Williams syndrome: Spontaneous use of gesture in a naming task", Neuropsychologia, 42(2):201-213, 2004.

[22] Hostetter, A. B., and Alibali, M. W.,"Visible embodiment: Gestures as simulated action", Psychonomic Bulletin and Review, 15:495–514, 2008.

[23] Goldin-Meadow, S., So, W. C., Özyürek, A., and Mylander, C., "The natural order of events: How speakers of different languages represent events nonverbally", Proceedings of the National Academy of Sciences, 105(27):9163-9168, 2008.

[24] Perniss, P., and Vigliocco, G., "The bridge of iconicity: from a world of experience to the experience of language." Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1651), 2014.

The Role of Left Inferior Frontal Gyrus in the Integration of Pointing Gestures and Speech

David Peeters [1], *Tineke M. Snijders* [2], *Peter Hagoort* [1, 2], *Aslı Özyürek* [1]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2] Radboud University, Donders Institute for Brain, Cognition, and Behaviour, Nijmegen, The Netherlands

David.Peeters@mpi.nl, Tineke.Snijders@let.ru.nl, Peter.Hagoort@mpi.nl, Asli.Ozyurek@mpi.nl

Abstract

Comprehension of pointing gestures is fundamental to human communication. However, the neural mechanisms that subserve the integration of pointing gestures and speech in visual contexts in comprehension are unclear. Here we present the results of an fMRI study in which participants watched images of an actor pointing at an object while they listened to her referential speech. The use of a mismatch paradigm revealed that the semantic unification of pointing gesture and speech in a triadic context recruits left inferior frontal gyrus. Complementing previous findings, this suggests that left inferior frontal gyrus semantically integrates information across modalities and semiotic domains.

Index Terms: pointing gesture, multimodal integration, reference, fMRI

1. Introduction

Pointing gestures are a fundamental part of human communication [1]. By producing them in everyday life we connect our communication to entities in the world around us [2]. In establishing a triadic link between child, caregiver, and referent, they play a crucial role in language acquisition [3] and impairments in the production and comprehension of pointing gestures are an early marker of the neurodevelopmental disorder autism [4]. From a phylogenetic viewpoint, it has been claimed that (declarative) pointing is a uniquely human form of communication in a natural environment [5].

Previous neuroimaging work investigating the comprehension of index-finger pointing gestures has presented the gestures in a context that lacked both a larger visual triadic context and co-occurring speech [6][7]. However, in everyday human referential communication pointing gestures often occur in a context in which one perceives not only the person pointing but also the referent she points at and the speech she may concomitantly produce. It is currently unclear how in such situations input from different modalities (visual: speaker, pointing gesture, referent; auditory: speech) is integrated in the brain. The lack of empirical neurocognitive research in this domain is surprising, because comprehending and integrating our interlocutors' referential (i.e. deictic) gesture and speech in a visual context is often critical to understand what they are talking about and a core feature of everyday communication [8]. The current study therefore investigates the neural mechanisms underlying the semantic integration of manual pointing gestures with speech in a visual, triadic context.

The majority of studies investigating the neural integration of gestures with co-occurring speech have focused on *iconic* co-speech gestures, i.e. hand movements that visually resemble the meaning of the linguistic part of the utterance they accompany [9]. It is relatively uncontroversial that LIFG, more specifically its pars triangularis, plays a role in the integration of speech and iconic gesture, possibly in interplay with MTG [10]. Willems et al. (2007) were the first to study the integration of speech and gesture using fMRI. In an orthogonal design, the ease of integration of linguistic and gestural information into a preceding sentence context was manipulated [11]. An increase in activation in LIFG was found when words and/or gestures were incongruent ("mismatch conditions") compared to when they were congruent ("match condition") with preceding speech. Such findings confirm LIFGs status as a multimodal integration site that plays a crucial role in the semantic unification of information from different modalities [12]. Such accounts argue, however, that LIFG is a node in a larger network that subserves the integration of gesture and speech, and also attribute a role to STS/STG and MTG in the perception and integration of speech-gesture combinations [10] [13].

As outlined above, in the current study we focus on a different type of gesture, namely (deictic) pointing gestures. Unlike iconic gestures, pointing gestures in exophoric use canonically create a vector towards a referent to shift the gaze of an addressee and establish a joint focus of attention [1]. Furthermore, whereas speech and iconic gestures often allow communicating about entities that are not immediately physically present ("displacement", [14]), pointing gestures in exophoric use play a crucial role in referential communication about entities that speaker and addressee may perceive in the immediate extra-linguistic context of a conversation. Therefore, the integration of speech and pointing gestures towards a referent need not necessarily recruit the same neural and cognitive mechanisms as in the integration of speech with iconic or other types of gestures.

Although it is currently unknown which cortical areas are involved in integrating pointing gestures and speech, a number of studies have looked at the neural correlates of comprehending pointing gestures in isolation and at their integration with other cues such as the gesturer's gaze direction. Sato et al. (2009), for instance, showed that the perception of a (meaningless) pointing hand, compared to a non-directional closed hand, elicits enhanced activation in a network of mainly right-hemisphere regions, including right IFG, right angular gyrus, right parietal lobule, right thalamus, and bilateral lingual gyri [7]. Materna et al. (2008) suggest that bilateral posterior STS is involved in following the direction of a pointing finger [6]. Conty et al. (2012) show that integration of pointing gestures and gaze direction in comprehension recruits parietal and supplementary motor cortices in the right hemisphere [15]. All in all, these findings suggest an extensive right-hemisphere dominant network that is activated when one perceives a manual pointing gesture that shifts one's attention.

Finally, Pierno et al. (2009) compared the observation of a static image of a hand pointing at an object to the observation of a hand grasping an object and to a control condition of a hand resting next to an object [16]. Compared to the control condition, the perception of the pointing hand and object elicited enhanced activation in left MTG, left parietal areas (postcentral gyrus and supramarginal gyrus) and left middle occipital gyrus. However, the pointing condition did not recruit significant differential activity compared to the grasping condition. Nevertheless these results suggest that, in addition to the right-lateralized network involved in perceiving a pointing hand, a left-lateralized set of cortical areas may be involved in visually integrating a pointing hand and an object.

1.1. The present study

In the present study, we investigated which cortical regions subserve the integration of pointing gestures with speech in a visual, everyday context. In an event-related functional magnetic resonance imaging (fMRI) study, participants were presented with images of a speaker who pointed at one of two different objects as they listened to her speech. We employed a mismatch paradigm, such that speech either referred to the object the speaker pointed at or to the other visible object. As such, speech and gesture were individually always correct, but there was congruence or incongruence when semantically integrated in the larger visual context. Thus, the match-mismatch comparison taps into the semantic integration/unification of pointing gestures and speech. Mismatch paradigms have been successfully used in the past to study the integration of iconic gestures and speech [13].

Because this is the first study investigating the neuronal integration of pointing gestures with speech in comprehension, predictions were derived on the basis of previous speech-gesture integration studies that used *iconic* gestures in their stimulus materials. If LIFG plays a key role in the semantic integration of gesture and speech [10] [13], it should show enhanced activation in the mismatch compared to the match condition. This is in line with a view of LIFG as a modality-independent multimodal integration site, with its pars triangularis specifically involved in semantic unification of information from different input streams [11] [12]. Conversely, if multimodal semantic integration of gesture and speech recruits the posterior part of the STS region [17], then this region should show enhanced activation in the mismatch-match comparison.

Finally, we included two conditions in which one of the two objects in the images was highlighted by an attentional cue in the absence of gesture. This allowed investigating whether the possible role of LIFG in semantic unification of speech and pointing gesture in a triadic context was dependent on the perceived communicative intentions of the gesturer. Research by Kelly and colleagues suggests that speech-gesture integration differs from the integration of gestures with actions more broadly because the former are generally viewed as more intended to accompany the speech signal compared to the latter [18]. Pointing gestures are shaped by the communicative intentions of the gesturer [19], and in that sense differ from other cues in the environment that may shift our attention. Therefore the integration of pointing gestures with speech may differ from the integration of other attentional cues with concurrently perceived speech. In sum, the current study thus aims to shed more light on the functional roles of different cortical areas involved in speech-gesture integration by investigating the integration of speech with a novel type of gesture, namely index-finger pointing.

2. Method

2.1. Participants

Twenty-three right-handed native speakers of Dutch (18 female; mean age 23.6, range 18-29) participated in the experiment. Data from three additional participants were discarded due to technical failure (n = 2) or drowsiness (n = 1). Participants had normal or corrected-to-normal vision, no language or hearing impairments or history of neurological disease. They provided written informed consent and were paid for participation.

2.2. Stimuli and Experimental Design

The experimental materials consisted of 40 spoken items in Dutch of the form “definite article + noun” (e.g., “het kopje”, *the cup*), 80 pictures in which a model (henceforth: the speaker) pointed (index-finger extended, [9]) at one of two objects presented at a table in front of her (henceforth “target pictures”), and 80 pictures that were the same except that one of the two objects was framed by a green box and that the speaker did not point (henceforth “attentional pictures”). The 40 spoken items were spoken at a normal rate by a female native speaker of Dutch, recorded in a sound proof booth, and digitized at a sample frequency of 44.1 kHz. They had an average duration of 837 ms (*SD* = 155 ms). In half of the target pictures the speaker pointed at the object at her left and in the other half of the target pictures she pointed at the object at her right. Similarly, in half of the attentional pictures the object at her left was framed and in the other half the object at her right. The 40 different table-top objects in the pictures were selected on the basis of a pre-test reported elsewhere [20] that confirmed that these objects elicited highly consistent labels (i.e. > 90% naming consistency for each object across 16 participants) across individuals from the same participant pool as the current participants.

The experiment consisted of three blocks. The *speech-only* block (AUDIO) consisted of the 40 spoken items. The *picture-only* block (VISUAL) consisted of 40 pictures in which the speaker pointed at an object. The *mixed block* consisted of 160 speech-picture pairs that made up four conditions. In the Bimodal Match (BM) condition, the spoken stimulus matched the object the speaker pointed at. In the Bimodal Mismatch (BMM) condition, the spoken stimulus did not match the object she pointed at but the other object. In the Attentional Match (AM) condition, the spoken stimulus matched the framed object. In the Attentional Mismatch (AMM) condition, the spoken stimulus matched the object that was not framed. Each condition consisted of 40 speech-picture pairs. The speech-only block and the picture-only block were included for a bimodal enhancement analysis that will be reported elsewhere. Figure 1 shows a subset of pictures used in the experiment.

2.3. Procedure

The three blocks were presented sequentially with specific instructions preceding each block. The order of presentation of the blocks was counterbalanced across participants. All stimuli were presented in an event-related design and in a randomized order. Twelve different randomized lists were used. The *speech-only block* consisted of the presentation of the 40 spoken stimuli. A trial in this block consisted of a fixation cross presented for a jittered duration of 2-6s followed by the presentation of the spoken stimulus. The *picture-only block* consisted of the presentation of 40 pictures in which the speaker pointed at one of the two objects. No speech was pre-

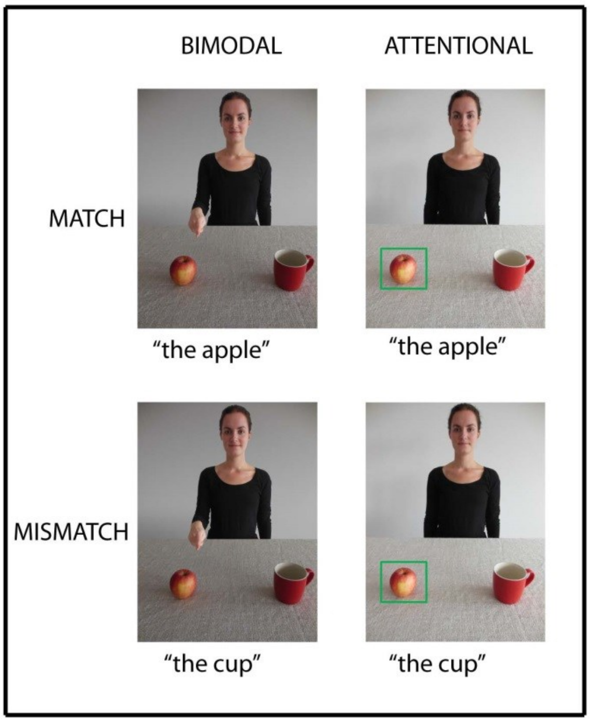


Figure 1: *Overview of the experimental conditions.*

sented during this block. A trial in this block consisted of a fixation cross presented for a jittered duration of 2-6s followed by the presentation of the picture for 2s. The *mixed block* consisted of 160 target trials in which a fixation cross (jittered duration of 2-6s) was followed by the presentation of a picture (for 2s) with a concurrently presented spoken stimulus. The onset of the spoken stimulus was 50 ms after the onset of the picture presentation. In both the picture-only block and the mixed block, the speaker pointed at the object at her left in half of the cases, and at the object at her right in the other half of the cases. In the mixed block, in half of the attentional pictures the object at the speaker's left was framed and in the other half of the attentional pictures the object at her right.

Pictures were presented on the screen using *Presentation* software (Neurobehavioral Systems) and speech was presented through nonmagnetic headphones that reduced scanner noise. Participants looked at the screen via a mirror mounted to the head coil. The size of the pictures on the screen was determined on the basis of judgments from two pilot subjects that did not participate in the main experiment. They confirmed that all objects, the speaker's gesture, and the attentional markers, were clearly visible while focusing on the center of the screen.

Participants in the main experiment were instructed to carefully listen to the speech and look at the pictures. They were asked to press a button with the middle finger of their left hand when an item (i.e. a spoken stimulus in the speech-only block, a picture in the picture-only block, and the picture-speech pair in the mixed block) was exactly the same on two subsequent trials. In the speech-only block and the picture-only block, four stimuli were repeated on two subsequent trials. In the mixed block 16 stimuli were repeated on two subsequent trials. The second presentations of such items thus served as catch trials eliciting a button press and were excluded from further MRI analyses. The experiment was preceded by a practice session.

2.4. fMRI data acquisition

Participants were scanned with a Siemens 3-T Skyra MRI scanner using a 32-channel head coil. The functional data were acquired in one run using a multiecho echo-planar imaging sequence, in which image acquisition happens at multiple echo times (TEs) following a single excitation [time repetition (TR) = 2250 ms; TE1 = 9 ms; TE2 = 19.5 ms; TE3 = 30 ms; TE4 = 40 ms; echo spacing = 0.51 ms; flip angle = 90 °]. This procedure broadens T2* coverage and improves T2* estimation. Each volume consisted of 36 slices of 3 mm thickness [ascending slice acquisition; voxel size = 3.3 x 3.3 x 3 mm; slice gap = 10 %; field of view (FOV) = 212 mm]. The first 30 volumes preceded the start of the presentation of the first stimulus and were used for weight calculation of each of the four echoes. Subsequently, the 31st volume was taken as the first volume in preprocessing. The functional run was followed by a whole-brain anatomical scan using a high resolution *T*1-weighted magnetization-prepared, rapid gradient echo sequence (MPRAGE) consisting of 192 sagittal slices (TR = 2300 ms; TE = 3.03 ms; FOV = 256 mm; voxel size = 1 x 1 x 1 mm) accelerated with GRAPPA parallel imaging.

2.5. fMRI data analysis

Data were analyzed using statistical parametric mapping (SPM8; www.fil.ion.ucl.ac.uk/spm/) implemented in Matlab (Mathworks Inc., Sherborn, MA, USA). The four echoes of each volume were combined to yield one volume per TR, after which standard pre-processing was performed [realignment to the first volume, slice acquisition time correction to time of acquisition of the middle slice, coregistration to T1 anatomical reference image, normalization to Montreal Neurological Institute (MNI) space (EPI template), smoothing with an 8 mm full-width at half-maximum (FWHM) Gaussian kernel, and high-pass filtering (time-constant = 128 s)].

Statistical analysis was performed in the context of the general linear model (GLM). Stimulus onset (i.e. the onset of the picture in all conditions, except the speech-only condition in which it was the onset of speech) was modeled as the event of interest for each condition. Each condition thus contained 40 events. The 6 condition regression parameters were convolved with a canonical hemodynamic response function. Additionally, 6 motion parameters from the realignment preprocessing step were included in the first-level model.

A whole-brain analysis was performed by entering first-level contrast images of each of the six conditions > baseline for each participant into a flexible factorial model at second-level [with factors Condition (6) and Participant (23)]. Two analyses were performed to compare semantic mismatch to semantic congruency. First, the bimodal mismatch condition was compared to the bimodal match condition (BMM > BM). Second, the attentional mismatch condition was compared to the attentional match condition (AMM > AM).

Whole-brain correction for multiple comparisons was applied by combining a significance level of p = 0.001 (uncorrected at the voxel level) with a cluster extent threshold using the theory of Gaussian random fields. All clusters are reported at an alpha level of $p < 0.05$ family-wise error (FWE) corrected across the whole brain.

We had the a priori hypothesis that LIFG would be recruited more in the BMM condition compared to the BM condition as this comparison arguably taps into semantic integration/unification of speech and gesture. However, it is unclear whether such a potential involvement of LIFG is specific to communicatively intended gestures and speech or, instead, generalizes to any semantic speech-referent relation as induced by an attentional cue (i.e. it would also show up in the AMM-AM comparison). Therefore, a region-of-interest (ROI) analysis was performed in LIFG. The ROI was an 8 mm sphere around centre voxels in LIFG taken from a meta-analysis on a large number of neuroimaging studies of semantic processing [13][21]. MNI coordinates were [-42 19 14]. Contrast estimates were calculated for each participant at first-level for the four conditions (AM, AMM, BM, BMM) using Marsbar (http://marsbar.sourceforge.net/).

3. Results

3.1. Behavioral performance

Participants detected 91.5 % of all catch trials. These data were not further analyzed.

3.2. Whole-brain analysis

We first compared the mismatch conditions to the match conditions at whole-brain level. Contrasting BMM with BM showed increased activations in left inferior frontal gyrus (Fig. 2 and Table 1). The reverse contrast (BM > BMM) did not show any significant cluster that survived the statistical threshold. Also contrasting AMM with AM did not show any areas that survived the statistical threshold (Table 1).

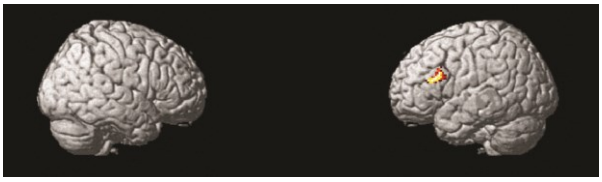


Figure 2: *Results from the whole brain analysis comparing Bimodal Mismatch (BMM) > Bimodal Match (BM). Results are displayed at p < .05, family-wise error corrected at the cluster-level.*

3.3. ROI analysis

An ROI analysis was performed comparing mismatch to match conditions in the predefined ROI (8 mm sphere around MNI coordinates -42 19 14) in LIFG. The interaction between cue (pointing gesture / attentional cue) and congruency (match / mismatch) failed to reach significance, F (1,22) = 2.10, p = .162. However, dependent samples t-tests revealed that there was enhanced activation in LIFG in mismatch vs. match conditions when the speaker's pointing gesture indicated the referent object, t(22) = -2.43, p = .024. There was no difference in activation in the ROI between the attentional mismatch and match conditions, t(22) = .48, p = .637. Figure 3 presents the contrast estimates for the four conditions.

4. Discussion

The present study investigated the neural integration of pointing gestures and speech in a visual, triadic context in comprehension. A mismatch analysis revealed that LIFG was sensitive to the congruence between speech and a concurrently presented pointing gesture towards a referent, whereas the posterior STS region was not.

Enhanced activation in LIFG has been found in previous studies that investigated the integration of iconic gestures with speech [10][11][13], pantomimes with speech [13], and metaphoric gestures with speech [22]. The common denominator in these studies is that an increase in semantic unification load led to an increase in LIFG activation [10]. For instance, gestures that are unrelated to concurrently presented speech require additional semantic processing because they are harder to semantically integrate with speech compared to iconic gestures that relate to the concurrently presented speech. Therefore, the former lead to enhanced LIFG activation compared to the latter [23]. The same holds for metaphoric co-speech gestures compared to iconic co-speech gestures [22]. Similarly, iconic gestures or pantomimes that are incongruent with speech activate LIFG more than iconic gestures and pantomimes that match the speech they accompany [11][13]. Confirming such previous findings, in the current study incongruence between speech and a visible object, as induced by a pointing gesture, led to enhanced activation in LIFG compared to a matched congruent condition.

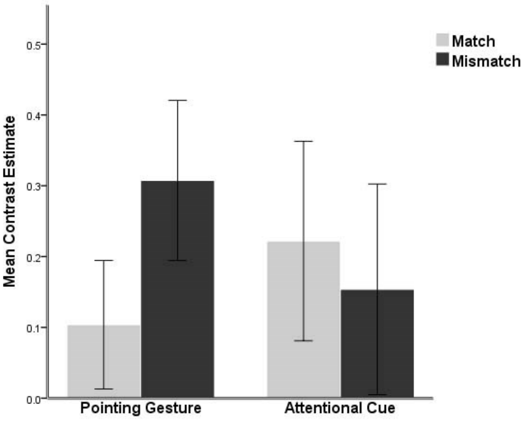


Figure 3: *ROI results. Mean contrast estimates for AM, AMM, BM, and BMM. Error bars represent standard errors around the mean.*

Previous studies have criticized the use of mismatch paradigms in gesture-speech integration studies, for instance arguing that "mismatches, which are rarely encountered in spontaneous discourse, may trigger additional integration processes which are not normally part of multimodal language comprehension"[17, p. 876], such that activations in LIFG may be a result of "the processing of unnatural stimuli and rather relate to error detection processes" [23, p. 3317]. There are convincing reasons to believe, however, that gesture-speech mismatch manipulations tap into semantic speech-gesture integration. For instance, LIFG activation is often also present in the "match" condition compared to baseline [13]. Furthermore, enhanced LIFG activation has also been found in speech-gesture integration studies that manipulated semantic load in a different way, not using a mismatch paradigm [10][24]. Dick et al. (2014), for instance, compared the integration of supplemental iconic gestures with speech to the integration of "redundant" iconic gestures with speech. The former gestures added information to the speech they accompanied (e.g. the verb in the phrase "Sparky attacked" was combined with a "peck" gesture) and therefore increased semantic processing and unification load compared to the latter gestures ("Sparky pecked" combined with a "peck" gesture). Indeed, a robust increase in activation was found in LIFG for the gestures that added information to the speech and therefore required additional semantic processing compared to the "redundant" gestures [10]. Crucially, both such gestures commonly occur in everyday interactions [9][25].

Table 1. *Results of the whole-brain analyses comparing congruent (match) to incongruent (mismatch) conditions.* p-*values are at the cluster-level, FWE-corrected.*

Contrast	*p*	k	*t*-value	MNI coordinates			Region/Peak
BMM - BM	.01	220	4.01	-46	20	20	LIFG (pars triangularis)
			3.72	-36	18	20	
			3.69	-50	28	18	
AMM - AM	-			-	-	-	

Abbreviations: AM, Attentional Match; AMM, Attentional Mismatch; BM, Bimodal Match; BMM, Bimodal Mismatch; k, extent (voxels).

LIFG plays a role not only in semantic unification of speech and gesture, but also in the semantic unification of word meaning and world knowledge into a preceding context in speech itself [26]. The current study extends previous work in showing that semantic unification recruits LIFG across semiotic domains. LIFG thus plays a crucial role in the case of an indexical semiotic relation between gesture, speech, and a referent (the current study), in addition to symbolic and iconic manners of signification (as in arbitrary word-meaning mappings and resemblance between iconic gestures/pantomimes/pictures and referents respectively). Furthermore, a core property of language (including iconic gestures) is that is allows for displacement, i.e. the ability to refer to entities that are not immediately present [14]. The current study shows that also when a referent is physically present in the immediate visual context, LIFG subserves the semantic unification of auditory and visual information at a higher-order semantic level. The involvement of LIFG in the case of pointing-speech integration may be dependent on whether transmitted information is semantic and/or communicatively intended, as it was not sensitive to the congruence between speech and an attentional cue around a visual object.

Finally, previous studies investigating the neural mechanisms involved in the perception of pointing gestures have focused on the gesture as a directional cue outside a speech context. Pierno et al. (2009), for instance, compared the observation of a static image of a hand pointing at an object to the observation of a hand grasping that object and to a control condition of a hand resting next to that object. Compared to the control condition, both types of actions activated a left-lateralized network that included parietal areas (postcentral gyrus and supramarginal gyrus) and left middle occipital gyrus [16]. Here we find that, when pointing gestures are produced with speech, LIFG is recruited and may be part of a larger network that comprises the areas found by Pierno et al. (2009). Furthermore, in that study no area was activated significantly more in the pointing condition compared to the grasping condition. Future work may therefore investigate whether the results of the current study generalize to situations in which a speaker grasps an object while concurrently producing speech. After all, in everyday life speakers may both point at an object and grasp and hold up or place an object to bring it into their addressee’s attention [2]. It is not unlikely that the extent of overlap between pointing-speech integration and grasping-speech integration might differ as a function of the perceived communicative intentions of the speaker (see [18]).

5. Conclusion

In sum, the current study investigated the neural integration of pointing gestures and speech in a visual, triadic context. We found that LIFG subserved the semantic unification of referential gesture and speech in a triadic context. This study can be informative as a starting point for studies investigating specific populations with impairments in the comprehension and integration of deictic speech and gesture and the subsequent establishment of joint attention in everyday life, as in autism spectrum disorders.

6. Acknowledgments

We would like to thank Paul Gaalman for assistance during the scanning sessions and Doris Deckers for help in creation of the stimuli.

7. References

[1] Kita, S. "Pointing. Where language, culture, and cognition meet", Hillsdale, NJ: Erlbaum, 2003.

[2] Clark, H. H., "Pointing and placing", In S. Kita [Ed], Pointing. Where language, culture, and cognition meet, 243-268, Hillsdale, NJ: Erlbaum, 2003.

[3] Carpenter, M., Nagell, K. and Tomasello, M., "Social cognition, joint attention, and communicative competence from 9 to 15 months of age", Monographs of the Society for Research in Child Development, 255: Vol. 63, 1-174, 1998.

[4] Baron-Cohen, S., "Perceptual role taking and protodeclarative pointing in autism", British Journal of Developmental Psychology, 7(2): 113-127, 1989.

[5] Call, J. and Tomasello, M., "Production and comprehension of referential pointing by orangutans (Pongo pygmaeus)", Journal of Comparative Psychology, 108(4): 307-317, 1994.

[6] Materna, S., Dicke, P. W. and Thier, P., "The posterior superior temporal sulcus is involved in social communication not specific for the eyes", Neuropsychologia, 46(11): 2759-2765, 2008.

[7] Sato, W., Kochiyama, T., Uono, S. and Yoshikawa, S., "Commonalities in the neural mechanisms underlying automatic attentional shifts by gaze, gestures, and symbols", NeuroImage, 45(3): 984-992, 2009.

[8] Bühler, K., "Sprachtheorie", Jena: Fischer, 1934.

[9] Kendon, A., "Gesture: Visible action as utterance", Cambridge: Cambridge University Press, 2004.

[10] Dick, A. S., Mok, E. H., Beharelle, A. R., Goldin-Meadow, S. and Small, S. L., "Frontal and temporal contributions to understanding the iconic co-speech gestures that accompany speech" Human brain mapping, 35: 900-917, 2014.

[11] Willems, R. M., Özyürek, A. and Hagoort, P., "When language meets action: the neural integration of gesture and speech", Cerebral Cortex, 17(10): 2322-2333, 2007.

[12] Hagoort, P., "On Broca, brain, and binding: a new framework", Trends in cognitive sciences, 9(9): 416-423, 2005.

[13] Willems, R. M., Özyürek, A. and Hagoort, P., "Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language", Neuroimage, 47(4): 1992-2004, 2009.

[14] Hockett, C. D., "The origin of speech", Scientific American, 203(3): 88-96, 1960.

[15] Conty, L., Dezecache, G., Hugueville, L. and Grèzes, J, "Early binding of gaze, gesture, and emotion: neural time course and correlates", The Journal of Neuroscience, 32(13): 4531-4539, 2012.

[16] Pierno, A. C., Tubaldi, F., Turella, L., Grossi, P., Barachino, L., Gallo, P. and Castiello, U., "Neurofunctional modulation of brain regions by the observation of pointing and grasping actions", Cerebral Cortex, 19(2): 367-374, 2009.

[17] Holle, H., Obleser, J., Rueschemeyer, S. A. and Gunter, T. C., "Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions", Neuroimage, 49(1): 875-884, 2010.

[18] Kelly, S., Healey, M., Özyürek, A. and Holler, J., "The processing of speech, gesture, and action during language comprehension", Psychonomic bulletin & review: 1-7, 2014.

[19] Peeters, D., Chu, M., Holler, J., Özyürek, A. and Hagoort, P., "Getting to the point: The influence of communicative intent on the kinematics of pointing gestures", In M. Knauff, M. Pauen, N. Sebanz and I. Wachsmuth [Eds], Proceedings of the 35th Annual Meeting of the Cognitive Science Society, 1127-1132, Austin, TX : Cognitive Science Society, 2013.

[20] Peeters, D., Hagoort, P. and Özyürek, A., "Electrophysiological evidence for the role of shared space in online comprehension of spatial demonstratives", Cognition, 136: 64-84, 2015.

[21] Vigneau, M., Beaucousin, V., Herve, P. Y., Duffau, H., Crivello, F., Houde, O., Mazoyer, B. and Tzourio-Mazoyer, N., "Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing", Neuroimage, 30(4): 1414-1432, 2006.

[22] Straube, B., Green, A., Bromberger, B. and Kircher, T., "The differentiation of iconic and metaphoric gestures: Common and unique integration processes", Human brain mapping, 32(4): 520-533, 2011.

[23] Green, A., Straube, B., Weis, S., Jansen, A., Willmes, K., Konrad, K. and Kircher, T., "Neural integration of iconic and unrelated coverbal gestures: a functional MRI study", Human brain mapping, 30(10): 3309-3324, 2009.

[24] Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C. and Small, S. L., "Speech-associated gestures, Broca's area, and the human mirror system", Brain and language, 101(3): 260-277, 2007.

[25] Holler, J. and Beattie, G., "Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener?", Gesture, 3(2): 127-154, 2003.

[26] Hagoort, P., Hald, L., Bastiaansen, M. and Petersson, K. M., "Integration of word meaning and world knowledge in language comprehension", Science, 304(5669): 438-441, 2004.

Aggressive Rhetoric: Prosodic and Kinetic Means

Nataliya Petlyuchenko [1], *Anna Artiukhova* [1]

[1] Chair of Germanic and Romanic Languages, National University "Odessa Law Academy", Odessa, Ukraine

npetljutschenko@onu.edu.ua, anna.art@rambler.ru

Abstract

This paper deals with the definition of aggressive rhetoric as the persuasive method in political communication, the transformation of the concept 'speech aggression' towards its positive semantic, the complex of aggressive speech means on the verbal and paraverbal levels. Verbal means of an aggressive rhetoric are the rhetorical figures, such as the antithesis that creates the greatest emotional stress due to its underlying semantic contrast. Paraverbal means of aggressive rhetoric are the prosody and co-speech gestures of the antitheses.

Index Terms: political communication, aggressive rhetoric, rhetorical figures, antithesis, verbal, prosodic, kinetic.

1. Introduction

This article attempts to define the concept of 'aggressive rhetoric' and describe its components on the verbal and paraverbal levels. The object of this study is German political discourse. The subject under analysis is the integral unit combining verbal, prosodic and kinetic means that add aggressiveness to political speech. The material for analysis was provided by the public speeches of Gregor Gysi and Joschka Fischer.

2. Aggressive rhetoric: the prosodic and kinetic aspect

Political communication is under the constant scrutiny of both ordinary citizens who themselves are parties to it, and researchers who study its mechanisms, types, and implementation methods. Political communication is of pronounced rhetorical nature. Rhetorical competence helps speakers convey their views to a wider audience, make contact, position themselves in a favorable light, convince the audience of the correctness of their views and encourage specific action. The more resolute, confident and aggressive a politician, the more persuasive his speech.

2.1. Verbal aggression

Initially verbal aggression was understood as a form of verbal behavior aimed at insulting or deliberately harming an individual or a group of people. It was accompanied by a highly emotional state of the speaker and the use of invective language. At the same time, it was noted that an essential feature of verbal aggression is its expressive and emotive coloration, which increases the persuasiveness of speech. Now researchers note the transformation of the concepts of 'aggression', 'aggressive' towards positive semantic content. Aggressiveness is becoming increasingly associated with persistent, ambitious, and charismatic. It should also be noted that these changes are caused by the propagandizing influence of the media, which specifically affect the formation of associative fields of various phenomena and provide them with the necessary focus. This means that the media tend to stereotype the word 'aggressive' and give its meaning a positive connotation. Thus, based on the current trend in interpreting verbal aggression, we can speak of aggressive rhetoric as the art of persistently, resolutely implementing speech impact in order to convince the public of the correctness of one's decisions and actions. It is particularly relevant in the context of political communication. Politicians should be aggressive as each party seeks to win. If a politician is not aggressive, he will simply be replaced by another one. What's more, the public likes aggressive politicians. Such politicians inspire admiration, trust, a sense of stability and security. Charismatic politicians have always been characterized by sharp statements, categorical views, and aggressive speech [1]. Aggressive rhetoric is inherent in politicians, whose position is contrary to the majority of the public, the opinion of their political allies, members of Parliament. We see its manifestations in acute crisis moments in political life.

2.2. Verbal means of aggressive rhetoric

Aggressive rhetoric is verbally implemented with rhetorical figures, which are markers of the *rhetorical force* of political speech.

As a stylistic device, antithesis is used to create contrasting characteristics of the described phenomenon and is widely used in the speeches of Gregor Gysi and Joschka Fischer. Antithesis is a rhetorical device in which an opposition or contrast of ideas is expressed [2]. Politicians use rhetorical figures in order to urge the audience adhere to his ideas and proceed to action. The antithesis is the prevalent rhetorical device, being meant to point to a strong conceptual opposition, with the ultimate goal to shock the audience. In this respect, the antithesis proves its rhetorical strength since it allows the orator the choice to point to those aspects that suit him best in positively and, respectively, negatively qualifying either term of the opposition. The force of the antithesis resides in the choice of the elements brought forth by the orator and in the way the latter constructs the oppositional relationship between them. Its persuasive effect is therefore measured in the “visibility” it provides a term of the opposition with, thus urging the audience to take action [3]

Being considered a brilliant figure of speech, the use of the antithesis is to be carefully pondered since, in case it is not artful, its effect fades away [4]. The lexical basis of this device is formed by antonyms, e.g.: *Und dann werden wir das Gegenteil von <u>Frieden</u> haben, sondern wir werden dauerhaft Instabilität, dauerhaft <u>Krieg</u>, dauerhaft Unterdrückung in dieser Region bekommen noch mit ganz anderen Konsequenzen (And then we will have the opposite of <u>peace</u> and we will get permanent instability, permanent <u>war</u>, permanent suppression in this region with absolutely different consequences).* [5]. This use of antithesis in the 'pure' form can be compared with the 'verbal game' used by a speaker to enhance the emotional and psychological impact of the opposition, e.g.: *Fragen Sie doch einmal einen Richter, ob ein Diebstahl <u>aus edlerem Motiv</u> im*

Vergleich zu einem Diebstahl aus unedlerem Motiv kein Diebstahl ist?(Just ask a judge whether a crime committed for a noble cause is not a crime if it is compared to a crime committed for a less noble cause) [6]. Often, instead of the classical antithesis built on the contrast of parallel structures and antonyms, speakers use emotional opposition. In this case, it is not antonyms in the proper sense that are opposed, but words, utterances, phrases, to which positive or negative appraisal is attributed in the context, e.g.: *Wir setzen darauf, und das, bitte ich euch, ist der Kern des Ganzen, nicht ob wir mit einem guten Gewissen nach Hause gehen, nicht ob wir uns mit Farbbeuteln beschmissen haben, sondern ob wir politische Entscheidungen treffen* (*We insist, and that, please note, is the core of everything, not that we go home with a clear conscience or paint bombs are thrown at us, but that we take political decisions)*[5]. This conceptual antithesis is used to highlight the importance of political decision-making, rather than protesting with cans of paint. We also see emotional contextual opposition in the antithesis used in G. Gysi's speach: *Die USA wollen mehr Einfluss gewinnen und vorhandenen verteidigen, und Russland will mehr Einfluss gewinnen und vorhandenen verteidigen (The US wants to gain more influence and defend the one it already has, and Russia wants to gain more influence and defend the one it already has)* [6]. The stress created by the opposition "*die USA*" and "*Russland*" is further reinforced through repetition. While evaluations are made by the speaker based on the arbitrary interpretation of the phenomena, realities, and facts. Emotional opposition serves to foreground substantive and axiological speech elements thus intensifying the effect of 'psychological pressure'. The use of emotional opposition, which we consider a kind of antithesis, contributes to establishing such a notional flow of the speech that does not allow the audience to make its own conclusions, since the candidates has already decided to foreground the axiological components. This speech structure allows 'imposing' one's views and expressing one's position on a particular issue, as well as establishing oneself as a 'reliable' politician, bearer of 'positive' qualities. In most cases, an antithesis is created not only by the semantics of the lexical units, but also the syntactic constructions [7]. Additional axiological and expressive functions of syntactic constructions were noted by T.A.van Dijk, who wrote that syntax reflects the distribution of the semantic roles of event participants: either through word order, or correlation of various functional elements (subject, object), or the use of active or passive forms, modality, modes [8]. In the antithesis *Ich bleibe aber der Meinung, dass die Abtrennung der Krim völkerrechtswidrig wäre, genauso wie die Abtrennung des Kosovo völkerrechtswidrig war (I remain of the opinion that the annexation of the Crimea would be contrary to international law, just as the secession of Kosovo was contrary to international law)* [6] the opposition is made on the level of changing mode and tense forms.

2.3. Paraverbal means of aggressive rhetoric

At the paraverbal level, the aggressiveness of rhetoric is created with certain phonatory and kinetic means. *Prosodic specifics* of political discourse are characterized by intensification of all its components (dynamic, tonal, and temporal). In experimental phonetics, this acoustic effect is referred to as 'prosodic intensity' [9], 'prosodic highlighting' (*prosodische Hervorhebung*) [10], 'prosodic emphasis' (*prosodische Emphase*) [11; 12]. This paper uses the term 'prosodic intensity' understanding it as abrupt changes in pitch, loudness, tempo variations, and pauses in certain speech sections. High prosodic intensity in certain sections as compared to others is an indicator of heightened emotionality, the speaker's involvement, an emphatic speech style [13].

Aggressive rhetoric is also formed by the *kinetic* (gesture) component that is in functional unity with the prosodic representation of speech making communication more effective. The gesture is the action or movement of the body through which one individual signals another individual about his presence, his intentions regarding objects [17]. Three main classes of *gestures* or *kinemes* can be singled out: (a) kinemes of independent lexical value capable of conveying meaning regardless of the verbal context, (b) co-speech kinemes accompanying verbal fragments, and (c) kinemes controlling the communicative process, i.e. establishing, maintaining and terminating communication [14]. The co-speech gestures of certain utterances are functionally deterministic, and the relationship between gesture and speech is of a twofold nature. Ensuring, on the one hand, the self-regulation of the communicative act, prominent (emphasizing) gestures accompany speech while simultaneously performing a communicative function; they are communicatively significant [15]. Prominent gestures accompany speech, so they are the markers of functional or meaningful components of spoken text and can serve as a tool for analyzing the structure of the text and its typological features [16].

Public political communications are characterized primarily by accentuating or illustrating gestures that represent movements of the body, especially the arms/hands, by which the speaker explains, complements his words, highlights the key points, emphasizes or amplifies a verbal utterance [17]. Aggressive rhetoric is also characterized by 'kinetic intensity'. Gestures make the speaker 'visible' increasing his image. The gesture is perceived by the addressee as a 'kinematic' form of verbal appeal through which he exercises his influence on his followers and/or opponents, encouraging them to actions aimed at achieving a particular goal.

2.4. The integral Verbal+Tone+Gesture Model of aggressive rhetoric

Consider the integrated model of aggressive rhetoric, which includes verbal, intonational and kinetic levels (verbal + ton + gesture / V+T+G). The antithesis of G.Gysi's speach *Entweder Zollunion mit Russland | oder Verträge mit uns! (*Either customs union with Russia | or contracts with us*)* [6] is characterized by particular prosodic and kinetic emphasis. The opposed parts of the antithesis are divided by a pause lasting 393 msec (see Figure 1), while in the beginning of the speech figure there is a sharp pitch increase F_{max} up to 400 Hz F_{min} 100 Hz and F_{mean} 240 Hz. This figure is also characterized by high intensity I_{mean} 70 dB, and an increase on the word *entweder* up to I_{max} 79 dB. With regard to kinetic emphasis, in the first part of the antithesis in the word *entweder* the main pitch accent [ʺɛnt], which is generally unstressed in German, is accompanied by an o-form gesture of the left hand; in the second part of the figure the main pitch accent [oː] is highlighted by a gesture of the right hand with a raised index finger (index finger gesture). Thus, we also observe the contrast characteristic of an antithesis on the gesture level using oppositional gestures (left : right, o-form : index finger). It is also worth noting that the phrase components containing information about Russia and the Customs Union are accompanied by an o-form gesture of the left hand, while the utterance about the EU agreement – by an index finger gesture of the right hand, which can be considered as approval.

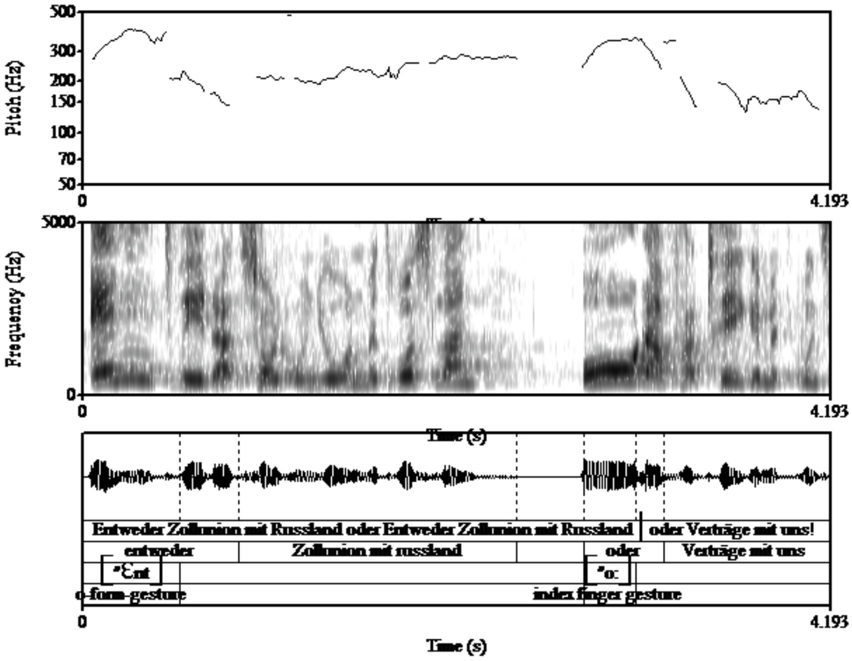


Figure1: *Antithesis example from G.Gysi's speech.*

Consider the integrated model of aggressive rhetoric on the example of the antithesis in J. Fischer's speech *Milosevic würde dann nur gestärkt und nicht geschwächt (Milosevic would then only be strengthened and not weakened)* [5]. This phrase (see Figure 2) is characterized by high volume and tone. I_{max} – 91 dB with I_{mean} – 86 DB, and F_{max} – 286 Hz with F_{min} – 164 Hz and F_{mean} – 244 Hz. The prosodic intensity of the figure is complemented by the kinetic intensity in the prominent parts of the speech. For instance, components of the antithesis *gestärkt*, main pitch accent [ʃtɛrkt] and *geschwächt*, main pitch accent [ʃvɛçt] are accompanied by o-form gestures of the left hand.

Thus, our figures allow some interesting observations: as we can see from the peak pitch and intensity values, emphasis is put on the key words of the antithesis *gestärkt* (strong) and *geschwächt* (weak), which are accompanied by gestures. At the same time, words that are semantically dependent, in our case components of the double conjunction *entweder* (either) … *oder* (or), are also highlighted. This shows that emphasis in this

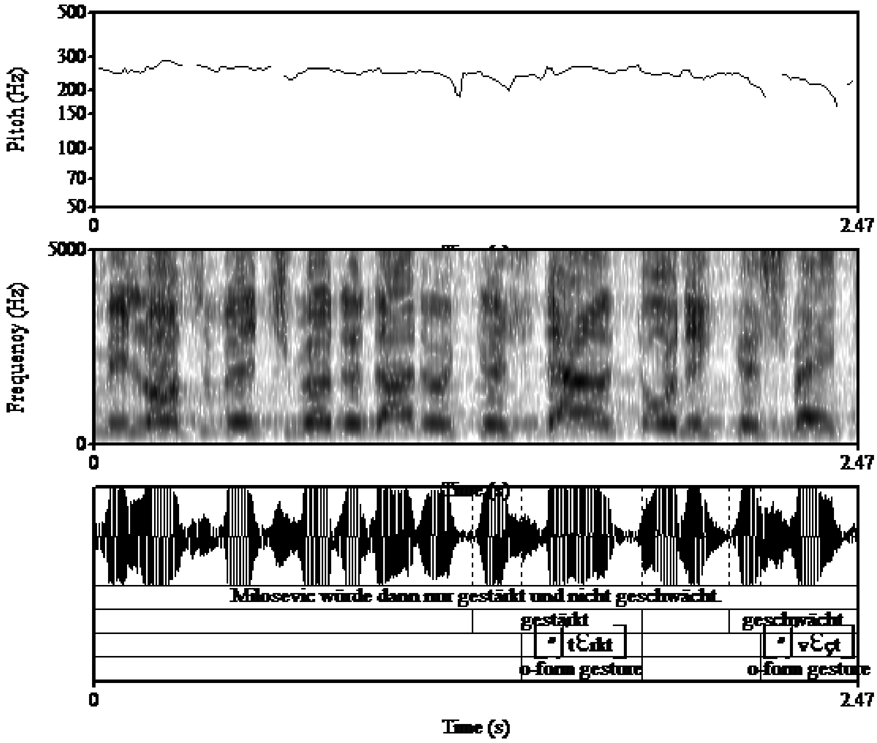


Figure 2: Antithesis example from J.Fischer's speech.

case is not content-related but its behavior is independent of this structure. This serves to justify the independence of what we call the aggressive rhetoric.

3. Conclusions

Thus, aggressive rhetoric is expressed in political communication in times of crisis and is characteristic of politicians, whose stand is not consistent with the majority position. This aggressiveness is created by using rhetorical figures, which are markers of the rhetorical force of speech, and their prosodic and gesture emphasis, which can be either content-related or not, indicating the independence of aggressive rhetoric. In addition, we observe the implementation of an integrated model of aggressive rhetoric on the verbal, prosodic and kinetic levels of speech.

4. References

[1] Ditko, H. Peter "Redner für die Freiheit", Norderstedt: Books on demand GmbH, 2009.

[2] Concise Oxford English Dictionary: Luxury Edition. Angus Stevenson, Maurice Waite [Ed], Oxford University Press, 2011.

[3] Zalenska, M. "Rhetoric and Politics: Central/eastern European Perspectives", Cambridge Scholars Publishing, 2012.

[4] Fontanier, P. "Les figures du discourse", Paris: Flammarion, 1977.

[5] Die Rede des Außenministers Joschka Fischer zum NATO-Einsatz im Kosovo, in: Heinrich Böll Stiftung, Archiv Grünes Gedächtnis, Hannover, 1999.

[6] Die Rede von Gregor Gysi im Bundestag am 13.03.2014 «Ukraine - Es gibt nur den Weg der Diplomatie. Antwort auf die Regierungserklärung der Bundeskanzlerin zur Situation in der Ukraine» Online: http://www.linksfraktion.de/reden/ukraine-es-gibt-weg-diplomatie, accessed on 11 Apr 2015.

[7] Plaksina, E. B. "Antiteza v zagolovkakh statey rossiyskoy i frantsuzskoy pressy", Dis. … kand. filol. nauk.: 10.02.20, Ekaterinburg, 2007.

[8] van Dijk, Teun A. "Discourse and Communication: New Approaches to the Analysis of Mass Media Discourse and Communication", New York: Walter de Gruyter, 1985.

[9] Petlyuchenko, N. V. "Kharizmatika: movna osobistist' i diskurs, Odesa: Astroprint, 2009.

[10] Essen, O. von. "Allgemeine und angewandte Phonetik", Berlin, 1966.

[11] Beck, H. R. "Politische Rede als Interaktionsgefüge: der Fall Hitler",Tübingen: Max Niemeyer Verlag, 2001.

[12] Selting, M. "Prosodie im Gespräch",Tübingen: Max Niemeyer Verlag, 1995.

[13] Kranich, W. "Phonetische Untersuchungen zur Prosodie emotionaler Sprechausdruckweisen", Frankfurt am Main: Lang, 2003.

[14] Krejdlin, G. E. "Muzhchiny i zhenshhiny v neverbal'noj kommunikacii", M.: Jazyki slavjan. kul'tury, 2005.

[15] Kendon, A. "Gesture: Visible Action as Utterance", Cambridge: University Press, 2004.

[16] Belikov, A. P. "Funkcional'noe vzaimodejstvie rechi i zhesta", avtoref. dis. ... kand. filol. nauk : 10.02.19, Odes. gos. un-t im. I.I. Mechnikova, Odessa, 1991.

[17] Cozzolino, M. "La comunicazzione invisible. Gli aspetti non verbali della comunicayione', Ediyioni Carlo Amore, 2003

Turn projection deficits in aphasic patients: Evidence from eye tracking

Basil C. Preisig [1], Noëmi Eggenberger [1], Giuseppe Zito [3], and René M. Müri [1,3,4,5]

[1] Departments of Neurology and Clinical Research, University Hospital Bern, Switzerland
[2] University Hospital of Old Age Psychiatry, University of Bern, Switzerland
[3] Gerontechnology and Rehabilitation Group, University of Bern, Switzerland
[4] Division of Cognitive and Restorative Neurology, Department of Neurology, University Hospital Bern, Switzerland
[5] Center for Cognition, Learning and Memory (CCLM), University of Bern, Switzerland

basil.preisig@dkf.unibe.ch, noemi.eggenberger@dkf.unibe.ch,
giuseppe.zito@ARTORG.unibe.ch, rene.mueri@insel.ch

Abstract

Humans are highly competent at managing the exchange of speaking turns during face-to-face interaction. Therefore, inter-speaker gaps and overlaps are minimal during conversation. Previous research showed that healthy subjects are able to predict the end of a turn and that this ability mainly relies on the recognition of the linguistic units within the conversation. In the present study, we compared the timing of transition related gaze shifts in aphasic patients with healthy controls while they watched video vignettes of natural conversations. Our results suggest that healthy controls adapt their gaze shift behavior depending on the dynamics of the turn transition in the video. They anticipate more dynamic turn transitions with overlapping speech that require faster gaze shifting and they anticipate less transitions with inter-speaker gaps. In contrast, aphasic patients did neither anticipate dynamic nor less dynamic turn transitions. These findings suggest that aphasic patients with acquired language comprehension deficits also have troubles with turn projection.

Index Terms: turn taking, aphasia, turn projection, eye tracking, eye gaze

1. Introduction

Sacks [1] suggested that people follow a basic set of rules (e. g. only one speaker at a time) in order to coordinate the exchange of speaking turns throughout conversation. He further argued that gaps and overlaps during conversation are only minimal because the conversation partners recognize the linguistic structure of a turn and thus can project where it ends.

Recent studies investigated turn processing in healthy subjects by tracking eye movements from a third person perspective [2-7]. In this experimental paradigm, a passive observer watches a pre-recorded dialogue episode on video while his or their eye movements are recorded. In this paradigm, the focus of the analysis lies on the timing of gaze shifts in relation to the turn transitions in the video. Indeed, the majority of the studies suggests that healthy subjects anticipate the end of the turn in the videos [2, 3, 6]. This means that they shift their gaze from the current to the next speaker in a time window between 500 ms prior to the end of the current turn and the first 200 ms after the beginning of the next turn. Westheimer [8] reported an average oculomotor reaction time of 150 ms, therefore it is assumed that gaze shift reactions within the first 200 ms are anticipatory [9]. To the best of our knowledge, turn projection has not been studied in aphasic patients. It has been shown in healthy subjects that the lexico-syntactic content of the conversation is essential for turn projection [10]. Thus, it could be assumed that aphasic patients have difficulties in turn projection because the lexico-syntactic content is not fully accessible for them due to deficits in semantic and syntactical processing [11].

In the present study, we investigated turn projection in aphasic patients in a third person perspective eye tracking paradigm. For data analysis, we distinguished two types of turn transitions: inter-speaker gap and inter-speaker overlap. Turns transitions with overlapping speech are more dynamic situations compared to transitions with inter-speaker gap because they require faster gaze shifting to keep up with the conversation. We hypothesized that healthy subjects are more likely inclined to show anticipatory gaze shifting on more dynamic turn transitions with overlapping speech compared to more static transitions with inter-speaker gap. In contrast, we expected that aphasic patients would not be able to adapt their reaction on inter-speaker overlaps and thus exhibit difficulties to project the end of the turn. Furthermore, we aimed to detect lesion sites that could be related to deficient turn projection in aphasic patients.

2. Material & Methods

2.1. Subjects

Eight aphasic patients with aphasia after first-ever left hemispheric stroke (mean age = 47.15 ± 9.88 years, 2 females) and 12 healthy controls (mean age = 49.13 ± 4.73 years, 5 females) were included in the study. All subjects had normal or corrected to normal visual acuity and an intact central visual field of 30°. Aphasia diagnosis was based on neurological examination and standardized diagnostic procedure conducted by professional speech-language therapists [12]. For all subjects, informed consent was obtained prior to study participation. The study was approved by the local ethical committee and conducted according to the latest guidelines of the Declaration of Helsinki.

2.2. Procedure

Subjects were seated comfortably in front of a 22” monitor with a remote 250 Hz infrared eye tracker (SensoMotoric Instruments GmbH, Teltow, Germany) at an operating distance of 60 to 80 cm. After a practice trial, four experimental trials consisting of four different videos with a length of 2 minutes each were presented. The videos depicted spontaneous unscripted dialogue between a female and a male actor. The topics were daily issues (food preferences, habitation, clothes, and sports) with different actors in each video.

2.3. Data analysis

After pre-processing with the SMI analysis software (BeGaze^{TM} SensoMotoric Instruments GmbH, Teltow, Germany) saccadic data was further processed with Matlab R2012b (Mathworks Inc., Natick MA). Direct gaze shifts (i.e. saccades) from the face area of one actor to the face area of the other actor were included in the analysis. In order to test the presence of turn projection, the analysis focused on gaze shifts near turn transitions in the video (inter-speaker gaps and inter-speaker overlaps). These events were defined through acoustic analysis and classified according to Heldner and Edlund [13]. First, gaze shifts were assigned to the nearest turn transition (inter-speaker gap and inter-speaker overlap) in time. Then, the mean gaze shift reaction time was calculated by subtracting the start of the gaze shift from the start of the assigned event. Thus, a negative value would indicate that the start of the gaze shift preceded the start of the assigned event, and vice versa.

Statistical analyses were conducted with IBM Statistics SPSS 21. We decided to use non-parametric tests given the small sample size and the violation of normality. A Friedman ANOVA was calculated to test the within-subject factor type of turn transition (inter-speaker gap and inter-speaker overlap). Consecutively, two-separate Wilcoxon tests were used for the *post hoc* analysis in each group (aphasic patients and healthy controls). Non-parametric correlations (Spearman's rank correlations) were calculated between Token Test scores as an index of aphasia severity and mean gaze shift reaction times. The level of significance was set to $p < .05$ (two-tailed). All values are expressed as mean ± standard error of the mean (SEM).

Lesion analysis of imaging data was conducted in MRICron [14]. The boundary of the lesions were delineated directly on the individual MRI image and then converted into the Talairach space using the spatial normalization algorithm provided by SPM5 (http://www.fil.ion.ucl.ac.uk/spm/). Further, lesions of patients with normal task performance comparable to healthy controls were contrasted with the lesions of patients with deficient turn projection. On this account, a conventional lesion subtraction analysis was conducted. This method is suitable for the descriptive analysis in small patient samples.

3. Results

As expected, the Friedman ANONA revealed a significant effect of transition type ($\chi^2(1) = 5.000$, $p = .025$). In general, subjects showed faster mean gaze shift reaction on inter-speaker overlaps compared to inter-speaker gaps (Fig. 1). Wilcoxon tests conducted for post hoc analyses within each experimental group revealed that healthy controls showed faster gaze shift reactions on transitions with overlapping speech ($z = -2.118$, $p = 0.034$, $r = -.43$). In contrast, aphasic patients showed no difference on their gaze shift reaction between transitions with inter-speaker overlap and inter-speaker gap ($z = -1.400$, $p = 0.161$, $r = -.35$).

The correlation analyses between Token Test scores and mean gaze shift reactions on the two types of turn transition revealed no significant results.

The lesion subtraction analysis between aphasic patients who showed comparable task performance as heathy controls and patients with deficits in turn projection indicates that delayed gaze shift reactions on inter-speaker overlap might be associated with lesions in the superior temporal lobe (Fig. 2).

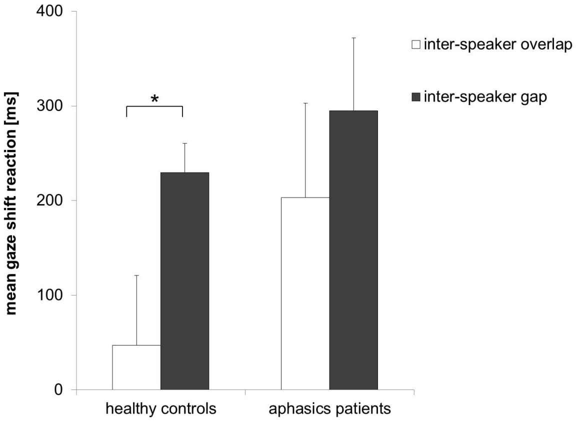


Figure 1: Mean gaze shift reaction on turn transitions with inter-speaker overlap and inter-speaker gap. The Asterisk indicates the level of significance (p < .05)*

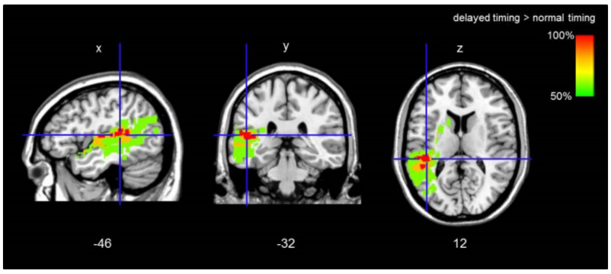


Figure 2: Illustration of the lesion subtraction analysis of patients with normal reaction on inter-speaker overlap vs. patients with delayed response times. The bar on the right indicates that (50-100%) of the patients with deficient turn projection have a lesion in this area, but not patients who performed normally.

4. Discussion

The present study investigated turn projection in aphasic patients with a third person perspective eye tracking paradigm. The main finding is that aphasic patients did not adapt the timing of gaze shifts from the current to the next speaker depending on the dynamics of the turn transition, while healthy controls reacted faster on transitions with overlapping speech compared to transitions with inter-speaker gap. On both types of transition, aphasic patients showed mean gaze shift reactions above 200 ms, which is over the average oculomotor reaction time of 150 ms [8]. This result suggests that aphasic patients did not anticipate the end of the turn for the majority of turn transitions. In contrast, healthy controls seemed to adapt their gaze shift strategy depending on the dynamics of the turn transition in the video showing anticipatory gaze shifting for turn transitions with overlapping speech. When the current turn was followed by an inter-speaker gap, they took more time to shift their gaze to the next speaker compared to situations where the next speaker already started to speak before the current speaker ended their turn. It seems as if healthy subjects would only show anticipatory gaze shift behavior when it is necessary. This might explain why other studies [4, 5] did not find turn anticipation in healthy subjects.

Our results indicate that aphasic patients have difficulties in turn projection. Possible consequences could be that aphasic patients react slower or even miss turns during face-to-face interaction with healthy speakers and are also less involved in multi-party interaction. Furthermore, this deficit could also hinder conversations over the phone, where nonverbal cues are missing. Nevertheless, aphasic patients still reliably detected turns and followed the current speaker with their gaze, despite their slowed reaction on turn transitions. These findings are in line with previous research which suggested that the fundamental communicative competences to establish turn taking behavior are preserved in aphasia [15, 16].

From the lesion subtraction analysis we gained first insights into the neural underpinnings of the turn projection deficits in aphasic patients. The contrast of the lesions from patients with delayed gaze shift reactions compared with the lesions from patients with normal task performance suggests that brain lesions in the superior temporal lobe could be predictive for the observed deficits. Previous research suggests that this area is part of a network which is involved in syntactic and semantic processing [17-19]. Aphasic patients with brain lesion to this area might thus encounter deficits on both, lexical retrieval and syntactical structure building. This might explain why aphasic patients with lesions in this area have more difficulties to recognize the structure of linguistic units during conversation and then cannot reliably project the end of the turn. Moreover, there is evidence that this area is not only involved in linguistic processing but also in social cognition. Ramnani and Miall [20] found brain activity in this area when subjects had to predict the actions of others. This is interesting because deficient turn projection in aphasic patients might not only be due to slowed syntactic [21] and lexical activation [22], or impaired lexical integration [23], but further rely on the ability to perceive social intentions. In the present, preliminary study, we tested only a small sample of patients. Therefore, further studies should target this aspect in larger samples that would allow also statistical lesion analysis.

Future research should also target turn projection in aphasic patients in real interactions. The results from a recent eye tracking study in healthy subjects during face-to face interaction [9] confirmed the evidence from previous studies that used the third person perspective eye tracking paradigm [2, 3, 6]. Holler and Kendrick [9] analyzed the turn projection of unaddressed participants that were participating in a free triadic conversation. Their results suggested that healthy subjects indeed anticipated the end of the current turn.

5. Acknowledgements

This study was supported by the Swiss National Science Foundation [Grant no. 320030_138532/1].

6. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language,* vol. 50, pp. 696-735, 1974.

[2] M. Tice and T. Henetz, "Turn-boundary projection: Looking ahead," in *The 33rd Annual Meeting of the Cognitive Science Society [CogSci 2011]*, 2011, pp. 838-843.

[3] M. Casillas and M. C. Frank, "Cues to turn boundary prediction in adults and preschoolers," in *SemDial 2012 (SeineDial)*, 2012, pp. 61-69.

[4] J. Edlund, S. Alexandersson, J. Beskow, L. Gustavsson, M. Heldner, A. Hjalmarsson*, et al.*, "3rd party observer gaze as a continuous measure of dialogue flow," 2012.

[5] L. Hirvenkari, J. Ruusuvuori, V. M. Saarinen, M. Kivioja, A. Perakyla, and R. Hari, "Influence of Turn-Taking in a Two-Person Conversation on the Gaze of a Viewer," *Plos One,* vol. 8, Aug 2013.

[6] A. Keitel, W. Prinz, A. D. Friederici, C. v. Hofsten, and M. M. Daum, "Perception of conversations: The importance of semantics and intonation in children's development," *Journal of Experimental Child Psychology,* vol. 116, pp. 264-277, 10// 2013.

[7] C. von Hofsten, H. Uhlig, M. Adell, and O. Kochukhova, "How children with autism look at events," *Research in Autism Spectrum Disorders,* vol. 3, pp. 556-569, Apr-Jun 2009.

[8] G. Westheimer, "Eye movement responses to a horizontally moving visual stimulus," *AMA Archives of Ophthalmology,* vol. 52, pp. 932-941, 1954.

[9] J. Holler and K. H. Kendrick, "Unaddressed participants' gaze in multi-person interaction: optimizing recipiency," *Frontiers in Psychology,* vol. 6, p. 14, Feb 2015.

[10] J. P. De Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation," *Language,* pp. 515-535, 2006.

[11] D. Caplan, G. Waters, G. Dede, J. Michaud, and A. Reddy, "A study of syntactic processing in aphasia I: Behavioral (psycholinguistic) aspects," *Brain and Language,* vol. 101, pp. 103-150, May 2007.

[12] W. Huber, K. Poeck, and K. Willmes, "The Aachen Aphasia Test," *Adv Neurol,* vol. 42, pp. 291-303, 1984.

[13] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics,* vol. 38, pp. 555-568, 10// 2010.

[14] C. Rorden, H. O. Karnath, and L. Bonilha, "Improving lesion-symptom mapping," *Journal of Cognitive Neuroscience,* vol. 19, pp. 1081-1088, Jul 2007.

[15] S. Schienberg and A. Holland, "Conversational turn-taking in Wernike aphasia," 1980.

[16] H. K. Ulatowska, L. Allard, B. A. Reyes, J. Ford, and S. Chapman, "Conversational discourse in aphasia," *Aphasiology,* vol. 6, pp. 325-330, 1992/05/01 1992.

[17] M. Vigneau, V. Beaucousin, P. Y. Hervé, H. Duffau, F. Crivello, O. Houdé, *et al.*, "Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing," *NeuroImage,* vol. 30, pp. 1414-1432, 5/1/ 2006.

[18] E. Jefferies and M. A. L. Ralph, "Semantic impairment in stroke aphasia versus semantic dementia: a case-series comparison," *Brain,* vol. 129, pp. 2132-2147, Aug 2006.

[19] A. D. Friederici, S. A. Kotz, S. K. Scott, and J. Obleser, "Disentangling Syntax and Intelligibility in Auditory Language Comprehension," *Human Brain Mapping,* vol. 31, pp. 448-457, Mar 2010.

[20] N. Ramnani and R. C. Miall, "A system in the human brain for predicting the actions of others," *Nature Neuroscience,* vol. 7, pp. 85-90, Jan 2004.

[21] P. Burkhardt, M. M. Pinango, and K. Wong, "The role of the anterior left hemisphere in real-time sentence comprehension: Evidence from split intransitivity," *Brain and Language,* vol. 86, pp. 9-22, Jul 2003.

[22] T. Love, D. Swinney, M. Walenski, and E. Zurif, "How left inferior frontal cortex participates in syntactic

processing: Evidence from aphasia," *Brain and Language,* vol. 107, pp. 203-219, Dec 2008.

[23] J. J. Choy and C. K. Thompson, "Binding in agrammatic aphasia: Processing to comprehension," *Aphasiology,* vol. 24, pp. 551-579, 2010.

How sound moves the body: Felt qualities of experience in gestural enactments of film sound

Linn-Marlen Rekittke [1], Dhana Wolf [1,2], Irene Mittelberg [1]

[1] Human Technology Centre, RWTH Aachen University, Aachen, Germany
[2] Department of Psychiatry, Psychotherapy and Psychosomatics, Medical School,
RWTH Aachen University
rekittke@humtec.rwth-aachen.de, wolf@humtec.rwth-aachen.de, mittelberg@humtec.rwth-aachen.de

Abstract

The approach presented in this paper aims to contribute to an embodied account of sound experience and multimodal expression. Here, gestures are assumed to be dynamically constructed and adaptive to the sounds perceived and described by their producers. Based on film retellings by 11 German native speakers, gestural re/presentations of sound experience were analyzed in terms of spatial and temporal features with the help of a motion-capture system. When describing sounds stemming from a cinematic context, study participants conveyed sound qualities and re-experienced sensations through co-speech gestures. Our findings put into relief ways in which felt qualities of this experience are expressed through gestural enactment.

Index Terms: gesture, sound, embodiment, film

1. Introduction

This study investigates how impressions of film sound and the sensations they evoke in the spectators translate into communicative gestural movements during multimodal film retellings. We are particularly interested in how speaker-gesturers relate and enact their subjective sensorial experience of being exposed to a movie in which sounds take center stage. The effects sound may have on the perceiver's emotional state have been described by previous research [1]. Smalley [2], for instance, reminds us of the link between music and emotional responses. When listening to music we sense a psychological quality reflecting qualities of the sound; the effect of the sound is psychophysical information.

In a first approach, it is necessary to distinguish basic emotions such as fear and happiness [3] from less clearly defined sensations, which are the main interest of the present paper. According to various psychological approaches [4], emotions generally comprise the following components, which may, but do not have to occur simultaneously: 1) subjective, 2) physiological, 3) behavioral (gesture and facial expressions), and 4) cognitive (evaluation). The subjective component is defined as a private experience that can only be described through words. Contradicting the assumption of emotions as being 'in-coming' sensations, Zlatev [5] argues that such an account 1) lacks vividness and does not connect emotion and self-motion and 2) is insufficiently intersubjective. Affectivity motivates the body to move, but by the same token, it may be expressed through bodily movement [5]. As gesture research from several disciplines suggests [6], [7] and as we hope to show in this paper, gestures are not only part of the behavioral component, but tend to also unite cognitive, subjective and other dimensions of emotions.

Besides spoken language, body movements and facial expressions, spontaneous manual co-speech gestures may indeed express emotional and mental states [8]. The interrelation of the cause, experiencer and effect of feelings and emotions is very complex [9]; gestures seem particularly apt at encompassing and expressing multiple dimensions due to their visio-spatial mediality and low codification. Besides the quality of the voice and intonation, spoken language alone can hardly convey subjective experiences holistically. Bodily semiotics, such as co-speech gestures, typically contribute additional dimensions. Furthermore, the subjective and physiological components may intermingle and as such are expressed – and also evoked – by gestures.

In addition, the cognitive component of emotions may be induced by manual gestures. Horst and colleagues [10] theoretically and empirically elaborate on how interlocutors share emotional experiences through their moving hands. This analysis of interaction leads to an embodied account of sense making through expressive movements. It challenges the assumption of pure conceptual and interpretative processes of understanding, suggesting dynamic sense-making processes through the perceiving, expressing and moving body.

Müller [8] has identified the three speech-functions of Bühler's organon model [11] interacting in co-speech gestures: representation, appeal, and expression. A gesture may represent an entity or event, impact the behavior of the interlocutor, and express feelings or affective stance. In this study the expressive function of gesture is focused upon, in the sense that the *how* rather than the *what* is foregrounded in the gesture [10]. In the following, special attention will be paid to the qualities of the moving hands and the manner of the movement (e.g. velocity, tension and torsion). In the sequences discussed below, the hands' movement qualities express sensations rather than emotions [12].

In line with embodied approaches [13], [14], this paper explores the perception of film sound and its interpretation in retellings. Scherer and colleagues [15] argue that films literally move the spectator. In addition to camera movement, temporal gestalt and montage rhythm, sound composition plays a crucial role in order to develop film images as movement patterns. Here, we are particularly interested in the connection between the experience of filmic sound events and their multimodal description: How are qualities of sound – and the affective responses provoked by them – expressed through the dynamic visual-spatial modality of co-speech gestures? Drawing on Mittelberg's [16] notion of the 'exbodied mind', we argue that gestures have the potential and tendency to express what Johnson [14, p. 31] calls "felt qualities of our experience, understanding and thought" in an immediate fashion [see also 17, 18].

2. Embodied sound perception and strategies of listening to film sound

In addition to speech and other vocal expressions such as screaming, we experience a variety of sounds in our everyday life. Film producers work with sounds to evoke certain expectations and sensations, e.g. fear in a horror movie. For instance, when sound becomes louder and shriller, we anticipate a terrifying event (see also [19]).

Huvenne [20] proposes four strategies of listening to sound in film, focusing on the role perception and imagination may play in these processes. The first strategy involves the emergence of a visual image when the sound causes the spectator to connect it to the memory of a formerly experienced scene. As a second strategy, a motor image is evoked when the spectator's body resonates with the sound's source movement. Here the emphasis on a pre-reflective way of listening is crucial: "Instead of a visual image being evoked, we listen first and foremost to the dynamic qualities of the movement and the physical efforts in the steps" [20, p. 140]. It's also important to note that sound experience is not an isolated event but always interwoven with a continuous flow of multimodal action in the film as a whole. Thus, Huvennes' [20] third listening strategy implies that the temporal composition of sound is related to the temporal composition of the film fragments. Here, the sound becomes meaningful as it connects with the possible reaction to it, e.g. the sound may be interpreted as pleasant or unpleasant. Based on a certain sound we expect a certain event to happen and thus an emotional affect may arise. Scherer and colleagues [15] highlight the temporal dimension of emerging affect in film perception. Especially the change of a sound (e.g. calm turns to frightening sound) creates tension and thus a gestalt like temporal unit. When sound is recorded from the protagonist's point of view, the spectator co-experiences it. This double experience of sound defines the fourth strategy, i.e., when the spectator resonates not only with the sound, but also identifies with the protagonist of the film. Here, the recorded sound positions the spectator in relation to the sound as well as to the action or object causing it.

The present study does not investigate instances of bodily action as the source of sound. However, these listening strategies are particularly relevant when exploring the re-experiencing of film sound during multimodal narrations, as they recruit the viewer's active perception and involvement. We suggest that the description of film sound is rather a relived experience than a description of sound qualities. It is the qualities of sound that merge with the qualities of feelings.

3. Multimodal descriptions of sound impressions

When describing sound, one may generally refer to 1) the **source** (e.g. environmental/naturalistic sound produced by nature; instruments; human audible action; machines or technical devices); 2) the **type** of sound (music, noise; see Smalley [3]); 3) the **volume** (loud, quiet); as well as 4) the physical (e.g. emotional) **reaction** to it (e.g. frightening).

Previous research has shown that when describing music and sound events, people tend to utilize gestures [21], [22]. A study by Caramiaux and colleagues [22] presents a quantitative, multimodal analysis of movement and sound. The authors explored the relation between audio descriptors and movement features by focusing on spontaneous body movements of their study participants, while they were listening to recorded sound. Results suggest a high correlation between loudness of sound and position as well as sharpness of the sound and both the velocity and acceleration of movement. They conclude that a gestural expression and the sound it tries to capture can be seen as being intertwined instead of separate entities. In another study, Caramiaux and colleagues [21] asked people to listen to sounds and to perform gestures describing the sounds. She investigated the difference of causal (action-derived) and non-causal sounds, whereby the causal sounds mainly induced gestures mimicking the sound-causing action. Non-causal sounds were rather represented by gestures tracing contours evoking specific acoustic features of a given sound. As we hope to show in this paper, in multimodal narrations qualities of sounds merge with felt qualities of meaning on the side of the spectator, thus motivating expressive, communicative movements of the hands and torso.

4. Research questions

Taking gesture as a starting point, this study explores the relationship between sound, emotion, and motion. It focuses on gestures produced in natural, uninstructed story retellings. Importantly, the sounds in question are contextually perceived within a story-environment and afterwards described from memory (see the third and fourth listening strategies according to Huvenne [20] discussed above).

Our main research question asks how sounds are described and enacted through co-speech gestures, and how sensations experienced when viewing the film are conveyed multimodally. For the purpose of the present paper, we focus on sound produced by non-bodily action: Small technical devices used in everyday life (e.g. coffee machine), heavy-duty machines (e.g. an electric saw), and natural environmental sound (e.g. the rustling of leaves).

In particular, we were interested in the following aspects: 1) How do gestures express sound when the source is not bodily movement but environmental sound produced, e.g. by plants or technical devices? 2) Do the observed gestures depict sound qualities as such, or rather express the impact the sound had on their bodies and emotional states? 3) To what extent may embodiment theory account for the observed bodily techniques of enacting/describing sounds?

As the retellings were recorded in a Motion-Capture Lab, precise capturing, visualization and comparison of the gestural motion features was possible. Similarly to Caramiaux and colleagues [22], we determined whether verbal descriptions of the volume of a given sound (e.g. quiet, loud) relate to measurable gestural features (e.g., torsion and velocity).

5. Study

The study was conducted in the Natural Media Lab[1] at RWTH Aachen University. 11 native-speakers of German (6 female) participated in the study. They were asked to watch a short movie and then to retell the story line and reflect on it while being recorded by video and infrared cameras using a motion-capture system.

5.1. Stimulus Movie

All participants watched the German short film 'The Archivist' by Michael Cherdchupan[2]. The film is 15.44 minutes long and shows the daily routine of an archivist who listens to and archives tape recordings of various kinds of sounds in a dark archive. Day by day he meticulously analyzes the recordings and documents their properties in a catalogue. Being alone and separated from social life the sounds ripple his imagination. All sounds (e.g. leaves rustling or noise coming from a sawmill) were edited to create a unique sound experience while watching the film. In correspondence with how the protagonist seems to perceive the various sounds,

[1] http://www.humtec.rwth-aachen.de/nmlab

[2] https://www.youtube.co m/watch?v=IlBRvLvHu1A

Figure 1: Stills from the stimulus film "The Archivist"

they are conveyed with exaggerated loudness and distinctiveness. Close-ups of the protagonist as well as of the items and events causing loud and striking sounds create an intimate relation between the viewer and the protagonist (as described by Huevenne's [20] fourth listening strategy). Some sounds are perceived before their source appears in the visual scene, e.g. the vibrating sound of a cellphone, thus giving the viewer time to anticipate the source and also possible effects.

5.2. Data recording

The narrations were captured with an HD video camera and an optical Motion-Capture system. The latter is an infrared camera system capturing – with the help of reflective markers – the position and movement of the hands in three dimensional space. It further allows the visualization of movement traces of gesturing hands (see Fig. 2 A and figures further below). Participants were asked to wear 42 reflective markers. For each hand we used 14 markers including one on the wrist. Additionally, markers were placed on the head, neck, shoulders, arms, and knees. During the narration task, the interviewer was sitting directly behind the video camera placed opposite the participants (see Fig. 2 B & C).

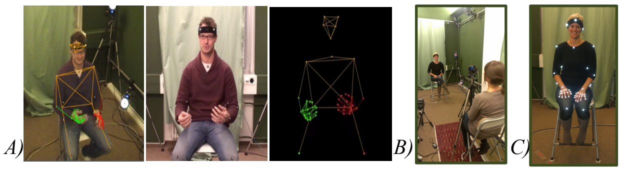


Figure 2: A) HD Video camera and Motion Capture Technology; B) Natural Media Lab; C) Marker Positions

5.3. Coding of the video data

The video data from the narration sessions were annotated and transcribed by a single coder using the software ELAN. All recorded gestures were segmented into gesture units, phrases and phases [23], which were coded independently for each hand. Only expressive gestures phases (i.e. strokes and holds; [24]) were considered, since they are taken to contribute most significantly to multimodal meaning construction. For each expressive phase, the four basic kinematic parameters – hand form, movement, position and orientation of the palm – were annotated [25], [26].

Speech was transcribed according to DuBois [27]. The transcription was added into the ELAN annotation. Transcripts were then searched for verbal terms referring to sound events. A description such as *you hear a cup clattering* (Germ. orig.:'*Tasse hört man klappern*') was counted as one verbal reference. Then the ELAN annotations were searched for expressive gestural phases co-occurring with these verbal expressions. Here, we only analyzed those gestures that express loud or quiet sound.

5.4. Processing and visualization of the motion-capture data

The Motion-Capture system records the participants' kinesphere and reduces the data into a three-dimensional matrix [28]. To describe the gestures' geometry and dynamics we lean on Caramiaux and colleagues [22] by deriving multi-dimensional coordinates of torsion (maximal distance of measured movement trace) and velocity (coordinates v_x, v_y, v_z, measured from the index finger) as two movement types from motion-capture data. The motion-capture data-based gesture visualization allows for a more objective account when comparing the gestures in terms of spatial and temporal features. The stick figures shown below are a way to map the body of the participants and to depict the trajectory of the moving hands through the blue lines. To reduce complexity, only the trajectory of the upper index finger marker was selected to trace the movement of the entire hand. The starting point of a trajectory is marked by the cones. The hands' locations are documented using McNeill's gesture space model [25] as well as information regarding the distance between the speaker's hands and body. Hand configuration is described as clenched, bend or stretched[1]. A statistic analysis of the data is not included in this paper; we will pursue a quantitative approach using these data in a follow-up study.

6. Results

In total, 11 narrations with 76 verbal references of sound experiences were analyzed. In their verbal descriptions, participants referred to either the source, the quality, the type, or the effect of the sound. They further express their individual involvement in speech and gesture. As previously observed by Caramiaux and colleagues [21], our study participants tend to imitate the action, thus the source of the sound, when verbally referring to sound produced by bodily action. For instance, the speaker moves her hand as if she was wiping a table. Only very few gestures represent gestural beats or point to the imagined location of the sound via pointing gesture. As mentioned earlier, our analysis focuses on those gestures that present the quality and effect of perceived film sound: either undefined surrounding sound or specific sound caused by, e.g., leaves or a sawmill. Special attention is paid to the description of loud versus low sound and its impact.

Gestures relating loud vs. low sound differed with respect to the hands' positions and torsions. A weaker difference was observed in the hand configurations and measurements of velocity of the hand movements (mean of the velocity of the stroke). In terms of **position,** gestures expressing loud sound were either produced close to the speaker's body or the body part perceiving the auditory information, e.g., the ears (see [18] on metonymic processes recruiting such body-part indices). Gestures that express a quiet and calming sound were either movements produced along a horizontal axis right above knee level or starting from close to the body, in the 'center center' of gesture space ('center center' is defined as the position of the gesture right in front of the upper body, between chest and navel of the gesturer; see [25]), going outward and downward. In terms of **torsion**, gestures expressing loud sounds tend to show several comparatively large twists, whereas gestures concerning low sound move only in one direction: either outwards producing a horizontal line, or away from the speaker downwards in a small slope. Comparing the **configuration** of the hands revealed only a slight difference in the sense that gestures expressing loud sound tend to be crooked, whereas gestures evoking low sound tend to be executed with either bent or straight hands. The **velocity** of the gestures was not strongly influenced by the sound quality. Here, the observations based on the video sequence suggested a difference in the senses that loud sounds

[1] We did not measure the muscle tension (for example using electrodes) and refrained from measuring the curvature (i.e. distance between the index, pinky and wrist) using the motion-capture data, as this would only make sense in an intra-subject comparison.

were portrayed by speedy gestures, whereas quiet sounds were expressed through gestures exhibiting less speed. The motion-capture data revealed that this perception was misleading, as the velocity of gestures expressing loud sound and those expressing low sound was not strongly different in the analyzed examples. In the following sub-sections, we will discuss a set of gestural examples in detail.

6.1. Gestures expressing loud sound

In the first example (see Fig. 3), the participant refers verbally to *'very loud surrounding sounds'* (Germ. orig.: *'sehr laute Umgebungsgeräusche*'). On *very loud* he clenches both hands while lifting them from the 'lower center' ('lower center' is defined as the position in gesture space right above the legs of the seated gesturer) to the 'center center' of gesture space relatively close to his chest. His hands go quickly up and down (left hand: 808 mm/s; right hand: 832 mm/s) with a comparatively larger torsion (170 mm). His clenched hands express the strong effect the loud noise had on him.

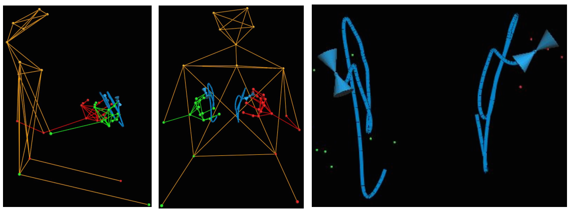


Figure 3: Gesture expressing loud surrounding sound

In the next example, the participant describes the impact of the *very loud* sound experience mentioned earlier on her own emotional state. She says the following: *there one notices how much sound may influence oneself, because* for example that* with the saw room, that immediately trou* or I myself was immediately troubled by it* (Germ. orig.: *'Da hat man mal gemerkt wie sehr Geräusche einen doch beinflussen können, weil* zum Beispiel das* mit dem Sägeraum da, das hat einen direkt beun* oder mich hat es direkt beunruhigt'*). Verbally she first mentions the general affect of the sound and then specifies her own reaction trigged by the sound. The first gesture co-occurs with the verbal phrase *sound may influence oneself* and is produced laterally on the right side with a small distance to her body (see Fig. 4). The dynamics of the gesture is comparable to the examples described above, also showing a quick pace (left hand: 680 mm/s; right hand: 921 mm/s) and larger torsion (151mm). The second gesture (Fig. 5) occurs with the words *immediately trou* or I myself* and is made very close to her own body with her finger tips point *at* her chest, thus directing the attention to her own experience. The velocity of this gesture is much less (left hand: 49 mm/s; right hand 43 mm/s) with a larger torsion (161 mm). In the next example (Fig. 6), the participant verbally refers to noise becoming louder and louder. On the mention of *then one hears more and more white noise* (Germ. orig.: *'dann hört man immer mehr Rauschen'*) she lifts both hands towards her ears. With a very quick motion of the right hand (left hand: 271 mm/s; right hand: 1038 mm/s) she draws a cycle-like trace. The hands are bent as in the two examples above; here with the fingers pointing at the ear. Thus an indexical relation between the gesturing hand and the body part per/receiving the sound is being created. Again the torsion is rather large (206 mm).

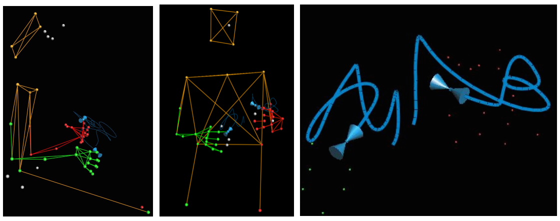


Figure 4: Gesture expressing the impact of the sound

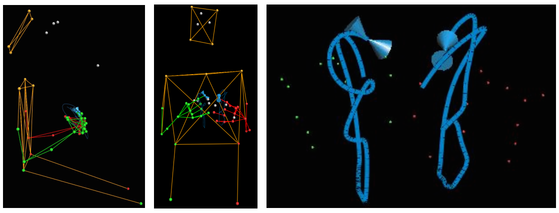


Figure 5: Gesture is performed close to speaker's body

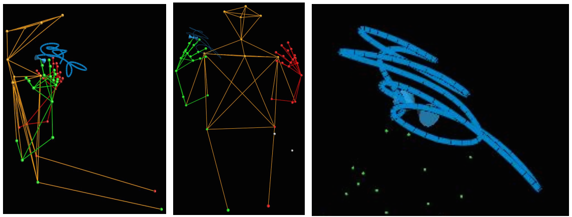


Figure 6: Gesture expressing noise becoming louder

6.2. Gestures expressing low/quiet sound

The participant compares the described loud sound with another scene in the forest, which was quiet. She connects the sound with its influence on her emotional state, as the sound calmed her down: *in the forest, for example, everything was very quiet, which immediately calmed one down* (Germ. orig.: *'Im Wald, die* zum Beispiel war alles sehr ruhig und das hat einen dann direkt beruhigt'*). This utterance is accompanied by two gestures. The first gesture in Figure 7 precedes the verbal reference to the sound in question – *very quiet* – and co-occurs with the verbal reference *everything*. The fact that the gesture precedes the verbal sound description may reflect that the experience was foremost physical and only secondarily verbal. From the captured motion trace, the arched trajectory of the gesture is clearly visible. The left hand starts from the 'center center' of gesture space close to the gesturer's body, moving out and downward in a well-formed arch. In comparison to the examples of the loud sound descriptions the torsion of this gesture is very small (11mm). The velocity (left hand: 527 mm/s; right hand: 105 mm/s) is slower than in the examples with participants describing loud sound; however, this difference is not strong. While moving the left hand is relaxed showing a bend configuration. The second gesture (Fig. 8) co-occurs with the verbal reference *immediately calmed down* with a low position of the hands and the wrist still leaning on the knees. The fingers of both hands are stretched forward and then draw a horizontal curve laterally outwards to both sides. Here, the velocity is rather high (left hand: 241 mm/s; right hand 907: mm/s), and the torsion small (58mm).

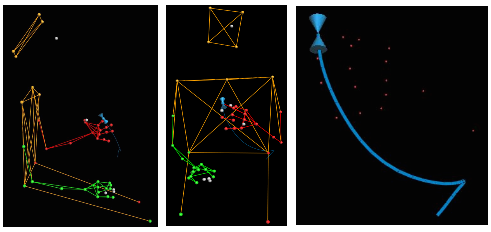


Figure 7: Gesture expressing quite surrounding

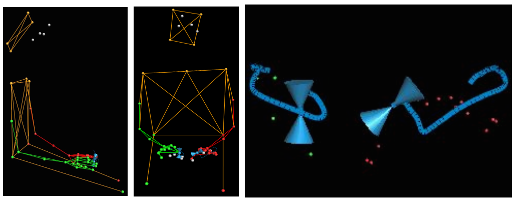


Figure 8: Gesture expressing the process of calming down

Interestingly, another participant (Fig. 9) is also leaning back with her shoulder while making a similar gesture evoking calmness. The trace of the shoulder is also depicted in Figure 9. Verbally she refers to her reaction to the sound *I noticed how I myself became very quiet to observe it* (Germ. orig.: *'Ich merkte wie ich selber ganz leise wurde um das aufzunehmen'*). Thus, the process of relaxing in response to the sound is described here. The concurrent gesture is similar to the gesture in Figure 8, which also draws a horizontal line. Here, the velocity is rather low (left hand: 497 mm/s; right hand 522 mm/s), with a rather small torsion (72mm).

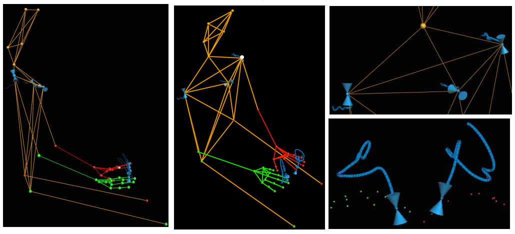


Figure 9: Participant leans back; gesture expressing her becoming quiet

7. Discussion

Using film sound as stimulus for analyzing expressive gestural movements in communicative acts has proven to be insightful, as the film retellings evoke sensorial qualities of experience as part of embodied listening strategies. While watching the movie, the participants experienced the sound. During the narrations they re-experience the sound and the effect it evoked in them. When reflecting on the filmic events they activate the physical dimensions of the listening and viewing experience. Our analysis has revealed that the volume of the sound and its physical reaction to it merge in accordance with Merleau-Ponty [29] who argued against an arbitrary relation between bodily experience and expression, suggesting a rather gestalt-like structure. The fact that the tension and relaxation of the speakers' hands reflect the loudness or calmness of the respective sounds supports the understanding of sound perception as a comprehensive sensorimotor and psychological experience [2].

As suggested in the beginning of the paper, gestures are far more than a behavioral component of emotions and particularly apt at expressing subjective stance and feelings. In Figure 7 the gesture precedes the verbal sound description underlining the assumption that the re-experiencing of sound qualities is foremost physical and only secondarily verbal. Whereas speakers tend to describe the quality of a given sound verbally, their gestures tend to express the effect of the sound. According to our view [16], such ‚exbodied' instances of experiential essence indeed invoke "felt qualities of our experience" [14, p. 31]. With these types of co-speech gestures, the gesturer does not refer to, but rather senses the sound through the way her or his own body resonates with the sound experience.

The strongest finding is the positioning of the gestures as well as their torsion in relation to the sound experience. In accordance with Huevenne´s fourth listening strategy two factors play a central role here: the participants 1) listened in an intersubjective way and 2) were situated listeners as the recorded sound invited them to react. Listening to a sound situates the listening body in space and time in relation to the sound, thus creating a certain nearness or distance to the sound. When perceiving a loud sound, one feels its closeness, vibration, and force. While retelling, the sound experience led to quick movements with bigger torsions located close to the body or to the ear as the body part perceiving it. Listening thus is an embodied practice of experiencing the sensorial environment including films as part of the semiotic world we live in.

The change in volume of sound gives the spectator the impression as if s/he was moving closer towards or further away from the source, thus leading to the embodied sensation the film evokes. The spectator experiences how his/her own body resonates with and is involved in the film. Thus, on the basis of bodily and affective sensing the participants grasp the story line and make sense of it cognitively and emotionally [30]. The participants found themselves attuned with the film´s sounds (see also [30] on this effect of images instead of sounds). When verbally referring to the process of 'calming down' the hands build up a tension, which gets released outward or downward. In a very engaged way, the participants empathize with the protagonist, e.g. through anticipating what a certain sound may forecast (see third listening strategy above). This kind of connection between the present moment and immediately following moments leads to a gestural (re-) presentation of the effect the sound evoked.

The sounds have an effect on their perceivers, and this effect is an affect in the sense of a physiological reaction that in turn motivates actual physical, communicative motion. In most of the examples, the sound talked about, movements and sensations merge in the multimodal description of the film scenes. The participants do not only refer to the sound quality itself but also express the influence the sound had on them and their bodily experience and memory of it. In summary, our analyses of multimodal film descriptions support (in line with Koch and colleagues [31]) an embodied account of perceived sound qualities, sound-induced sensations, as well as of their multimodal enactments.

As a next step, a quantitative analysis of the narrations is planned. A detailed observation and measurement of all gestures expressing sound experiences may give further insights into the active perception of film sound. We are interested in whether the tendencies identified in this study also holds for the entire data set. Furthermore, we plan to look at the multi-modal descriptions of additional sound events and sources.

Another question that arose from this study concerns the relation of metonymy and metaphor when describing and enacting (the influence of) sound qualities multimodally. The linguistic expression "the sound moved me" reflects, according to Conceptual Metaphor Theory (CMT), a conceptual metaphor, whereby the emotional response evoked by the perception of a sound is described in terms of 'physical movement' [32]. Foolen [9] argues that the figurative expression of emotion has itself an expressive motivation and suggests that the predominant form of figurative language about emotions is metonymic as an 'effect for cause' relation. In a similar vein, Mittelberg and Waugh [33], [18] have described how metaphoric gestural portrayals are rooted in fundamental metonymic processes. We thus will investigate whether multimodal descriptions of sound experience provide further evidence for metonymic/metaphoric relations between sound, sensation, motion, and meaning.

8. Conclusions

In this paper, we used auditory stimuli for multimodal descriptions to see how speakers describe with words and gestures sound impressions. Sound as such is invisible, but the sounds' sources and their effects on our physical and mental state can be described and perhaps imitated. Compared to traditional video analysis – relying on visual observation – Motion-Capture system allowed us to visualize and measure movement qualities more objectively. As these preliminary results show, there is a distribution of semantic and emotive dimensions across modalities. When describing from memory sounds stemming from a film context, participants conveyed sound qualities and the sensations they evoked through both speech and gestures. Our observations suggest that sensations caused by sound impressions link sound with motion and meaning physically. That is, it is not the sound that moves the person watching the movie, but the sensations and feelings evoked by the sound. Multimodal descriptions of experience and understanding as discussed in this paper may help the interlocutor not only to understand different aspects of the sound experience put into words, but also the speaker's sensorial and emotional experience while watching the film scenes. While the work cited in this paper no doubt represents a considerable step towards a fuller understanding of gesture's capacity to convey felt dimensions of experience, expression, and meaning more empirical research is needed to see how these various dimensions interact when people interact face-to-face or try to make sense of semiotic universes such as films and other multimodal forms of communication.

9. Acknowledgements

This study was funded by the Excellence Initiative of the German Federal and State Governments. We thank the members of the Natural Media Group for valuable input at various stages of this project, especially Bela Brenger and Ugur Töman for technical assistance, and the study participants for supporting our research.

10. References

[1] Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A. and Turnbull, D., "Music emotion recognition: A state of the art review", in Proc. ISMIR, Utrecht, The Netherlands, 2010.

[2] Smalley, D., "Spectromorphology: explaining sound-shapes", in Organised Sound 2, 107-126, 1997.

[3] Ekman, P., "Methods for measuring facial action", in K. R. Scherer and P. Ekman [Eds], Handbook of Methods in Nonverbal Behaviour Researh (pp. 45-90), Cambridge: Cambridge University Press, 1982.

[4] Brandstätter, V., Schüler, J., Puca, R. M. and Lozo, L., "Motivation und Emotion. Allgemeine Psychologie für Bachelor", Berlin: Springer, 2013.

[5] Zlatev, J., "Prologue", in Foolen, A. et al. [Ed], Moving Ourselves, Moving Others, 1-25, 2012.

[6] Müller, C., Cienki, A., Fricke, E., Ladewig, S. H., McNeill, D. and Teßendorf, S. [Eds], "Body – Language – Communication: An International Handbook on Multimodality in Human Interaction", Handbooks of Linguistics and Communication Science (38.1). Berlin: Mouton de Gruyter, 2013.

[7] Müller, C., Cienki, A., Fricke, E., Ladewig, S. H., McNeill, D. and Bressem, J. [Eds.], "Body – Language – Communication. An International Handbook on Multimodality in Human Interaction", Handbooks of Linguistics and Communcation Science (38.2), Berlin/Boston: De Gruyter Mouton, 2014.

[8] Müller, C., "Gestures as a medium of expression: The linguistic potential of gestures", in C. Müller et al. [Eds.], 202-217, 2013.

[9] Foolen, A., "The relevance of emotion for language and linguistics", in A. Foolen et al. [Ed], Moving Ourselves, Moving Others, 349-368, 2012.

[10] Horst, D., Boll, F., Schmitt, C. and Müller, C. "Gesture as interactive expressive movement: Inter-affectivity in face-to-face communication", in C. Müller et al. [Eds.], 2112-2125, 2014.

[11] Bühler, K., "Sprachtheorie: Die Darstellungsfunktion der Sprache", Stuttgart, Fischer. First published [1934], 1982.

[12] Damasio, A., "Looking for Spinoza: joy, sorrow and the feeling brain", New York, Harcourt Brace, 2003.

[13] Gibbs, R. W. Jr., "Embodiment and cognitive science", Cambridge University Press, Cambridge, 2006.

[14] Johnson, M.,"The philosophical significance of image schemas", in B. Hampe [Ed.], From Perception to Meaning: Image Schemas in Cognitive Linguistics. Edited in cooperation with J. Grady. Berlin/New York, Mouton de Gruyter, 15–33, 2005.

[15] Scherer, T., Greifenstein, S., and Kappelhoff, H., "Expressive movements in audio-visual media: Modulating affective experience", in Müller et al. [Eds], 2080-2092, 2014.

[16] Mittelberg, I., "The exbodied mind: Cognitive-semiotic principles as motivating forces in gesture", in C. Müller et al. [Eds], 750-779, 2013.

[17] Mittelberg, I., "Balancing acts: Image schemas and force dynamics as experiential essence in pictures by Paul Klee and their gestural enactments", in M. Bokrent, B. Dancygier, J. Hinnell [Eds.], Language and the Creative Mind, 325-346, 2013.

[18] Mittelberg, I. and Waugh, L. R., "Gestures and metonymy", in C. Müller et al. [Eds], 1747-1766, 2014.

[19] Boltz, M., "Musical soundtracks as a schematic influence on the cognitive processing of filmed events", Music Perception, 18: 427-455, 2001.

[20] Huvenne, M., "Sound in film as an inner movement. Towards embodied listening strategies", in H. De Preester [Ed], Moving Imagination, John Benjamins, 133-148, 2013.

[21] Caramiaux, B., Bevilacqua, F., Bianco, T., Schnell, N., Houix, O. and Susini, P., "The role of sound source perception in gestural sound description", ACM T. Appl. Percept.,11(1), 2014.

[22] Caramiaux, B., Bevilacqua, F. and Schnell, N., "Towards a gesture-sound cross-modal analysis", in S. Kopp, I. Wachsmuth [Eds], Lecture Notes in Computer Science, Embodied Communication and Human-Computer Interaction, 8th International Gesture Workshop, 158-170, Heidelberg, 2010.

[23] Kendon, A., "Gesture: Visible action as utterance", Cambridge u.a.: Cambridge University Press, 2004.

[24] Kita, S., Gijn, I. van and Hulst, H. van der., "Movement phases in signs and co-speech gestures, and their transcription by human coders", in I. Wachsmuth and M. Fröhlich [Eds], Gesture and Sign Language in Human-Computer Interaction, 23–35, Berlin/ Heidelberg, Springer, 1998.

[25] McNeill, D., "Gesture and Thought", Chicago, London, University of Chicago Press, 2005.

[26] Bressem, J., "A linguistic perspective on the notation of form features in gestures", in C. Müller [Ed], 1079–1098, 2013.

[27] DuBois, J. et al., "Outline of discourse transcription", in J. A. Edwards and M. D. Lampert [Eds], Talking Data: Transcription and Coding in Discourse Research. Mahwah, Lawrence Erlbaum Assoc., 45-87, 1993.

[28] Priesters, M., Mittelberg, I., "Individual differences in speakers' gesture spaces: Multiangle views from a motion-capture study", Proceedings of the Tilburg Gesture Research Meeting (TiGeR), June 19-21, 2013.

[29] Merleau-Ponty, M., "Phenomenology of Perception", London/New York: Routledge. First published [1945], 2005.

[30] Schmitt, C. and Greifenstein, S., "Cinematic communication and embodiment", in Müller et al. [Eds], 2061-2070, 2014.

[31] Koch, S. C., Fuchs, T., Summa, M. and Müller, C [Eds], "Body Memory, Metaphor and Movement", Amsterdam: John Benjamins: 2012.

[32] Lakoff, G. and Johnson, M., "Philosophy in the flesh. The embodied mind and its challenge to western thought", Basic Books, New York, 1999.

[33] Mittelberg, I. and Waugh, L. R., "Metonymy first, metaphor second: A cognitive-semiotic approach to multimodal figures of speech in co-speech gestures" in: C. Forceville and E. Urios-Aparisi [Eds], Multimodal Metaphor, Berlin/New York, Mouton de Gruyter, 229-356, 2009.

Multimodal analysis of hand gesture back-channel feedback

Jorane Saubesty [1,2], *Marion Tellier* [1]

[1] Aix-Marseille Université, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France
[2] Brain and Language Research Institute, Aix-en-Provence, France

Jorane.Saubesty@blri.fr, Marion.Tellier@lpl-aix.fr,

Abstract

This study explores the distribution of back-channel feedback in terms of modality in a doctor-patient interaction. The literature studied revealed that previous studies on feedback hardly considered hand gestures among the modalities that serve this function. In this study we therefore investigate whether hand gestures can perform a feedback function or not. Preliminary results show that back-channel feedback is mostly expressed through two modalities: *speech* and *head movements*, separated or together. However, we also observe some hand gesture feedback. Interestingly, we find that adaptors can be used as feedback when produced for a communicative purpose.

Index Terms: back-channel feedback, multimodality, hand gesture, speech, head movements.

1. Introduction

In the medical context, breaking bad news is a complex task. On the one hand, it is very stressful for doctors because they have to deliver bad news to patients who usually trust them [1]. On the other hand, it is very important for the patient to understand his/her health situation to be able to recover properly [2].

In France, this issue has been increasingly discussed for the past ten years and has started to be taken into account by the French health authorities. In 2004, a law was passed to make it compulsory to inform patients about their health situation and the law further stipulates that the announcement be made in a certain setting [3]. The High Authority of Healthcare (HAS) also published a report on how to break bad news [4] and another on the announcement of damage due to healthcare [5]. Moreover, a call for the conception of training tools to develop doctor’s communication skills was launched in 2009. One of the tools developed since, is the training session in which doctors train to break bad news to patients/actors [6].

Breaking bad news is not strictly speaking a treatment, but an aspect of care. Understanding what the doctor is saying is of primary importance. Thus, when breaking bad news, doctor and patient work together to enable understanding: the doctor uses everything he/she can to be understood while the patient gives feedback showing his/her understanding or misunderstanding. Hence, feedback is the keystone of understanding between the doctor and the patient and the key to a successful interaction, and in some respects to a successful treatment. In this kind of interaction, where the need for understanding and being understood is very strong, studying feedback could provide information on how the interaction works and on how mutual understanding is constructed.

While speaking, participants keep giving and eliciting “information about their attention, perception, understanding and reactions to what is said by others” [7:118]. In other words, they keep giving and eliciting feedback. The latter is both visible and audible and can include, among other things, expressions such as “mhm”, “yeah”, “no”, nods, smiles, or dramatic intakes of breath [8]. It can be produced on the main channel but its privileged channel is the back-channel (i.e. a channel used to speak without disrupting the main speaker). In this study, we consider back-channel feedback only [9].

We approach this study from a multimodal perspective, meaning that we are considering all the modalities one may use to convey a message [10]. The term *modality* here is referring to the sensory modalities which are *visual*, *auditory* and *tactile*. In the multimodal perspective we are considering the visual (i.e. kinesic) and auditory (i.e. oral) modalities. We go further, distinguishing five different sub-modalities, which one may call canals: *verbal* and *prosody* for the auditory modality, and *hand gestures*, *postures, facial gestures and head movements* for the visual one. For convenience, we will refer to these sub-modalities as *modalities* throughout this paper. When we speak, we do not only say words but we say them with a certain melody (i.e. prosody). We can reinforce those words with facial gestures (i.e. gaze, smile, eyebrows, etc.) or illustrate them with hand gestures, for instance. All the modalities may work together to convey the message (see Figure 1).

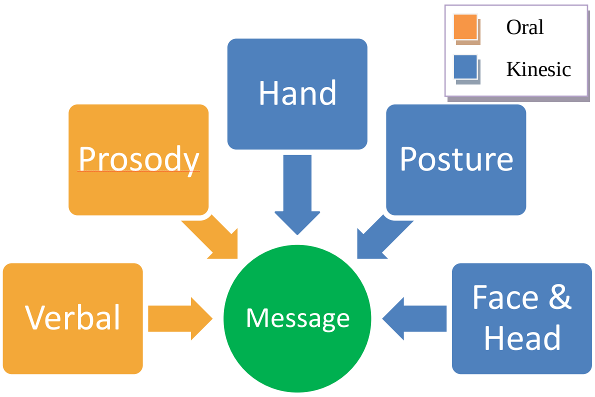


Figure 1: Multimodality scheme

Most of the studies on feedback focus on speech, prosody and head gestures [11, 12, 13]. Studies have explored the distribution of speech utterances used for feedback, the distribution of the different types of head movements, or the semantic functions of such feedbacks [11, 14, 15]. Some of them have also sexamined postures [16], but to our knowledge, very few of them mention hand gestures used as feedback. Bertrand et al. [17] and Allwood and Cerrato [12] do consider hand gestures when coding their data but [17] only analyzed head movements and facial expressions as backchannels and [12] do not mention hand gestures in their results (that might be because there is no hand gesture used for feedback in their data). Malisz, and Karpiński [18], however, do annotate hand gestures and find some but very few “iconic/deictic” gestures [18:3] overlapping or produced near to the verbal responses produced by the instructions giver.

Given the fact that the literature does not focus on the exploration of the production of hand gestures as feedback, **we propose to investigate whether hand gestures can have a feedback function or not and if so, we will be interested in what types of gestures they might be[19].** To answer these questions we explore the distribution of back-channel feedback depending on their modality. Hand gestures analysed in this study are co-speech gestures and emblems, but also self adaptors when they are used to convey feedback. We will first present our methodology, and then we will give a descriptive analysis of the data. Before concluding, we will also propose qualitative analyses of the hand gestures used as back-channel feedbacks encountered in our corpus.

2. Materials and method

2.1. Corpus

The corpus was created by and belongs to the Institut Paoli-Calmette (Marseille, France) and is an authentic corpus of training sessions for doctors involved in role plays with an actor playing the role of a patient. It is composed of 6 audio-video recorded interactions involving one doctor in training and one actor/patient. Each dialogue is around 15 minutes long.

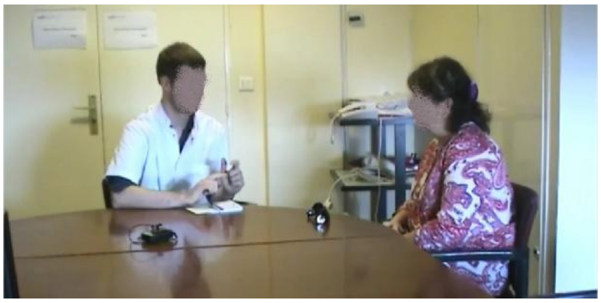


Figure 2: corpus overview

In the beginning of each training session, participants receive information to guide them in their role play. The doctor has to break bad news to the actor/patient. In our case, the actor actually plays the role of the patient's wife.

The corpus was not recorded for research purposes but to allow a debriefing on the doctor's actions. However, it offers an interesting corpus for a linguist in that it is an ecological corpus of training sessions. Due to the nature of the corpus, the camera's view is focused on the doctor and does not allow us to see the actor's face. Furthermore, the quality is too low to annotate facial expressions, as well as a fine-grained prosody annotation, so these will not be taken into account although they are of importance for a multimodal analysis.

2.2. Annotation steps

Different tools were used in order to annotate the corpus. Speech was manually transcribed using PRAAT [20] and was automatically segmented into inter-pausal units (IPUs) and tokens using SPPAS [21]. Discourse and kinesic categories were manually annotated using ELAN [22].

2.2.1. IPUs and tokens

We automatically segmented the signal into IPUs, which are speech units each separated by at least a 200-millisecond pause. The duration of silence used to segment IPUs is the mean duration of silences in French; 200 milliseconds correspond to a plosive's start duration. If less than 200 milliseconds were used to segment IPUs, automatic segmentation tools might cut into a plosive consonant. Thus, using IPUs allows us to get utterance units free from any interpretation [23].

We manually transcribed each participant's speech on different tiers using the TOE convention (Transcription Orhographique Enrichie / Enriched Orthographical Transcription [24]).

IPUs were then segmented into tokens. Items such as "hmm" or "euh" were considered as tokens.

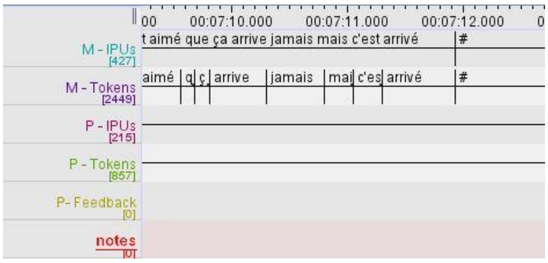
2.2.2. Feedback

To annotate feedback, we followed the definition given in the introduction, that is, utterances produced in the back-channel in order to give information about the speaker's attention, perception, understanding and reaction to the main speaker's speech.

First, we manually segmented feedback utterances for each participant. Then we labeled each of the segments depending on their modality. We pointed out which modality was used for each of the utterances. We coded "S" for speech, "G" for hand gestures, "H" for head movements and "Post" for postures. We used the mathematical symbol "+" for feedback produced in several modalities. For example, if a participant said "hmm" while nodding, we annotated "H + S".

2.3. Inter-coder agreement

In order to validate the segmentation and the coding of feedback, 5 minutes of the corpus were annotated by 3 more annotators. Two of the annotators were experts in multimodal analysis while the last one was a non-expert. We gave the annotators an ELAN [22] file, the audio/video recording and an annotation guide. The ELAN file contained six tiers: Doctor IPUs and tokens (D – IPUs, D – tokens), Patient IPUs and tokens (P – IPUs, P – tokens), an empty tier dedicated to feedback annotation (P – Feedback) and a note tier (note) so coders could give comments about the coding (see Figure 3). A quick explanation on how to use ELAN [22] was given to the naïve annotator.



t aimé que ça arrive jamais mais c'est arrivé | #

aimé | q | ç | arrive | jamais | mai | c'es | arrivé | #

Figure 3: Screenshot of ELAN – annotator file.

To calculate the inter-coder agreement, we coded each segment using numbers. For example, if a segment was coded "1", it meant only one annotator identified the utterance as feedback. On the contrary, if the segment was coded "3+1", all the annotators agreed on the identification of the segment, but only 3 of them agreed on the labeling (i.e. head, gesture, speech, etc.).

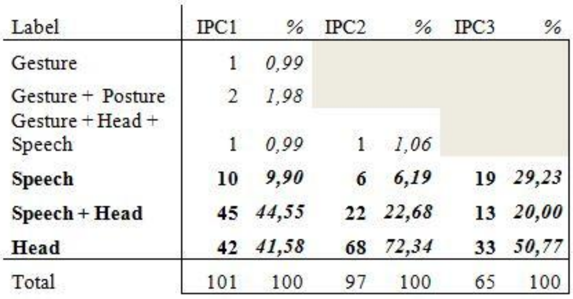
We then extracted occurrences of each code with SPPAS and we calculated two agreements. Annotators agreed on 67.21% of the identification of segments, while they agreed on 75.41% of the labeling.

3. Analysis and results

3.1. Descriptive analysis

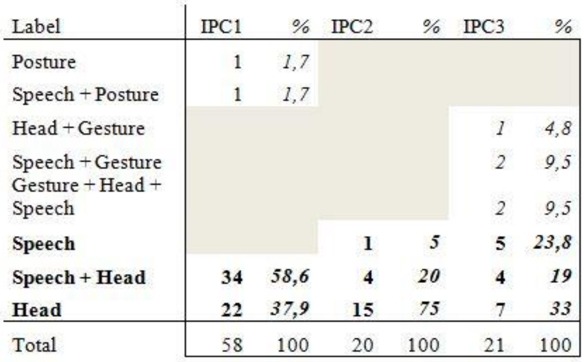
The results present here are preliminary. Analyses were conducted on three interactions only. Because we are currently unable to present statistical results, we are presenting descriptive statistics (number of occurrences and percentages). We extracted the data from the tiers "P – Feedback" and "D – Feedback", respectively for the patient and the doctor using SPPAS [21]. We found 263 occurrences of feedback produced by the patients and 99 occurrences of feedback produced by the doctors.

In order to analyze the distribution of feedback modalities through the corpus, we calculated the percentage of occurrences of each modality actually observed compared to the total number of occurrences (table 1 and 2).



Label	IPC1	%	IPC2	%	IPC3	%
Gesture	1	*0,99*				
Gesture + Posture	2	*1,98*				
Gesture + Head + Speech	1	*0,99*	1	*1,06*		
Speech	**10**	***9,90***	**6**	***6,19***	**19**	***29,23***
Speech + Head	**45**	***44,55***	**22**	***22,68***	**13**	***20,00***
Head	**42**	***41,58***	**68**	***72,34***	**33**	***50,77***
Total	101	100	97	100	65	100

Table 1: Distribution of the patient's feedback per modality per dyads



Label	IPC1	%	IPC2	%	IPC3	%
Posture	1	*1,7*				
Speech + Posture	1	*1,7*				
Head + Gesture					*1*	*4,8*
Speech + Gesture					*2*	*9,5*
Gesture + Head + Speech					*2*	*9,5*
Speech			**1**	***5***	**5**	***23,8***
Speech + Head	**34**	***58,6***	**4**	***20***	**4**	***19***
Head	**22**	***37,9***	**15**	***75***	**7**	***33***
Total	58	100	20	100	21	100

Table 2: Distribution of the doctor's feedback per modality per dyads.

Three modalities or combination of modalities clearly appear more frequently: *speech*, *head movements* and *speech and head movements*. Interestingly, no feedback was produced with "speech only" by the doctor IPC1. We can see that most feedback is produced through two categories only, which are *head movements* only, and *speech and head movements* (223 occurrences, 84.8% for the patients and 86 occurrences, 86.6% for the doctors). On the contrary, hand gestures and postures are hardly represented (3 without speech, and 2 with speech for the patients and 1 without speech and 4 with speech for the doctors, all produced by the same doctor). We thus notice that even in this specific interaction, speech and/or head movements are the main modalities used to convey feedback, which is consistent with previous findings on other kinds of interaction (different task, setting and language [11,12,13]). Hand gestures exist but are scarce.

3.2. Qualitative analysis

In order to better understand the production of hand gesture back-channel feedback, in this section we will present some qualitative analyses concerned with the patient, as well as the doctor. Then, we will pay attention to a particular type of feedback which was not discussed above.

Example 1

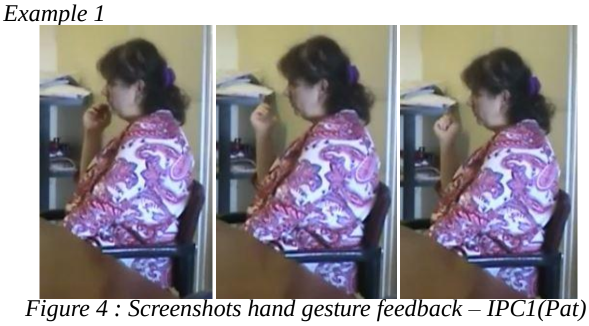


Figure 4 : Screenshots hand gesture feedback – IPC1(Pat)

Doctor: alors là pour l'instant c'est c'est c'est plus la priorité là pour l'instant c'est **la priorité** c'est d'amélio- de l'améliorer sur le plan respiratoire afin qu'on puisse enlever **[#379ms# la sonde d'intubation et qu'il se]** remette à respirer normalement

So here for the moment it's it's it's not the priority anymore now for the moment it's ***the priority*** *is to amelio- to ameliorate him on the respiratory level so we can take of* ***#379ms# the breathing tube out and he*** *starts to breathe normally again*

Interlocutor : **Vous êtes sûre ?** #3420ms# **[geste manuel]**

Are you sure? *#3420ms#* ***[hand gesture]***

Note: English translation is in italics. "[ ]" mark the beginning of an overlap with a given feedback. "{}" mark the beginning of an overlap with the speech part of a given feedback. Bold marks the overlap of doctor and interlocutor productions. Pauses are noted between "#".

The doctor just told the patient's wife that her husband is in the resuscitation unit after a problem which occurred during surgery. The patient's wife is worried about the cancer her husband might have, so she asked the doctor what the doctors can do about the cancer now that her husband is stable in the resuscitation unit. The doctor thus answers that right now, the cancer is not a priority. The patient's wife looks sceptical and asks the doctor if she is sure of what she says. The doctor does not take this intervention into account and continues her explanation. At this moment, the patient's wife has her hand under her chin; she then produces a hand gesture, index finger pointed up with a rotation of the wrist. The doctor still does not take into account the intervention of the patient's wife. The latter is not relieved by the doctor's answer, and asks again about the cancer. In this example, we can see that the patient's wife is having concerns about her husband's cancer. The doctor is trying to tell the patient's wife that the cancer is not a priority but is not successful. Indeed, the patient asks the doctor for a confirmation of what she says is right, then produces a hand gesture that could be interpreted as "what?" then asks again about the cancer. The patient's wife is using different strategy in order to get an answer. This gesture here is almost an emblem, in that we can interpret it without any speech and is a feedback in that it is a reaction, produced in the back-channel, to what the doctor is saying.

Example 2

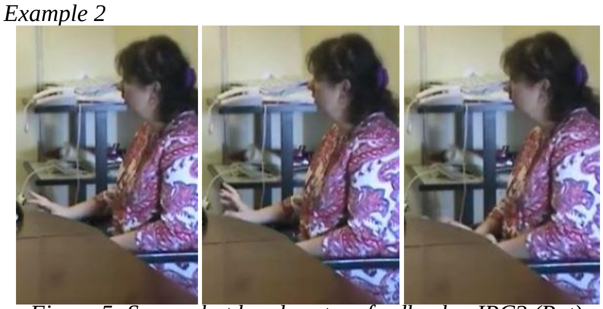


Figure 5: Screenshot hand gesture feedback – IPC2 (Pat)

Interlocutor: même le cancer de la vessie qu'il a pour l'instant on s'en occupe plus c'est ce que vous me **dites aussi #1756ms#**

Even the bladder cancer he has for now we do not take care of it anymore it's also what you are telling me #1756ms#

Doctor : **on peut** pas #316ms# [**on peut pas s'en occu{per**

We *can't #316ms#* ***[we can't take care of {it***

Interlocutor : **d'accord}]**

O.K.}]

This example is taken from a different pair of interlocutors but the moment of the interaction is the same. The actress playing the patient's wife is the same actress. The doctor and the patient's wife are discussing the treatment of the husband's cancer. The patient's wife asks for a confirmation that the doctor has stopped the treatment of the cancer; she then finishes her question with an interactive gesture toward the doctor and she holds it while the doctor is answering. In this example the patient's wife is producing a feedback using a combination of modalities. She first starts nodding while the doctor is answering and does it during the whole production of the feedback utterance, and then does a movement upwards with her hand and finally she says "O.K." The hand gesture is a co-speech beat appearing slightly before the speech it accompanies ("O.K.").

Example 3

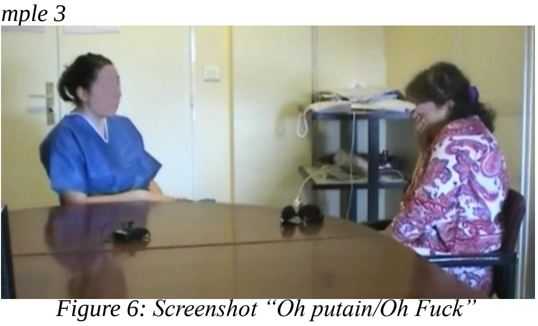


Figure 6: Screenshot "Oh putain/Oh Fuck"

Doctor: Donc euh donc ça a entrainé une #703ms# ce qu'on appelle une détresse respiratoire hein il a il **s'est mis à** voilà à plus bien respirer du tout **[on a dû {l'intuber} #1467ms# là actuellement il est intubé #534ms# hein il est intubé avec une ma]**chine qui le fait respirer parce qu'il a des il peut il peut pas bien respirer d'accord

So hmm so it evolved into a #703ms# what we call a respiratory distress uh he has ***he started to*** *well not to breathe well at all* ***[we had to {intubate him} #1467ms# right now he is intubated #534ms# uh he is intubated with a ma]****chine that helps him breath because he has some he can he can't breathe well O.K. ?*

Interlocutor : **Il arrivait plus à respirer ?** #1310ms# **[Oh putain]**

He couldn't breathe anymore? *#1310ms#* ***[Oh fuck]***

In this example the doctor is finally talking about the main reason why the patient's wife was called to the hospital (the interview began 3 minutes ago). The patient had an emergency surgery, and during that surgery, some liquid finds its way into the lungs and causes a respiratory distress. We can see in this example that the doctor's pauses are very long, we can suppose the doctor has trouble saying what happened. She starts her explanation using the connector "so", then stops for 703 milliseconds. When she starts again, she immediately uses a medical term "respiratory distress". She is conscious it is a medical term, and she begins to explain it right away but she stutters. At this point the patient's wife asks "he couldn't breathe anymore?".The doctor accepts the proposition included in the question by saying "voilà". Then the situation is unfrozen; the doctor repeats the terms used by the patient's wife, then moves on, describing the situation a little more, saying they had to intubate the patient. When she says that, the patient's wife breaks down, she curves, passes her hand over her face and says "oh fuck" (see Figure 4). She keeps that posture until the end of the feedback. In this example, the feedback is not given with head movements only; instead the patient's wife uses her whole body to give feedback. That moment of the interaction is really important for the patient's wife; because it is the moment when she actually finds out what happened to her husband. The hand gesture that the patient is using is typically a self adaptor. Self adaptors are non-communicative self-contact gestures [24]. The patient is touching her face, but she does it in a very dramatic way. Doing so, she is using a non-communicative gesture in a communicative way to signify the news is really breaking her down.

Example 4

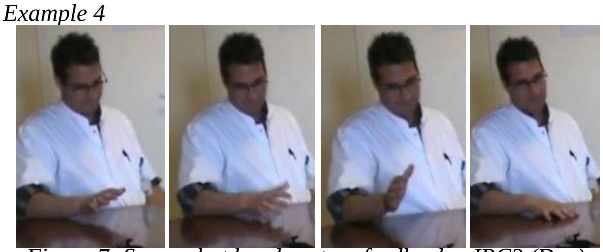


Figure 7: Screenshot hand gesture feedback – IPC3 (Doc)

Interlocutor: quand vous quand on l'a anesthésié on pouvait pas e- vider l'estomac avant pour éviter qu'il #421ms# que qu'il y ait qui qu'il y ait ces liquides qui passent dans les poumons et que #492ms# **[et qu'il se retrouve en réanimation ou][#439ms# ou #579ms# prévoir avant que ça se]** bouche et mettre la prothèse ou

When you when he was anesthetized we could not a-empty the stomach before to avoid that he's #421ms# that there is that there is that liquids going into the lungs and that #492ms# ***[and that he ends up in the resuscitation unit or][#439ms# or #579ms# to prevent it before it]*** *clogs and install the prosthesis or*

Doctor : **[geste manuel + mouvements de tête][geste manuel + mouvements de tête]**

[hand gesture + head movements][hand gesture + head movements]

So far, we have only examined examples produced by the patient. But in this corpus, the patient is always the same person (since it is an actor). One could argue that the use of a hand gesture as feedback might be idiosyncratic. In this example, we now focus on the doctor's feedback. He is the participant who is producing the most hand gestures as feedback. The patient's wife here is asking if anything could have been done to avoid the respiratory distress caused by the surgery. The doctor is helpless. He does not answer, letting her finish her sentence. However, he shakes his head repeatedly and produces beats with his hand on the vertical axis. He also produces another hand gesture, open hand, palm down, with a movement of the wrist and the forearm to reach an almost palm-up position, then returns to a rest position. This gesture might be interpreted as an emblem showing the helplessness of the doctor. This emblem is reinforced by the shaking head.

Example 5

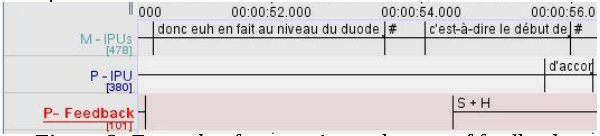


Figure 8: Example of annotation - absence of feedback (ELAN)

Doctor : Donc euh en fait au niveau du duodénum #535ms# c'est-à-dire le dé**[but de l'intestin grêle après l'est{omac**

*So hmm in fact at the duodenum level #535ms# that is to say the be****[ginning of the small intestine after the sto{mach***

Interlocutor: **D'accord}]**

O.K.}]

Finally, this example of a feedback raises a lot of questions and will probably lead us to rethink our annotation scheme. In this example, the doctor is telling the patient's wife why her husband had an emergency surgery. Since the beginning of the interaction, the patient's wife has been listening carefully to the doctor and has been producing feedback regularly. But in this example, the doctor says something using the specific medical term "duodenum" which is probably unknown to her interlocutor, and then she pauses for 535 milliseconds. During this time, the patient's wife does not answer, nor gives feedback anymore. So the doctor continues her turn, starting a new IPU with "that is to say", the interlocutor knows then that the doctor is now defining the specific medical term. Indeed, almost immediately after this sentence, the patient's wife starts to nod again, and does so until the end of the doctor's utterance where she finally says "O.K." This example shows that the absence of feedback is somehow giving feedback anyway (i.e. here to show her not understanding). The doctor was expecting feedback from the patient's wife. As she did not give it, the doctor initiated a repair sequence retelling what she just said, probably in a more legible way. Thus, this example shows that the absence of feedback can be interpreted as negative feedback.

Through these examples, we have seen that hand gestures can be used to give feedback, by the listener as well as by the main speaker when he/she is in the listener's position. Hand gesture feedback can appear in different dimensions such as co-speech gestures but also emblems. More interestingly, they can also be self adaptors used for a communicative purpose. Unfortunately, they are quite scarce, at least in this corpus, and the number of utterances is not sufficient to propose a functional analysis.

4. Discussion and conclusion

The aim of this study was to explore the distribution of back-channel feedback and its forms in terms of modality. The literature studied revealed that hand gestures have almost never been considered as a modality to express feedback. Our main goal was thus to find out whether hand movements could be used by the listener to react to the main speaker speech. We considered co-speech gestures, emblems and adaptors when used for communicative purposes. Our results show that back-channel feedback was expressed mostly through two modalities: *speech* and *head movements*. But we also found back-channel feedback expressed by hand gestures; however, the use of hand gestures in the function of feedback is relatively scarce. In their study, Malisz and Karpiński, [18] found more hand gestures feedback than we did. Ths can be due to the task. Indeed, the participants of Malisz and Karpiński,'s study [18] are involved in a folding paper task. Moreover, they analyse only the productions of the instruction giver. The nature of the task itself might explain the different results. These findings allow us to conclude that hand gestures can have the function of feedback, but the number of such feedbacks in our data is not sufficient to fully understand their functioning. Thus, future work will consist of fully annotating the corpus at our disposal, then extending the research to different types of corpus in order to corroborate the findings of [18].

This study also shows that more than just *speech* or *head movements* modalities be used for feedback; in fact all modalities can be combined in one way or another to express feedback.

The last example of the qualitative analysis shows something even more important, that is, not only can all the modalities convey feedback, but the absence of any modality may be a feedback as well, as long as the speaker is expecting some. The qualitative analysis also showed that adaptors can have a communicative purpose and that, when doing so, they can be used as feedback. Those two examples lead us to rethink the annotation scheme for back-channel feedback in particular and the multimodal annotation in general. Indeed, until now, feedback was annotated thanks to clues effectively present in the interaction (i.e. speech, prosody, head movements). This way of annotating feedback, although it is efficient, does not allow for identifying the absence of feedback as negative feedback. Besides, the multimodal studies mostly focus on hand gestures as co-speech gesture and adaptors are often not considered by gesture studies at all. But we have shown that adaptors can have a communicative function when needed. We believe that in this particular setting, when the interaction is very emotional and the interlocutor is producing a lot of adaptors (almost during the whole interaction), analysing them might be very important.

In future work, we will extend our analysis to the other interactions in our corpus taking into account the findings of this study. We will also conduct a functional analysis of back-channel feedback. This future work aims at answering more deeply the question of the distribution of back-channel feedback in doctor/patient interaction while breaking bad news and thus aims at improving our knowledge of such interaction.

5. Acknowledgements

This work was supported by the Labex Brain and Language Research Institute (ANR-11-LABX-0036) and has benefited from support from the French National Agency of Research (ANR), under the project title ‘Investments of the Future’ A*MIDEX (ANR-11-IDEX-0001-02). We would like to thank Daniel Mestre for his helpful suggestions.

6. References

[1] Ptacek, J. T. and Eberhardt, T. L., “Breaking bad news: a review of literature.” Jama, 276.6:496-502, 1996.

[2] Hannawa, A. F., "Disclosing medical errors to patients: effects of nonverbal involvement.", Patient and education counseling, 94.3: 310-313, 2014.

[3] Loi n°2002-303 du 4 mars 2002 relative aux droits des malades et à la qualité du système de santé.

[4] HAS, “Annoncer une mauvaise nouvelle”, 2008.

[5] HAS, “Annonce d’un dommage associé aux soins, guide destiné aux professionnels de santé exerçant en établissement de santé ou en ville”, 2011.

[6] Raper, S. E., Resnik, A. S. and Morris, J. B., “Simulated Disclosure of a Medical Error by Residents: Development of a Cours in Specific Communication Skills”, Journal of Surgical Education, *71*(6):116-126, 2014

[7] Bunt, H. , “The semantics of feedback”, iIn 16th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2012) 118-127, 2012.

[8] Bavelas, J. B., Coates, L., & Johnson, T., “Listener responses as a collaborative process: The role of gaze”, Journal of Communication, 52.3:566-580, 2002.

[9] Prévot, L. and Bertrand, R., “CoFee - Toward a multidimensional analysis of conversational feedback, the case of French language”, Interdisciplinary Workshop on Feedback Behaviors in Dialog, 2012

[10] Colletta, J. M., “Le développement de la parole chez l’enfant âgé de 6 à 11 ans”, Corps, langage et cognition. Sprimont : Pierre Mardaga Editeur, 2004.

[11] Paggio, P., and Navarretta, C., “Feedback in head gestures and speech”, in LREC 2010 Workshop Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, 1-4, 2010.

[12] Allwood, J. and Cerrato, L., “A study of gestural feedback expressions”, First Nordic symposium on multimodal communication, 2003.

[13] Battersby, S. A., and Healey, P. G., “Head and hand movements in the orchestration of dialogue.”, in Thirty-Second Annual Conference of the Cognitive Science Society, 2010.

[14] Boholm, M., and Allwood, J.,“Repeated head movements, their function and relation to speech”, in Proceedings of LREC workshop on multimodal corpora advances in capturing coding and analysing multimodality, 6-10, 2010.

[15] Truong, K. P., Poppe, R., Kok, I. de and Heylen, D., “A multimodal analysis of vocal and visual backchannels in spontaneous dialogs”, Proceedings of INTERSPEECH 2011, International Speech Communication Association, 2973-2976, 2011.

[16] Hadar, U., Steiner, T. J., and Rose, F. C., “Head movement during listening turns in conversation”, Journal of Nonverbal Behavior, 9.4:214-228, 1985.

[17] Bertrand, R., Ferré, G., Blache, P., Espesser, R. and Rauzy, S., “Backchannels revisited from multimodal perspective”, Auditory-visual Speech Processing, 1-5, 2007.

[18] Malisz, Z. and Karpiński, M., “Multimodal aspects of positive and negative responses in Polish task-oriented dialogues”, Speech Prosody 2010 – Fifth International Conference, 2010.

[19] McNeill, D., “Hand and mind: What gestures reveal about thought”, University of Chicago Press, 1992

[20] Boersma, P. and Weenik, D., “Praat: Doing phonetics by computer.[Computer Software] Amsterdam: Department of Language and Literature, University of Amsterdam”, http://www.praat.org/, 2011.

[21] Bigi, B., “SPPAS: a tool for the phonetic segmentations of Speech”, in The eight international conference on Language Resources and Evaluation, ISBN 978-2-9517408-7-7, 1748-175, 2012.

[22] Nijmegen: Max Planck Institute for Psycholinguistics, “ELAN - linguistic annotator. Language archiving technology portal [computer software].” http://www.lat-mpi.eu/tools/elan/, 2011.

[23] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A. and Den, Y. “An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs”, Language and speech, 41(3-4):295-321, 1998.

[24] Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B. and Rauzy, S., “Le CID- Corpus of Interactional Data – Annotation et exploitation multimodale de parole converationnelle”, Traitement Automatique des Langes, 49(3):1-30, 2008.

[25] Ekman, P., & Friesen, W. V., “The repertoire of nonverbal behavior: Categories, origins, usage, and coding.” Semiotica, 1.1:49-98, 1969.

Age-related differences in multi-modal audience design: Young, but not old speakers, adapt speech and gestures to their addressee's knowledge

Louise Schubotz[1,2], *Judith Holler*[1], *Aslı Özyurek*[1,3]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
[2] Max Planck International Research Network on Aging
[3] Radboud University Nijmegen, Nijmegen, the Netherlands
Louise.Schubotz@mpi.nl, Judith.Holler@mpi.nl, Asli.Ozyurek@mpi.nl

Abstract

Speakers can adapt their speech and co-speech gestures for addressees. Here, we investigate whether this ability is modulated by age. Younger and older adults participated in a comic narration task in which one participant (the *speaker*) narrated six short comic stories to another participant (the *addressee*). One half of each story was known to both participants, the other half only to the speaker. Younger but not older speakers used more words and gestures when narrating novel story content as opposed to known content. We discuss cognitive and pragmatic explanations of these findings and relate them to theories of gesture production.

Index Terms: co-speech gesture, aging, audience design, common ground

1. Introduction

The assumption that communicative competence deteriorates with advancing age, due to cognitive or biological decline, is widespread both among younger and older adults, and also in the scientific community, see [1]. Yet, to date little is known about the every-day language use of older adults in personal, face-to-face interactions. Importantly, face-to-face interaction has two inherent features which are frequently overlooked in laboratory investigations of language production: face-to-face language use is a *multi-modal* activity in the sense that it comprises communicative channels beyond the mere speech signal, such as manual co-speech gestures; and it is produced for and targeted at an addressee, shaped by a process called *audience design* [2]. Previous research with younger adults shows that speakers adapt both their speech and their co-speech gestures to an addressee's perceived communicative needs (e.g. [3], [4]). Likewise, addressees are able to perceive, integrate, and interpret information that is presented in these two modalities (e.g. [5], [6]). It is currently unclear whether, and if so, how older adults use these multiple communicative channels when designing utterances for others. However, the findings of previous research presented in the following paragraphs suggest that language produced by older adults may differ systematically from that of younger adults.

1.1 Audience design in speech and co-speech gesture

In younger adults, effects of audience design are frequently investigated by manipulating the amount of conversational *common ground*, i.e. knowledge, beliefs and assumptions that conversational partners believe to be mutually shared and that allows for the appropriate adaptation of utterances [7]. Generally, the more information is shared between interactants, the less needs to be put into words, characterized e.g. by shorter utterances, less complex syntax, or less informational content (e.g. [8], [9]).

Older adults' ability to engage in audience design based on mutually shared knowledge has been addressed in a number of studies employing a director-matcher card game: Participants are required to establish mutual reference to a limited set of objects over the course of several trials, thereby building up *local* or *emerging* common ground ([10], [11], [12]). Younger adults' interactions become increasingly more efficient, indicated by shorter utterances and task-completion times. Older adults are also able to establish common reference, however, they are less efficient, indicated by longer utterances, longer task-completion times, more errors produced, and more idiosyncrasy when compared to younger adults. [10] suggest that this may be due to age-related cognitive limitations, specifically difficulties in retrieving partner-specific information from memory.

Findings from studies manipulating global addressee characteristics such as age ([13], [14]) or mental retardation [15] suggest that older adults are able to adapt their speech based on these *a priori* or *global* aspects of common ground and do not differ significantly from younger adults. Arguably, memory demands are much lower in these paradigms as opposed to the director-matcher tasks, which may account for older adults' better performance here.

One shortcoming of nearly all of these studies is that they ignore the multi-modal character of face-to-face language use.[1] Yet, information conveyed in the visual domain is essential to face-to-face interaction. Especially *representational* co-speech gestures, i.e. gestures that depict information imagistically, contribute to the semantic content of a message and are sensitive to social context variables. For example, speakers can produce representational gestures to clarify verbal ambiguity for their addressee [16], and representational gesture rate (i.e. the number of gestures produced per 100 words) is sensitive to visibility between speaker and addressee (e.g. [17], [18]), as well as to dialogue and addressee feedback (e.g. [19], [20]), suggesting that speakers take their addressee's communicative needs and abilities into account when designing multi-modal utterances.

Studies investigating the effect of common ground on gesture production often obtain effects that parallel the findings for speech. For example [4] used a cartoon narration task in which a speaker narrated one story three times, first to a naïve addressee, and then again to either the same addressee (common ground) or a to different addressee (no common ground). In second narrations, speakers produced significantly fewer, smaller, and less precise representational gestures for same addressees than for different addressees. Using a similar paradigm, [21], Exp. 1, also found that speakers produce fewer representational gestures when narrating the same comic story three times to either the same addressee or an addressee who could also see the story (common ground) as opposed to addressees who were not familiar with the story (no common ground), again indicated by a decrease in gesture rate. Similar effects for common ground on gesture rate or quality have been obtained by e.g. [22], [23], and [24] amongst others. However, others have found no effects of common ground on

[1] With the exception of [12], who take eye-gaze into consideration.

gesture rate ([25], [26], [27]), or even opposite effects, such that participants gesture more in relation to speech when common ground was present ([28], [29]).

1.2 Co-speech gesture production in older adults

Research on co-speech gesture production in older adults is to date limited to the two studies summarized in the following. [30] asked younger and older women to describe four physically co-present objects to a video camera and found that older women produced representational co-speech gestures at a significantly lower frequency than younger women. The authors explain this significant age difference by referring to the idea that older adults are less involved with mental imagery during speaking. This assumption was explicitly tested by [31]. In their study, younger and older participants responded to questions thought to evoke mental imagery (visual and motor). Again, older adults produced representational gestures at a significantly lower rate (computed as number of gestures per five-second time window) than the younger experimental group, but the imagery content of their speech was comparable to that of younger adults. Hence, a lack of mental imagery seems to be an unlikely explanation for older adults' decreased use of representational gestures. Rather, [31] argue that older adults prefer simpler gestural forms when facing the task of speaking and gesturing concurrently, possibly due to cognitive limitations, although the authors do not elaborate on this issue further.

However, neither of these studies used an interactive paradigm in which an addressee's knowledge state must be taken into account for successful communication. It is therefore unclear how older adults use speech and co-speech gestures concurrently in these types of situations.

1.3 The present study

The main aim of our research was to find out whether, and if so, how older adults adapt their speech and their gestures to mutually shared knowledge between speaker and addressee. In order to investigate this, we designed a comic narration task in which a primary participant (the *speaker*) narrates six short comic strips to a secondary participant (the *addressee*) who would then answer a question based on this narration. Common ground was manipulated by showing both participants one half of each strip (either the first or the second half) at the beginning of each trial. Only the speaker would subsequently see the full story, meaning that one half of the story content was mutually known to both participants (common ground or CG), while the other half was only known to the speaker (no common ground or no-CG).

Our two main dependent measures were number of words and the gesture rate per condition/narration. In line with previous findings, we expected an effect of our common ground manipulation on speech production such that younger adults would use fewer words when narrating shared story content and more words when narrating novel content (e.g. [3]). Based on the results obtained by [11], [12], and [13] we expected this effect to be smaller in older adults. Given the mixed findings in the gesture literature on common ground, gesture rates could decrease (e.g. [4], [21]), stay constant (e.g. [25]), or increase (e.g. [28]) with an increase in common ground. In analogy to our predictions for an age effect on speech, we did hypothesize that if there is an effect for younger adults on gesture production, this should be smaller for older adults. Also, we expected older adults to produce fewer representational gestures overall, in line with [30] and [31].

2. Method

2.1 Participants

64 participants took part in the study, 32 younger adults (16 women) between 21 and 30 years old (*M* = 24.31 years, *SD* = 2.91 years) and 32 older adults (16 women) between 64 and 73 years old (*M* = 67.69 years, *SD* = 2.43 years). All participants were recruited from the participant pool of the Max Planck Institute for Psycholinguistics and received between € 8 and € 16 for their participation, depending on the duration of the session. All participants were native Dutch speakers with self-reported normal or corrected-to-normal vision and hearing.

2.3 Design

We employed a 2 x 2 design, with the between-participant variable age (young vs. old), and the within-participant variable common ground (CG vs. no-CG).

2.3 Materials

Seven black-and-white comic strips from the series "Vater und Sohn" were used to elicit narratives. Each strip consisted of a self-contained story, either four or six frames long, and centered around a father and a son and their activities.

2.4 Procedure

Participants came to the lab in pairs. We tested same age and same sex pairs only. The role of speaker and addressee were pre-assigned randomly and kept constant across the entire experiment. Upon arrival, speaker and addressee were asked to sit in designated chairs at a table at 90° from each other. Two video cameras were set up on tripods at a small distance from the table, one of them getting a full frontal view of the speaker, and the other one positioned such that it captured both speaker and addressee (see Figure 1 for a still from the second camera). Sound was recorded with an additional microphone suspended from the ceiling over the table and connected to the speaker camera.

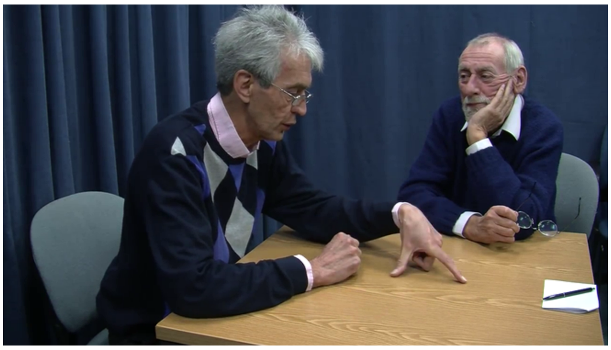


Figure 1. *Speaker (left) and addressee (right) seated at the table.*

Participants were introduced to each other and received a description of the experiment. This and all subsequent instructions were given both in writing and verbally to ensure that all participants received and understood the necessary information to successfully participate in the experiment. Signed consent was acquired from all participants. Before the start of the actual experimental sessions, participants played a warm-up game to get familiar with each other as well as the experimental set-up. Following the warm-up game, the experiment continued with one of two experimental tasks: a comic narration task (present experiment) and a building block task (reported elsewhere), the order of the two tasks was counterbalanced across dyads.

All participants completed one practice trial and six experimental trials, narrating a total of seven stories. At the beginning of each trial, both participants were presented with either the first or the second half of the comic strip (counterbalanced across the experiment) and were instructed to look at it together for 10 seconds without talking. Eight experimental lists determined the order of story presentation. Each list was tested four times, once for each age/sex pair. Subsequently, the drawings were removed and a screen was put up on the table between speaker and addressee. The speaker then received the full story to look at. Once the speaker signaled that she had understood and memorized the story, both drawings and screen were removed again and the speaker narrated the entire story to the addressee. The speaker was instructed to narrate the full story, keeping in mind that the addressee had already seen part of it. Addressees were instructed to listen to the narrations and ask all clarification questions at the end. Then the screen was put back up and the addressee answered a question about the content of the story in writing. Depending on the dyad, the task took about 20 to 30 minutes.

2.5 Transcription and coding

All recordings from the two cameras were synchronized and subsequently segmented into trials. Transcription of speech and annotation of gestures was done in Elan [32]. For all segments, the speaker's initial narration was identified. All analyses reported here are based on these initial narrations only, discarding repetitions or clarifications elicited by the addressee. Speech from the speaker was transcribed verbatim, including disfluencies such as filled pauses and word fragments. However, these disfluencies are excluded from the word counts presented in the results section. We also distinguished between speech belonging to the narrative proper, i.e. relating story content, and non-narrative speech such as statements about the task or comments relating to the speaker or the addressee. Among the non-narrative speech, we identified explicit references to common ground, i.e. statements such as "this time we saw the first half together".

For the gesture coding, we first identified all co-speech gestures produced by the speaker and accompanying narrative speech, disregarding irrelevant movements that were not gestures as well as gestures accompanying non-narrative speech. We then categorized these gestures according to their function. Globally, we distinguished between *representational* and *non-representational* gestures (see [17]).

For our purposes, representational gestures include *iconic* gestures, which depict shape or size of concrete referents or represent specific physical movements or actions; [2] furthermore *metaphoric* gestures, which relate to speech in a more abstract manner; and finally pointing gestures or *deictics*. We distinguished between *concrete deictics*, i.e. finger points to a physically co-present referent, and *abstract deictics*, i.e. finger points to a specific location in space, e.g. that of a story character.

All other gestures were considered non-representational and include what are frequently called *beat* gestures, i.e. biphasic movements of the hand e.g. to add emphasis, furthermore *interactive* gestures relating to the structuring of the conversation. As non-representational co-speech gestures occurred very infrequently and were not the primary interest of the present study, we decided to not investigate them further here.

A second coder blind to the experimental hypotheses coded 10% of the trials. Inter-rater agreement on stroke identification was 92.3%. Inter-rater agreement on gesture categorization was 97.9%, Cohen's Kappa = .949.

To normalize for differences in speech rate, we computed the gesture rate as the number of gestures per 100 words.

3. Results

Table 1 lists the mean values and standard deviations for the various measures by age group and condition. We first computed an average value per participant and condition, and then computed means and standard deviations based on these averages. Like words and gesture rate, explicit reference to common ground was computed as the average number of explicit references made across trials per condition. For all analyses, we performed a 2 (age: old vs. young) x 2 (common ground: CG vs. no-CG) ANOVA as well as pair-wise comparisons using t-tests or, where applicable, Wilcoxon tests, in combination with Bonferroni corrections. All p-values are two-tailed unless clearly stated otherwise.

Table 1. *Means and standard deviations of various measures per trial for age groups and conditions.*

	CG		no-CG	
	Young	Old	Young	Old
Narrative words	44.13 (21.37)	52.39 (12.45)	65.59 (20.23)	54.59 (12.47)
Representational gesture rate	5.67 (4.28)	5.95 (4.01)	7.62 (3.92)	4.86 (3.72)
Non-rep. gesture rate	1.91 (1.93)	1.25 (1.12)	1.84 (1.74)	1.08 (1.02)
CG reference	.72 (.59)	.11 (.23)	.03 (.09)	0

3.1 Speech

3.1.1 Narrative words

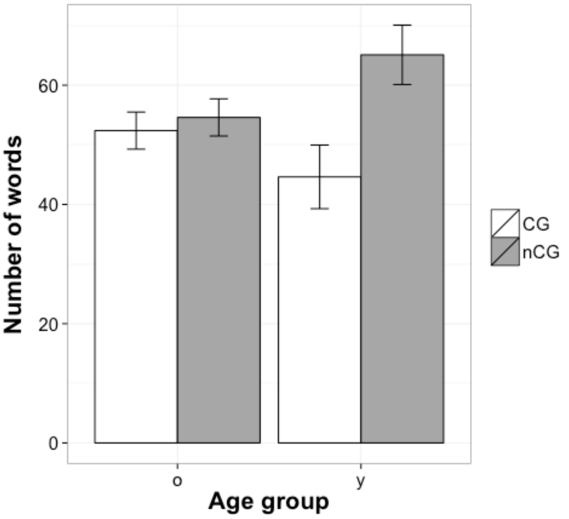


Figure 2. *Average number of narrative words per age group and common ground condition. Error bars represent the SE.*

Figure 2 shows the average number of narrative words per age group and common ground condition. The results of the ANOVA yielded a significant interaction of age by common ground, *F(1,60) = 5.043, p = .028*. The main effect of common ground was also significant, with participants producing more words in the no-CG condition than in the CG condition, *F(1,60) = 7.621, p = .008*, but the main effect of age was not, *F(1,60) = .102, p = .75*. To explore this

[2] Note that "re-enactments", i.e. movements of the body that represented specific actions of the stories' characters, were also coded as iconic gestures, even if they did not include manual movements.

interaction further, we calculated four pairwise comparisons, adopting a Bonferroni-corrected alpha level of .0125. First, we compared the number of narrative words per common ground condition within age groups using one-tailed paired t-tests. As predicted, young adults used significantly fewer words to describe CG content vs. no-CG content, *t(15) = 4.852, p < .001*. For old adults, the difference between CG and no-CG was not significant, *t(15) = .746, p = .23*. We then compared the average number of narrative words used per common ground condition across age groups. Young and old adults did not differ significantly in the number of words used to describe CG content, *t(24.145) = 1.256, p = .221*, or no-CG content, *t(25.186) = 1.786, p = .086*.

3.1.2 Explicit reference to common ground

As the data were not normally distributed, we used Wilcoxon rank sum and signed rank tests to explore the differences between age groups and conditions in the explicit reference to common ground. Young adults made significantly more explicit references to common ground in the CG condition than in the no-CG condition, *Z = 3.189, p = .001*. Also, young adults used significantly more explicit references than old adults across conditions, *Z = 2.903, p = .003*. None of the other pairwise comparisons were significant (all *p*'s > *.05*).

3.2 Representational co-speech gesture

Figure 3 shows the average representational gesture rate per age group and common ground condition. The ANOVA revealed no significant main effects of common ground, *F(1,60) = .19, p = .66*, or age, *F(1,60) = .1.554, p = .21*, and no significant interaction, *F(1,60) = 2.342, p = .13*. Since we tested specific hypotheses, we computed four pairwise comparisons, adopting an alpha level of .0125 throughout. Young adults used significantly more representational gestures in the no-CG vs. CG condition, *t(15) = 4.136, p < .001*. For old adults, the trend goes in the opposite direction, however, this difference was not significant, *t(15) = 1.981, p = .06*. The comparison of age groups within conditions also did not yield significant differences (both *p*'s > *.05*).

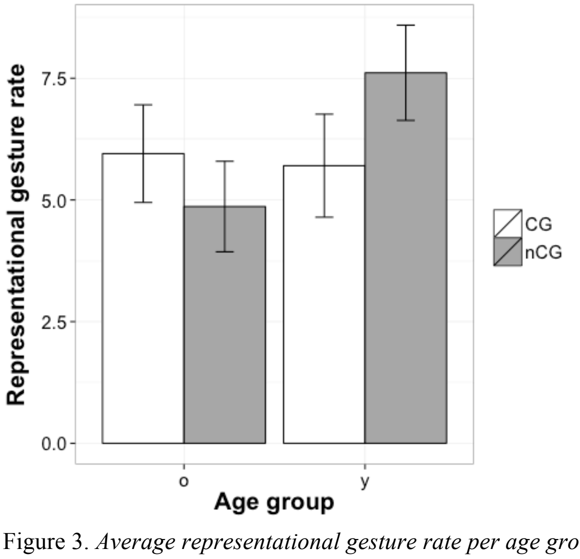


Figure 3. *Average representational gesture rate per age group and common ground condition. Error bars represent the SE.*

4. Discussion

We investigated how younger and older adults adapt their speech and co-speech gestures to an addressee's knowledge state when narrating short comic strips. Younger, but not older adults produced more words and more representational co-speech gestures when relating content that was novel to the addressee. The individual results will be discussed in more detail in the following.

The expected effect of common ground on speech, i.e. fewer words to narrate known story content and more words to narrate novel content was only significantly present in younger adults. For the younger adults, this is in line with some previous findings on speech and common ground in similar narration tasks (e.g. [3], [28]), supporting the idea that the more knowledge interactants assumed to be mutually shared, the faster and more efficient their communication gets. The fact that younger adults frequently referred to the common ground explicitly when relating familiar content, e.g. by stating "you've already seen the first half so I'll go through it quickly" clearly indicates that they were aware of their addressee's knowledge state. Hence we can safely assume that our manipulation of common ground worked as we intended. Older adults, on the other hand, hardly differed in the number of words they used to narrate familiar vs. novel story content, and made very few explicit references to common ground. Two explanations are conceivable: Older adults may not be able to engage in audience design based on common ground as we induced it here due to cognitive factors, but they may also have different communicative goals than younger adults, as laid out in the following to paragraphs.

From a cognitive perspective, it may be that older adults simply do not remember what does and what does not constitute common ground, i.e., in the present task, which half of the story they had inspected together with the addressee at the beginning of the trial.[3] Alternatively, they may still remember which information is mutually shared between them and the addressee, but then are unable to use this knowledge when designing their utterances, potentially due to a failure to retrieve the relevant information in time in order to plan the utterance accordingly (as suggested by [10]). Remembering which knowledge is in common ground and designing one's utterances accordingly is arguably more challenging than taking global addressee features like addressee age into account, a task which older adults have been shown to be able to do successfully (e.g. [13], [14]). Still, the small difference between the number of words used to narrate familiar vs. novel story content for the older adults is surprising, given that previous research using quite complex manipulations of emerging common ground did find an effect ([10], [11], [12]).

One should therefore also consider the possibility that older adults' communicative goals differ from those of younger adults (see also [13]). Older adults may *choose* to give equal weight to both known and unknown story content in their narrations. Whereas young adults may have the goal to enable their addressee to correctly answer the question he would receive after the narration, focusing on providing information that the addressee does not yet have, older adults may have the primary goal of narrating "a nice story". For example [33] found that older adults are judged to be better at story telling than younger adults. We did not obtain objective ratings of narration quality, but the first author's personal impression was that older adults, more so than younger adults, largely enjoyed the task, putting considerable effort into narrating the stories in an entertaining and animated way. Older adults frequently added additional material to their story, e.g. attributing intentions and feelings to the characters, whereas younger adults were more likely to include information on smaller, visual details of the individual frames which they thought might be relevant to answering the

[3] Unfortunately, we did not assess whether participants remembered which part they had inspected together at the end of the task, so this interpretation remains speculative.

question. Obviously, both cognitive *and* pragmatic factors may influence how younger and older adults speak for an addressee.

There were no age differences for representational gesture rate. This is contrary to previous findings by [30] and [31], which suggest that older adults use significantly fewer representational gestures than younger adults. We propose that this is due to the more communicative design we employed here. Whereas participants in the two previous studies either had no addressee at all or an experimenter-addressee, in the present study we used naïve, real addressees. Research with younger adults indicates that the presence of a visible, attentive addressee increases the production of representational gestures (e.g. [21], [34]). This suggests that, given the appropriate context, older adults have sufficient cognitive capacities to produce potentially complex gestural forms concurrently with speech. It should also be noted that the older adults in our sample were a little younger ($M = 67.69$ years) than in [30] and [31], where the mean age was about 70 years.

Crucially, with respect to the hypotheses we set out to test, older adults' representational gesture production was not sensitive to their addressee's knowledge state. Whereas younger adults produced representational gestures at a higher rate when relating novel as compared to mutually shared content, this was not the case for older adults. Our finding for younger adults replicates some of the earlier findings on common ground and gesture production in studies using comparable tasks ([4], [21]). The fact that representational gesture rate is influenced by contextual factors such as mutually shared knowledge between a speaker and an addressee lends support to views of gestures as communicatively motivated [35]. Additionally, it is in line with accounts of gesture production claiming that speech and gesture are part of a single, integrated system [36] in which both modalities tightly interact with each other and information conveyed in gesture is semantically coordinated with information conveyed in speech (Interface Hypothesis, [37]). In the current study, we found that an increase in information conveyed in the spoken modality is coupled with an increase in information conveyed in the gestural modality. The idea that "more speech goes with more gesture, less speech with less gesture" ([26], p. 243) is expressed in the "hand-in-hand" hypothesis of gesture production as formulated by [38] who propose that speech and gesture behave in a parallel fashion. Although in our case, more speech goes with *even* more gesture, our findings support the notion of a parallel increase in both modalities.

Thus considered, it is also not surprising that older adults' representational gesture rate is not influenced by the presence of common ground. As they show no sign of audience design in their speech, why should they do so in gesture? The same cognitive and/or pragmatic factors that influence older adults' verbal behavior may also influence their gestural behavior, again underlining the tight parallel between the two modes of communication.

5. Conclusions

The results of the present study suggest that there is an age-related difference in how speakers adapt their speech and co-speech gestures based on mutually shared knowledge with an addressee. Younger, but not older adults, convey more information both in their speech *and* in their gestures when there is common ground as opposed to when there is not. Whether these differences in verbal and gestural behavior have an impact on how older adults are comprehended by others, and on the overall quality of their interactions remains to be investigated.

6. Acknowledgements

LS was supported by a doctoral stipend by the Max Planck International Research Network on Aging. JH was supported through ERC advanced grant INTERACT #269484. We would also like to thank the MPI for Psycholinguistics, Nijmegen in general, for providing funding for participant payments, and Nick Wood and Peter Nijland in particular, for extensive support with the video editing and with Elan. Additionally, we would like to thank two anonymous reviewers for their comments on an earlier version of this paper.

7. References

[1] Thornton, R. and Light, L.L. (2006). Language comprehension and production in normal aging. In J. E. Birren and K. W. Schaie (Eds.). *Handbook of the Psychology of Aging* (6th Ed.) (pp. 261-287). San Diego, CA: Elsevier Academic Press.

[2] Clark, H.H. and Murphy, G.L. (1983). Audience design in meaning and reference. In J.F. LeNy and W. Kintsch (Eds.), *Language and comprehension*, pp. 287-299. Amsterdam: North-Holland Publishing Co.

[3] Galati, A. and Brennan, S. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62, 35-51.

[4] Galati, A. and Brennan, S. (2014). Speakers adapt gestures to addressees' knowledge: implications for models of co-speech gesture. *Language, Cognition and Neuroscience*, 29(4), 435-451.

[5] Kelly, S. D., D. J. Barr, R. B. Church, & K. Lynch (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, 40: 577-592.

[6] Kelly, S.D., Özyürek, A. & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, 21(2), 260-267.

[7] Clark, H.H. (1996). *Using Language*. Cambridge University Press, Cambridge.

[8] Fussel, S.R. and Krauss, R.M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology*, 62, 378-391.

[9] Isaacs, E.A. and Clark, H.H. (1987). References in conversations between experts and novices. *Journal of Experimental Psychology: General*, 116, 26-37.

[10] Horton, W. S. and Spieler, D. H. (2007). Age-related differences in communication and audience design. *Psychology and Aging*, 22(2), 281-290.

[11] Hupet, M., Chantraine, Y., and Nef, F. (1993). References in conversation between young and old normal adults. *Psychology and Aging*, 8(3), 339-346.

[12] Lysander, K. and Horton, W. S. (2012). Conversational grounding in younger and older adults: The effect of partner visibility and referent abstractness in task-oriented dialogue. *Discourse Processes*, 49(1), 29-60.

[13] Adams, C., Smith, M.C., Pasupathi, M., and Vitolo, L. (2002). Social context effects on story recall in older and younger women: Does the listener make a difference? *Journal of Gerontology: Psychological Sciences*, 57B(1), 28-40.

[14] Keller-Cohen, D. (2014). Audience design and social relations in aging. *Research on Aging*, 1-22.

[15] Gould, O.N. and Shaleen, L. (1999). Collaboration with divers partners: How older women adapt their speech. *Journal of Language and Social Psychology*, 18, 395-418.

[16] Holler, J. & G. Beattie (2003). Pragmatic aspects of representational gestures: Do speakers use them clarify verbal ambiguity for the listener? *Gesture*, 3(2), 127-154.

[17] Alibali, M. W., Heath, D.C., & Myers, H.J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language*, 44, 169-188.

[18] Mol, L., Krahmer, E., Maes, A. & Swerts, M. (2011). Seeing and Being Seen: The effects on gesture production. Journal of Computer-Mediated Communication 17(1), 77-100.

[19] Bavelas, J., Gerwing, J., Sutton, C., and Provost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58, 495-520.

[20] Holler, J., & Wilkin, K. (2011). An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. Journal of Pragmatics, 43, 3522-3536.

[21] Jacobs, N. and Garnham, A. (2007). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language*, 56, 291-303.

[22] Gerwing, J. and Bavelas, J. (2004). Linguistic influences on gesture's form. *Gesture*, 4(2), 157-195.

[23] Holler, J. & R. Stevens (2007). The effect of common ground on how speakers use gesture and speech to represent size information. *Journal of Language and Social Psychology*, 26(1): 4-27.

[24] Parrill, F. (2010). The hands are part of the package: Gesture, Common ground and information packaging. In S. Rice and J. Newman (Eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*, pp. 285-302.

[25] Campisi, E., & Ozyurek, A. (2013). Iconicity as a communicative strategy: Recipient design in multimodal demonstrations for adults and children. Journal of Pragmatics, 47, 14-27.

[26] de Ruiter, J.P., Bangerter, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, 4(2), 232-248.

[27] Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., and Swerts, M. (2015). Reduction in gesture during the production of repeated references. *Journal of Memory and Language*, 79-80, 1-17.

[28] Holler, J. and Wilkin, K. (2009). Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task. *Language and Cognitive Processes*, 24(2), 267-289.

[29] Holler, J., Tutton, M. and Wilkin, K. (2011). Co-speech gestures in the process of meaning coordination. In *Paper presented at the 2nd GESPIN – Gesture and Speech in Interaction Conference*, Bielefeld.

[30] Cohen, R. L. and Borsoi, D. (1996). The role of gestures in description-communication: A cross-sectional study of aging. *Journal of Nonverbal Behavior*, 20(1), 45-63.

[31] Feyereisen, P. and Havard, I. (1999). Mental imagery and production of hand gestures while speaking in younger and older adults. *Journal of Nonverbal Behavior*, 23(2), 153-171. [32] Wittenburg & al., 2006

[33] James, L.E., Burke, D.M., Austin, A., and Hulme, E. (1999). Production and perception of "verbosity" in younger and older adults. *Psychology and Aging*, 13, 355-367.

[34] Kuhlen, A. K., Galati, A., & Brennan, S. E. (2012). Gesturing integrates top-down and bottom-up information: Effects of speakers' expectations and addressees' feedback. *Language and Cognition*, *4*, 17-41.

[35] Melinger, A. & W. J. M. Levelt (2004). Gesture and the communicative intention of the speaker. *Gesture*, 4(2): 119-141.

[36] McNeill, D. (1992). *Hand and mind*. Chicago, London: The Chicago University Press.

[37] Kita, S. & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48: 16-32.

[38] So, W.C., Kita, S., and Goldin-Meadow, S. (2009). Using the hands to identify who is doing what to whom: Gesture and speech go hand-in-hand. *Cognitive Science*, 33, 115-125.

Ostensive signals: markers of communicative relevance of gesture during multimodal demonstrations to adults and children

Anita Slonimska [1], *Asli Özyürek* [1,2], *Emanuela Campisi* [3]

[1] Radboud University, Nijmegen, The Netherlands
[2] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[3] Lund University, Lund, Sweden

anita.slonimska@mpi.n, asli.ozyurek@mpi.nl, emanuelacampisi82@gmail.com

Abstract

Speakers adapt their speech and gestures in various ways for their audience. We investigated further whether they use ostensive signals (eye gaze, ostensive speech (e.g. *like this, this*) or a combination of both) in relation to their gestures when talking to different addressees, i.e., to another adult or a child in a multimodal demonstration task. While adults used more eye gaze towards their gestures with other adults than with children, they were more likely to use combined ostensive signals for children than for adults. Thus speakers mark the communicative relevance of their gestures with different types of ostensive signals and by taking different types of addressees into account.

Index Terms: gesture, recipient design, ostensive signals, eye gaze, ostensive speech

1. Introduction

In a face-to-face communication, the visual modality has provided us with communicative tools like our hands, facial expressions and eye gaze, in addition to the verbal modality. The final message that we produce is the interplay of all of these and other communicative channels. This implies that we are free to manipulate them and encode our message according to our preference or necessity [1]. It also implies that we are likely to accentuate the core information and attend to the modality in which it is present or more relevant [2]. In other words, considering that we are capable of shifting information from one channel to another and also employ different channels simultaneously, we are able to accentuate the relevant information wherever it might be encoded.

Even though hand gestures accompanying speech have been intensively studied in the context of communication and research has provided extensive support for its communicative function [3,7,8,9 among many], almost nothing is known concerning the circumstances under which speakers foreground information in their gestures. For example, it has been proposed that gesture use is designed differently for children in comparison to adults in order to enhance the understanding of the message [3,4]. Very little is known, however, about why and how speakers make their gestures communicative or, more precisely, alter the level of their communicativeness. Within this study we therefore set out to unravel the use of ostensive signals, namely eye gaze and/or ostensive speech (e.g. *like this, this*), relating to gesture in face-to-face communication, and investigate whether it is implemented as a strategy to highlight the communicative relevance of the information expressed by gesture. Secondly, we are interested in the contexts where such highlighting of gesture occurs and how this highlighting is achieved.

Accordingly, we aim to answer the following questions:

- What are the patterns of the ostensive signals (eye gaze and/or ostensive speech) used to emphasize information in gestures during a demonstration task?
- Do these patterns change in certain communicative contexts (e.g. demonstration to a child or an adult)?

2. Background

Addressees tend to pick up the most relevant stimuli from the speaker in order to process a message and understand its meaning [5]. Accordingly, speakers are likely to produce such stimuli that will be relevant to the addressee and that will help the addressee in deriving the meaning of what is expressed more efficiently. Such inferential communication is rooted in an attempt to make the addressee recognize that the speaker has intended to affect the state of her knowledge by manifesting such intention. Therefore, this communication is not merely inferential but also ostensive. It results in a communicative act having two intentions - informative and communicative. A speaker's *informative intention* is the intention to inform the addressee about something. A *communicative intention*, in turn, informs the addressee that the speaker intends to provide the information. Such communicative intention can be realized through *ostensive signals* [6]. By means of the ostensive signals the speaker invites the addressee to attend to what she has referred to and by doing so she informs the addressee that the particular piece of information is relevant in the processing of the meaning of the message. Ostensive signals go beyond the verbal domain of lexical/semantic information, word order, morphology and prosody, manifesting themselves also visually [2, 5, 7, 8].

It is important to note that in face-to-face interaction a variety of ostensive signals combine or compete in a discourse. The role of the speaker is a crucial one in selecting the most adequate signal in order to manifest her communicative intention [7]. Therefore, the speaker has to assess the context of the interaction and decide to implement a particular communicative strategy (recruiting multi-modal signals) from which the addressee would benefit the most, allowing the addressee to process the message more efficiently. In other words, recipient-design plays an important role in the decision of how to encode the message [9].

Regarding the age of the recipient, the importance of ostensive signals in communication has been investigated mostly in relation to very young children (infants and toddlers) and the way mothers change their communicative patterns. Studies have shown that infants become aware of the ostensive signals at an early stage (during the first year of their life) and react to the communicative nature of the message when it is transmitted not only through the verbal modality like child-directed speech or naming [10,11,12], but also eye gaze [11,12,13], object exchange [11], object demonstrations, object displays and pointing [4]. Even though picking up the communicative signals to process the message does not end at infancy and is not limited to input of the mothers alone, to date there is only one study that looked into multi-modal communicative strategy implemented towards school-age children (12 years old) in comparison to adults.

[3] looked into differences in gesture use according to the age of the addressee – child or adult. Participants in the study did not have an actual addressee present but they had to imagine the addressee in order to describe an action (a child, an adult who knows how to prepare coffee (*expert*) and adult who does not know how to prepare coffee (*novice*)). Therefore, differently from experiments with very young children, where the object is usually manipulated (displayed, demonstrated or exchanged) or at least is present during interaction, in this study speakers had to bear in mind that addressees would have to interpret the message based solely on a mental representation. This study showed that speakers tended to use more iconic gestures with children rather than with adults who knew how to prepare coffee. This result shows that iconic gestures might have been implemented as an informative tool in order to provide a clearer message. A follow-up study showed that iconic gestures addressed to the children were considered more informative and bigger than gestures in both adult conditions, thus suggesting that implementing iconic gesture might serve as a strategy to improve the effectiveness of the message for children [3]. Such an assumption seems plausible if we take into account that integrating gesture and speech makes the interpretation of the message easier for very young children [4] and therefore, this aspect of ease might stand for older children as well.

It has been proposed that gesture is a communicative strategy used in order to determine a discourse referent [8]. Unlike linguistic signals that compete between each other in order to manifest the relevant information, gesture and linguistic signals tend to combine. It was presumed that these signals may cluster to help to identify the referent and the more clustering there is the more likely that the referent is determined [8]. Considering that gesture may be seen as a bridge that combines world knowledge and language, in the sense that it encodes world knowledge visually and relates it to the co-occurring verbal expression, it serves as a strong signal to the information that was expressed by the speaker. In such contexts, gesture serves as an ostensive signal to the co-occurring speech. According to [8], the main idea is that the core information is contained in speech - which is the informative intent of the speaker - while gesture functions as intensifier, a communicative intent by letting the addressee know that the information contained in the speech is relevant [8].

Although there has been some interest in investigating ostensive signals that lead to attending to gesture from the perception point of view [7,14], to our knowledge there is no research that centers exactly on production of those signals (except [16] who investigated how speakers attend to their gestures after the feedback from the addressee). However, without understanding how and why speakers attend to their gestures, research on perception of these gestures and signals renders such studies somewhat incomplete. In fact, in order to make a judgment on why addressees attend to particularly highlighted gestures, it is first necessary to find the answers on how this highlighting occurs. According to the literature, the main ostensive signals to gesture are eye gaze and ostensive speech [7,14, 2].

[2] points out that gestures are not solely internal conceptualizations, but also images that can be perceived as material objects and therefore pointed to by means of an eye gaze. The hand creating a gesture becomes a representational artifact that is meant to be seen by others. Moreover, by attending with eye gaze to the self-produced gesture, the speaker signals that the gesture is meant to provide information. The main point made by [2] is that mere research of speech and gesture is not entirely reliable in order to investigate the function of gesture. In fact, there is the third aspect, eye gaze that has to be taken into account when making any inferences about intended communicativeness of the gesture. [14] presents results indicating that addressees fixate very few gestures produced by the speaker; they are, however, more likely to fixate the gestures that the speakers gazed at themselves, and therefore they conclude that eye gaze serves as an effective attention drawing device for the addressees.

Eye gaze is not the only tool used by the speakers to direct attention of the addressees to their gestures. For example, [2] refers to the ostensive speech (demonstratives e.g. *like this, that*) as a signal to the relevance of gesture in conversation. When a speaker ostensively refers to a gesture, this becomes the core of the utterance because ostensive speech is not providing any concrete information but is only pointing to the gesture as the one that possesses it [2]. Also, [16] report that participants used ostensive speech significantly more when they were requested to elaborate on the information they had provided before, which signals that ostensive speech, indeed, is implemented as a tool to direct attention toward the gesture in order to provide a clearer message.

As stated above, a communicative act requires two intentions – the informative intention, which is the content of the message, and the communicative intention, which is a manifestation of willingness to transmit the message. According to the reviewed literature, it is plausible to suggest that, if gesture is communicatively intended in the sense that it possesses relevant informative value, it should be manifested through ostensive signals like the eye gaze (2, 7, 14] and/or the ostensive speech [2, 4]. The gesture might serve not only as the ostensive signal to the co-occurring speech in order to determine the referent [7, 8], but as the content of the information in its own, which is manifested by means of ostensive signals. Thus, if the speakers attend to their own gestures, they mark the relevance of the information contained in them and not only the information contained in the speech.

The purpose of this study is to extend existing research on the communicative function of representational gestures (gestures that have a semantic relation to their referent) from the encoding point of view. Furthermore, it aims to be the first systematic study to stress the role of ostensive signals in the light of gesture's relevance to the encoding of the message. We reviewed the study of [3] who stated that using iconic gesture might be used as a strategy with children to render the message during a demonstration more comprehensible. Thus we predict that:

- Speakers attend more (by means of eye gaze and/or ostensive speech) to their gestures with children than adults during a demonstration task.

We make no predictions about which type of the ostensive signals will be used, nor about the preference for one ostensive signal over another due to the lack of literature on which to base such predictions.

3. Method

3.1. Participants

Forty-eight right-handed Italians, born and raised in Sicily, participated in the study. Thirty-two participants were undergraduate students, ranging from 20 to 30 years. Sixteen participants were school-age children, ranging from 9 to 10 years. None of the adult participants had experience with children (they had no children, no very young siblings, no teaching experience). None of the participants knew each other before the experiment. All of the participants were informed about presence of the cameras before the recording and gave their written consent to the use of recorded material.

3.2. Material

The material for the study was a game called „Camelot“. The idea of the game is to create a path through which a prince, who is located on one tower, can arrive at the princess, who is located on the other tower. The game consists of wooden blocks that have to be put together (either in line horizontally on the wooden plate with 6 spaces for each block, or vertically with a block on another block) to create a path without gaps from one tower to the other, which are situated at the extremes of a wooden plate. Some of the blocks have the shape of a stair, which is the difficult part of the game. Player’s task is to understand how to put these stairs appropriately in order to create a path without gaps. This game was chosen to provide speakers with stimuli to recruit gesture use, as the rules, which the speaker has to explain, require a lot of form, location and motion explanations.

3.3. Design

A within-subject design was used. Sixteen randomly chosen Italian adult participants (12 female, 7 male) were assigned the role of the *speakers* in the study, while another 16 adults and 16 children had a role of the *addressees,* which represented two conditions (adult and child). The order of the conditions was counter balanced. Only *speakers* were analyzed for the present study.

3.4. Procedure

The speakers were introduced to the rules of the game by the experimenter who presented them in written form. After, the speakers had to complete the game on their own to be able to explain it to the addressees at a later time.

The addressees were asked not to ask questions during speaker‘s description and they were informed that they could ask them, if there were any, once the speaker has finished speaking. The reasoning behind this arrangement was that dialog might affect gesture use frequency [15] and addressee’s feedback might result in changes of the frequency with which speakers gazed at their gestures [16]. It is important to note, that even though such preventives secured a lower chance of verbal feedback, it could not limit non-verbal feedback of the addressees (e.g., face expressions). At the end of the session the addressees were asked to fill in a questionnaire and answer the question *how difficult was it to understand the speaker's descriptions?* by using Likert scale responses from 1 to 5 (1= *very easy*, 5= *very difficult*).

Before the experiment, each speaker had a warming-up session with both addressees (one at a time), during which they had small talk on random topics. The speaker and the addressee were seated at adjacent sides of the table (squared shape) and in chairs without armrests.

Data were recorded on two cameras from two different angles. The first camera recorded a frontal, diagonal view, so both participants were visible. The second camera recorded from the top down and to the left of the speaker; this view covered the whole surface of the table.

3.5. Coding

Data were coded for speech and gesture, as well as for ostensive signals investigated in the present study, which are eye-gaze and ostensive speech. All the data were coded using the video annotation software ELAN [21], developed at the Max Planck Institute for Psycholinguistics.

3.5.1. Speech

The speech of the speaker was transcribed and divided in words [3]. Then it was coded for *description* and *question/answer* segments. Few instances occurred where the addressees interrupted the speakers while they were describing the rules (in total 5 occurrences in adult condition, and 6 occurrences in child condition). The answers to the question and interruptions of the addressee were excluded from the analysis.

3.5.2. Gesture

All gesture strokes during the description were coded, where the stroke is considered the part of the gesture that conveys the most meaningful part and requires the most effort [17, 18]. In this study, the focus is put on representational gestures that, in our corpus, consist of iconic gestures and abstract deictic gestures [17]. Pragmatic gestures were coded but were not considered in the present study (9% of all gestures in both conditions), as they do not provide information on the content itself but they mark pragmatic aspects of the speech act [18]. Representational gestures were labeled as iconic (a gesture reflecting the property of the referent, e.g., a speaker traced a line with a hand palm down to represent a path) or abstract deictic (a gesture referring to abstract location, e.g., a speaker pointed with finger to the right side of the table to refer to the location of the prince) [17]. Only gestures with co-occurring intelligible speech were taken into account for analysis. In other words, gestures that were produced during disfluencies were not analyzed due to the possibility of disfluencies affecting the way speakers gestured [19].

3.5.3. Eye gaze as ostensive signal

Speakers’ eye gaze was coded for “bouts”. We considered a bout as an eye-gaze that was directed to the gesture regardless of its duration [5]. Eye gaze bouts were divided into two categories: bouts to the iconic gesture and bouts to the pointing gesture. For example, a speaker uttered “a path has to be straight” and accompanied *to be straight* with an iconic gesture by tracing a straight line on the table with both hands and directed her eye gaze to the gesture as she was tracing the line. This gesture was assigned one eye gaze bout to the iconic gesture. Instances where, due to the hindered visibility of the eye gaze direction (shadow, eye glasses), it was not clear whether or not eye gaze bouts referred to the gesture, they were not coded as a bout.

3.5.4. Ostensive speech

Ostensive speech utterances (e.g. speaker uttered "like this", "this figure") that accompanied gestures were annotated. Within this annotation it was distinguished whether ostensive speech referred to a complementary or a redundant gesture. For example, if the speaker said “the stair can’t be positioned like this“ and while uttering *like this* performed an iconic gesture that represents the stairs’ peak touching the ground, ostensive speech was annotated as complementary, due to the fact that the position of how the stair is located was present in the gesture, but not in the speech. However, if the speaker said “this knight” and simultaneously performed an iconic gesture representing the knight, it was annotated as redundant, because *the knight* was present in the speech and in the iconic gesture. Furthermore, ostensive speech was annotated whether it referred to the iconic or pointing gesture.

3.5.5. Combination of eye gaze and ostensive speech

The last type of ostensive signal assessed in our study was the combination of eye gaze and ostensive speech. Namely, if the same gesture received an eye gaze bout and was accompanied by ostensive speech it was coded as a combination of eye gaze and ostensive speech. Importantly, we considered using both ostensive signals – eye gaze and ostensive speech – in combination as a separate type of ostensive signal. In other words, three types of the ostensive signals – eye gaze only, ostensive speech only, and combination of eye gaze and ostensive speech were counted separately.

3.6. Reliability

All data were transcribed and coded by the first author of the present study who has a near-native proficiency in Italian language (certified CEFR-C2 level). Randomly selected 20% of the data (3 speakers for each condition) were coded by a second coder. Third author of the present study - a native speaker of Italian language, coded gesture segmentation in strokes and their classification. The agreement between coders on the gesture segmentation was 88%. The strength of agreement between coders on gesture classification was good as indicated by Cohen’s Kappa = .76. Ostensive signals (eye gaze bouts to the gesture and ostensive speech utterances) were coded by another native speaker of Italian language, naïve to the hypothesis of the study. The agreement on eye gaze bouts was almost perfect as indicated by Cohen’s Kappa = .96, and there was a total agreement on ostensive speech utterances (Cohen’s Kappa=1).

4. Results

4.1. Analysis

All analyses, except number of words, were performed on arcsine transformed proportions. Based on previous literature, analyses of number of words, rates of iconic gestures (per 100 words) and proportion of total ostensive signals used were planned comparisons with a prediction that proportions in the child condition will be higher than in adult condition (1-sided) based previous literature and findings of [3]. In regard to the particular type of the ostensive signals no predictions were made.

4.2. Speech and gesture

Speakers produced comparable amount of words (F(1,15)=2.104, ns, one-sided) in child (M=130.44, SE=8.34) and adult (M=118.56, SE=10.31) conditions during descriptions of the rules of the game.

Speakers did use slightly more iconic gestures in the child condition (M=23.10, SE=1.2) in comparison to the adult condition (M=21.15, SE=1.45). However, this difference did not reach significance (F(1,15)=2.532, ns, but approaching *p*=.07, one-sided, η2=.144). Speakers used pointing gestures with comparable frequency (F(1,15)=.093, ns, two-sided) in child (M=2.89, SE=0.67) and adult conditions (M=2.62, SE=0.74).

Due to the fact that pointing gestures were used scarcely and 4 out of 16 participants did not produce any pointing gestures at all, they were excluded from the further quantitative analyses. In further analyses only iconic gestures were considered.

4.3. Overall use of ostensive signals with iconic gestures

The amount of iconic gestures attended to by means of eye gaze and/or ostensive speech during description differed according to the age of the addressee. The prediction was that speakers would use more ostensive signals with children than adults. Indeed, when talking to a child, speakers, on average, highlighted 33% (M=0.33, SE=0.04) of their iconic gestures but 25% (M=0.25, SE=0.03) when talking to an adult. A simple contrast revealed this difference to be significant (F(1,15)=4.268, p=.03, one-sided, η2=.222).

4.4. Eye gaze

Eye gaze was the preferred ostensive signal in both conditions. However, speakers used significantly more eye gaze (F(1,15)=6.766, p=.02, two-sided, η2=.311) with the adults (M=0.89, SE=0.03) than with the children (M=0.77, SE=0.04).

4.5. Ostensive speech

Complementary and redundant ostensive speech utterances were collapsed into a single variable due to the scarce use (3 out of 16 participants) of complementary ostensive speech.

When describing the rules of the game, speakers used ostensive speech to highlight their gestures more with children (M=0.15, SE=0.01) than with adults (M=0.10, SE=0.03), but the analysis showed that this difference was not significant (F(1,15)=1.871, ns, two-sided).

4.6. Combination of eye gaze and ostensive speech

The strategy to use a combination of the eye gaze and ostensive speech to the gesture was almost exclusively used with the children (M=0.08, SE=0.02) rather than with adults (M=0.02, SE=0.01). The difference was statistically significant (F(1,15)= 6,019, p=.03, two-sided, η2=.286).

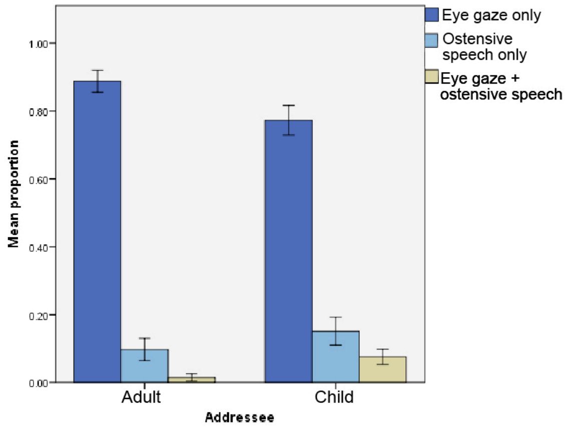


Figure 1: Mean proportion of ostensive signal types to iconic gesture. Error bars display +/- 1SE of the Mean

5. Discussion and conclusions

There has been no quantitative research investigating the signals that speakers use to attend to their gestures. Therefore, the finding that speakers do highlight approximately 33% of their gestures with children and 25% with adults by means of eye gaze, ostensive speech or combination of both during a demonstration task merits attention in its own right. We also confirmed that speakers use more ostensive signals to the iconic gesture with children than adults. The main findings are discussed below.

5.1. Words and gesture

In the present study, speakers produced a comparable amount of words and iconic gestures in both conditions while in [3] the difference for both reached significance. The absence of the difference in our study might be rooted in the similar state of the addressees' knowledge of the task in both conditions. Research has shown that common ground between the speaker and the addressee plays an important role in the way speakers encode their message, also in regard to iconic gesture use [20]. [3] had three different conditions, which compared three imaginary addressees: an *adult - expert*, an *adult - novice*, and a *child - novice*. A difference between the *expert* and *child* was observed, resulting in speakers producing more words and more iconic gestures with children. However, the amount of words and gestures produced with the *novice* did not reveal significant difference in comparison to the child condition. In our study, all addressees, adults and children, did not know the rules of the game. It is therefore possible that the difference was not observed in our study due to the fact that both conditions employed novices as in [3].

5.2. Ostensive signals to gestures

It is plausible to suggest, after examining the present data, that some iconic gestures are used more intentionally than others and therefore not all iconic gestures are intended as equally communicative. Speakers decide which of the iconic gestures are more important to accentuate and they alter the level of the gesture's communicativeness. Just like speech, gesture is at the disposal of the speaker to convey the information. Furthermore, if a particular gesture, according to the speaker's judgment, is able to provide information more efficiently than speech, then attention of the addressee is brought to it by means of ostensive signaling. This assumption is in accordance with [18] who considers gesture as an „equal partner" of speech in utterance formation, where the speakers are free to construct their utterance by means of both modalities and give preference to one or the other according to the context.

Speakers attended to approximately one-third of iconic gestures produced with the children, while with the adults one-fourth of the total iconic gestures was accentuated with an ostensive signal. Our results show that the average rate of iconic gesture use did not differ significantly across conditions (not expected); nevertheless, speakers highlighted more iconic gestures with children than adults. They attended more to gestures with children and an explanation of this may be that, by doing so, they prompted the young addressees to ground the concepts expressed in speech with their referents in gesture and as a consequence provide more diversified input to ease comprehension. Our results cannot say anything about beneficial effects on the addressees. It is nevertheless possible to conclude that speakers, when referring to the children, were significantly more active in providing signals to bring their attention to the iconic gestures they produced.

5.3. Eye gaze

Although total use of ostensive signals was more frequent in the child condition, the analysis showed that within the three different types investigated in this study, speakers used eye gaze to highlight the relevant gesture significantly more with adults. Here, perhaps, one needs to think not about why speakers used more eye gaze with adults, but rather why speakers used less eye gaze to iconic gestures with children compared to adults. A possible answer to this is an increased necessity to control the child's attention. [11] shows that mothers, when demonstrating objects, gaze significantly more and longer at their children than when performing a demonstration task with the adults. This was explained by the need to monitor the children's attention and maintain interest in the activity (demonstrating an object) they were performing.

This study shows that, in general, adults do feel the need to control the attention of the older children regardless of speakers' previous experience with them. It is common sense to assume that it is more difficult to maintain the attention of a child than that of an adult, who is cognitively more disposed to concentrating on the task. It is important to note that even though speakers' eye gaze to the addressee was not coded, throughout the data, when speakers did not look at their gestures, their eye gaze was mainly directed to the addressee.

It is possible to argue that when speakers used eye gaze to look at their gestures, this may have served as a self-assuring or cognitive strategy for the speaker herself rather than as an attention-directing strategy for the addressee. Namely, the speaker might have gazed at her gesture to make sure for herself that she is representing the concept in a precise way. However, since eye gaze to gesture differed between adults and children, this corroborates the communicative function hypothesis. Otherwise, speakers would be as likely to look at their own gestures in both conditions.

5.4. Combination of the ostensive signals: eye gaze and ostensive speech

It is quite plausible to assume that it would require extra effort to provide a clear message to a child in comparison with an adult. The significant trend in the use of combinations of ostensive signals to highlight the gesture mainly with children seems to signal that, indeed, this is the case. In fact, it seems that the most efficient strategy for adults to highlight attention to gesture is not eye gaze or ostensive speech alone but to use eye gaze and ostensive speech together.

Our results support the claim of [3] that speakers might use iconic gesture as a communicative strategy with children to transfer the message more efficiently. How this is achieved, however, seems to differ when the address is imagined versus when the addressee is present. In the present study, speakers highlight iconic gestures that are relevant to the message, namely the gestures that might lead to better understanding of the message. Results of this study demonstrate that adults are aware of the possible benefits of the gesture in comprehension of the message. They are also aware that children might require more guidance to locate this information. Therefore, they use more signals overall, and they combine these signals to render their referent more salient to children.

To summarize, during a demonstration task speakers tended to use more eye gaze to accentuate the relevant gestures with adults compared to children, but a combination of eye gaze and ostensive speech was clearly a strategy designed for children.

6. Conclusions

The scope of this study was to investigate whether the speakers use verbal and visual deixis to make the information expressed in their iconic gestures more salient during a demonstration task. Moreover, the study was aimed at exploring the strategies of ostensive signal use implemented by the speakers when referring to different addressees, namely another adult or a school-aged child. The results are in line with the main hypothesis of the study and support the assumption that speakers use ostensive signals, namely ostensive speech, eye gaze and combination of both to augment the informative relevance of the iconic gesture during demonstration task and they do it differently according to the age of the addressee.

The fact that speakers highlighted their gestures shows that gesture can function as a main constituent of the message (rather than as co-speech), at least in the context of a demonstration task. We found that speakers use more ostensive signals to their gesture with children compared to adults. Eye gaze was the preferred type of ostensive signal in both conditions, followed by ostensive speech. Combination of eye gaze and ostensive speech was almost exclusively used with children. It is plausible that this multiple articulators strategy was used to ensure that the child attends to the gesture.

Furthermore, it was found that eye gaze as an ostensive signal to the gesture was used more with adults than with children. This finding shows the importance of maintaining the attention of the addressee during a face-to-face interaction. Considering that this is more difficult to achieve with children, speakers, when communicating with them, chose to mark their gestures by means of eye gaze and ostensive speech together. On the other hand, bringing attention to the relevant piece of information with adults is more easily achieved. Therefore, to direct the attention of the adults to the gesture, speakers used eye gaze alone.

Further research is needed to be able to extend our knowledge about the use of ostensive signals as markers of the communicative relevance of gestures in other communicative tasks such as narratives, conversation etc. Also, further research is needed to investigate the response of addressees of different ages to these signals and whether the strategy implemented by the speakers is actually efficient.

7. Acknowledgements

The data used in the present study constitutes a part of database that belongs and was collected by Dr. Emanuela Campisi for a Marie-Currie Actions (FP7-PEOPLE-2011-IEF) funded project „Child Directed Demonstrations Across Cultures". The present study was a part of the master's thesis of the first author for MA programme in Linguistics at Radboud University, Nijmegen.

8. References

[1] Levinson, S. C. and Holler, J., "The origin of human multi-modal communication", Philosophical Transactions of the Royal Society of London B: Biological Sciences, 369 (1651), 2014.

[2] Streeck, J., "Gesture as communication I: Its coordination with gaze and speech", Communications Monographs 60 (4): 275-299, 1993.

[3] Campisi, E. and Özyürek, A., "Iconicity as a communicative strategy: Recipient design in multimodal demonstrations for adults and children" Journal of Pragmatics 47(1): 14-27, 2013.

[4] Schmidt, C. L., "Scrutinizing reference: How gesture and speech are coordinated in mother-child interaction" Journal of Child Language 23(2): 279-305, 1996.

[5] Wilson, D. and Sperber. D., "Linguistic form and relevance", Lingua, 90(1): 1-25, 1993.

[6] Wilson, D. and Sperber, D., Relevance Theory. In L. Horn & G. Ward (eds.), The Handbook of Pragmatics, 607-632, Oxford, Blackwell, 2004.

[7] Gullberg, M. and Holmqvist. K., "Keeping an eye on gestures: Visual perception of gestures in face-to-face communication", Pragmatics & Cognition, 7(1): 35-63, 1999.

[8] Gullberg, M., "Giving language a hand: gesture as a cue based communicative strategy", Lund Working Papers in Linguistics, 44: 41-60, 2009.

[9] Holler, J. and Stevens, R., "The effect of common ground on how speakers use gesture and speech to represent size information", Journal of Language and Social Psychology, 26(1): 4-27, 2007.

[10] Brand, R. J., Baldwin, D.A. and Ashburn L. A., "Evidence for 'motionese': modifications in mothers' infant-directed action", Developmental Science, 5(1): 72-83, 2002.

[11] Brand, R. J., et al. "Fine-grained analysis of motionese: Eye gaze, object exchanges, and action units in infant-versus adult-directed action", Infancy, 11(2): 203-214, 2007.

[12] Csibra, G., "Recognizing communicative intentions in infancy", Mind & Language, 25(2): 141-168, 2010.

[13] Csibra, G. and Gergely, G., "Social learning and social cognition: The case for pedagogy", Processes of change in brain and cognitive development. Attention and performance XXI 21: 249-274, 2006.

[14] Gullberg, M. and Kita, S., "Attention to speech-accompanying gestures: Eye movements and information uptake", Journal of nonverbal behavior, 33(4): 251-277, 2009.

[15] Beattie, G. and Aboudan R., "Gestures, pauses and speech: An experimental investigation of the effects of changing social context on their precise temporal relationships", Semiotica, 99(3-40): 239-272, 1994.

[16] Holler, J. and Wilkin, K., "An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses", Journal of Pragmatics, 43(14): 3522-3536, 2011.

[17] McNeill, D., Hand and mind: What gestures reveal about thought, University of Chicago Press, 1992.

[18] Kendon, A. Gesture: Visible action as utterance, Cambridge University Press, 2004.

[19] Seyfeddinipur, M. and Kita, S., "Gesture as an indicator of early error detection in self-monitoring of speech", ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech, 2001.

[20] Holler, J. and Wilkin, K., "Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task", Language and cognitive processes, 24(2): 267-289, 2009.

[21] Sloetjes, H., & Wittenburg, P. (2008, May). Annotation by Category: ELAN and ISO DCR. In *LREC*.

Does gesture add to the comprehensibility of people with aphasia?

Karin van Nispen [1], *Kazuki Sekine*[2], *Miranda Rose* [3]

[1] Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, The Netherlands
[2] Departments of Psychology, University of Warwick, Coventry, the United Kingdom
[3]School of Allied Health, La Trobe University, Melbourne, Australia
k.vannispen@uvt.nl, kazuki@tkc.att.ne.jp, m.rose@latrobe.edu.au

Abstract

Gesture can convey information co-occurring with and in the absence of speech. As such, it seems a useful strategy for people with aphasia (PWA) to compensate for their impaired speech. To find out whether gestures used by PWA add to the comprehensibility of their communication we looked at the information conveyed in gesture (similar to speech, additional to speech or essential information that is absent in speech), produced by 34 PWA and 5 non-brain damaged participants (NBDP) during semi-structured conversation. There were no significant differences found between PWA and NBDP, or between aphasia types. The total number of gestures and the use of similar gestures correlates with the information PWA can convey in speech. Essential gestures are used in instances of speech break down. These findings suggest that gestures used by PWA may add to the compressibility of their communication and that some PWA may use gesture compensatorily.

Index Terms: gesture, speech production, aphasia, compensation

1. Introduction

Gesture can convey information additional to speech or even in the absence of speech [1]. As such, gesture seems a useful compensatory tool for situations in which speech is difficult. For instance, in a bar, where the music is very loud, one could make a drinking gesture to ask whether someone wants a drink. Intuitively this seems a logical strategy. However, people do not easily seem to stop verbal communication and switch to another modality to convey their message [2]. Therefore, in daily life it may be more usual to see people try to shout as loud as possible under these circumstances. Such observations have led researchers to believe that people do not use gesture compensatorily [2] and that the comprehensibility of gesture may be a useful side effect, but not an intended function. For many people with aphasia (PWA), it is no longer possible to convey information in speech. If non-brain damaged people (NBDP) do not use gesture in a compensatory manner does this mean that PWA will not do this either? The present study sets out to find out whether 1) gestures produced by PWA can add to the comprehensibility of their communication and 2) whether, differently from NBDP, PWA use gesture during instances of speech break down.

1.1. Information in speech and gesture

Intentionally or not, gesture can convey information, useful for an interlocutor [3]. This information can be supplementary in that a gestures sometimes conveys information that is not expressed in speech, such as in example 1. Consider a child who tells his mother that he came straight home and accompanies this message with a gesture in which the arms are swinging as if running. The gesture here provides the mother with information additional to the information in speech, namely the manner of the child's return. In some cases, speech may be incomplete and gesture may provide essential information for understanding a message. For instance, as in example 2, when a patient says to the doctor that "he has pain here", while pointing to his leg. Here the gesture is essential for the doctor to understand where the pain is situated.

Example	1	2
"speech"	"I came straight home"	"I have pain here"
gesture	*Arms swing as if running*	*Point to leg*

1.2. Do NBDP compensate using gesture?

Observations such as described above, in which gesture conveys information in addition to that contained in speech, have been used to support the claim that gesture has a communicative function [4].

This communicative function hypothesis is much debated as gesture may also serve other functions, such as aiding cognition [5] or facilitating speech production [6-8]. Following the latter two hypotheses the comprehensibility of gesture may be a useful side effect for an interlocutor, but It may not be its main function. This facilitation hypothesis is supported by evidence showing that people do not deploy gesture in cases of speech difficulties. Gullberg and colleagues showed that gesture production usually stops when speech stops and if gestures are produced during speech break down, these are more often pragmatic (commenting on the fact that there is a speech break down) than representational (depicting the information missing in speech) [2, 9]. They also showed that although representational gestures convey information for an interlocutor, mostly these gestures convey information which is similar to the information conveyed in speech. Therefore representational gestures complement speech but do not replace it [2]. Furthermore, Mayberry and colleagues showed that the production of gesture stops with the production of speech during dysfluent speech in children who stutter [10].

These findings are in line with models that assume that the production of speech and gesture are two highly connected processes [7, 11]. Difficulties in one modality (speech) would be reflected in the other (gesture) restricting the compensatory use of gesture.

1.3. Gesture by PWA

If healthy speakers do not usually compensate for speech difficulties using gestures, can we expect PWA to behave differently? Various studies have shown that PWA use gestures [12, 13] and more importantly that these gestures may benefit their communication [14-17]. Substantial individual differences are reported which, according to Sekine and Rose [12], may be explained by two factors that influence whether PWA use gesture 1) the ability to use gesture and 2) the need to use gesture.

The ability to use gestures: Though research has shown that PWA use gestures, huge individual differences have been reported. The ability to access and select semantic knowledge seems to be an important predictor of PWA's ability to use gesture [14, 17-19]. These findings support the notion that gesture and speech are related processes, but only partly [8]. PWA with difficulties in verbal expression resulting from a semantic impairment, are likely to have difficulties in the production of gesture. PWA with difficulties in verbal expression not resulting from a semantic impairment, for example, a phonological access impairment on the other hand may be able to use gestures still.

The need to use gestures: The studies [2, 9, 10] discussed above claiming that people do not use gesture compensatorily may be explained by the notion that there is no real need for them to put information in gesture. Particularly for second language learners and individuals who stutter the primary goal may be to succeed in putting information in speech despite the struggle to do so. PWA on the other hand, are more often aware of the fact that they will not be able to convey information in speech and instead try other means of communication. Still, there may be differences among different types of aphasia in their need to use gesture [12, 13]. For example, for people with mild or anomic aphasia there may be a need for gesture in cases of word retrieval. A gesture can help replace the missing word ("I would like coffee and ……." + *gesture using an imaginary spoon to scoop something*, in order to communicate the word 'sugar'). For PWA with very limited speech production abilities, the need for gesture may be even larger. Although it might be difficult to convey a full message in gesture, providing some aspect of the message in gesture might increase the likelihood of successful communication of someone otherwise unable to communicate ("…….." + *drinking gesture*, in order to communicate a request for something to drink).

Importantly, gesture may also be comprehensible in cases of unintentional use. Gesture naturally co-occurs with the production of speech and may often convey information [20]. In cases where speech is planned but not produced, gesture might still be produced. This is illustrated by a case-study by Van Nispen and colleagues [21] where an individual with Wernicke's aphasia produced incomprehensible speech, but fairly normal co-speech gestures. Although the individual probably did not intentionally plan to produce the gestures, these gestures still greatly improved his message comprehensibility.

Finally, we wish to point out that the use of gesture may also depend on a third factor; the type of information needed to be conveyed. Gesture seems most useful to convey information regarding actions, movements or shapes [22], but may be more limited for other categories of referents. For instance, one may use gesture to communicate about hobbies (reading a book, cycling, watching television), but it may be more difficult to use gesture to explain your political viewpoint.

1.4. Present study

If healthy speakers do not compensate for speech difficulties using gestures, can we expect PWA to do this? Sekine and colleagues [12, 13] have revealed the type and frequency of gesture used by PWA. Its communicative value remains understudied. Therefore, the present study looks into the communicative value of gestures used by PWA and aims to determine whether these add to the information conveyed in speech. Furthermore, we will look at whether compensatory gestures are used during instances of speech break down.

For this study we examined the gestures used by 34 PWA and 5 NBDP previously analyzed in two earlier studies by Sekine and colleagues [12, 13]. We compared the information conveyed in gesture to the information conveyed in speech by using a coding scheme developed by Colletta and colleagues [23, 24]. The present paper presents preliminary results of this study.

2. Method

2.1. Participants

This study uses data from an online database; AphasiaBank [25], also analyzed in two studies by Sekine and colleagues [12, 13]. The present paper reports on 34 PWA (19 male, age 34-73) and 5 NBDP (1 male, age 36-84). For a detailed description of inclusion and exclusion criteria see [12, 13].

For PWA we examined two variables, both based on the Western Aphasia Battery, WAB [26]:

1) Aphasia type: Broca (n=6), Wernicke (n=8), Anomic (n=8), Transmotor (n=4) and Conduction (n=8);
2) The ability to convey information in speech, based on WAB Spontaneous speech information content.

2.2. Design

Participants were videotaped during a semi-structured interview. An experimenter asked four questions about the participants' recovery and an important event in their lives following a strict protocol (see www.aphasiabank.com):

1) How do you think your speech is these days?
2) Do you remember when you had your stroke?
3) Tell me about your recovery. What kinds of things have you done to try to get better since your stroke?
4) Thinking back, can you tell me a story about something important that happened to you in your life?

Questions for NBDP were comparable. Here the interviewer asked the participant to tell her about an illness or medical condition that they had and whether they had experience with people with language difficulties:

1) Could you tell me what you remember about any illness or injury you've had?
2) Tell me about your recovery from that illness (or injury). What kinds of things did you do to get better?
3) Have you had any experience with people
who have a difficult time communicating? Please tell me what the problems were and what you did about it.
4) Thinking back, can you tell me a story about something important that happened to you in your life?

Table 1. Categories for communicative value of gesture related to speech and their definitions with examples.

Category	Gesture label	Definition	Example	
		The information in gesture………….(fill in definitions given below) the information in speech	**"speech"**	***gesture***
similar	i Reinforce	is identical to	"me"	*point to self*
	ii Integrate	adds precision to	"drinking"	*pretend to drink*
additional	iii Supplement	adds new information (not essential for understanding the message)	"cake"	*draw round shape*
essential	iv Complement	brings a necessary complement to the incomplete	"I have pain here"	*point to leg*
	v Contradict	contradicts	"Five"	*show four fingers*
	vi Substitute	replaces (missing)	"….."	*thumbs up gesture*
			"slowly"	*move hands upwards*

2.3. Coding

All gestures used by participants were coded for their communicative value. For this we recoded the data from the previous studies by Sekine and colleagues [12, 13], who determined what type of gestures people used. For the present study, we added a second label to every gesture determining its communicative value. For this aim we used a coding scheme developed by Colletta and colleagues [24] which determines the relation of a gesture to the corresponding speech (see Table 1 for short definitions of the labels used). All coding was performed using the software ELAN [27].

2.4. Analyses

For the analyses we collapsed the six gesture labels into three categories; similar, additional, or essential. Similar is defined as information in gesture is similar to that in speech, 2) Additional was categorized if gestures add additional information to information in speech and 3) Essential refers to gestures that are essential for understanding a message (see Table 1). Essential gestures do not necessarily occur in the absence of speech (the gesture: *a hand moving upwards* in combination with the speech "slowly" is essential for understanding the message; there is improvement).

In the analyses we looked at the total number of gestures used, the number of times people used a certain category and the proportion of each category considering the total number of gestures used.

In a quantitative analysis, using ANOVA, we first examined the potential differences in the total use of gestures and gesture categories of communicative value (number and proportions) between PWA and NBDP, and within PWA for aphasia type using Bonferoni's post hoc analysis. Second, we performed correlational analyses for information in speech (WAB spontaneous speech score) and the total number of gestures used and the different gesture categories (number and proportion). Finally, in a qualitative analysis we looked at whether essential gestures occurred during instances of speech break down.

3. Results

3.1. Quantitative analyses

No significant differences were found for the use of similar, additional or essential gestures between NBDP and PWA (see Figure 1), power ranges from .05 to .38 for the dependent variables. Within the group of PWA there were no significant differences for Aphasia Type (see Figure 2), power varies from .11 to .24 for the dependent variables. Information in speech correlated with the total number of gestures r=.36, p=.04 and the number of similar gestures r=.32, p=.06 (trend).

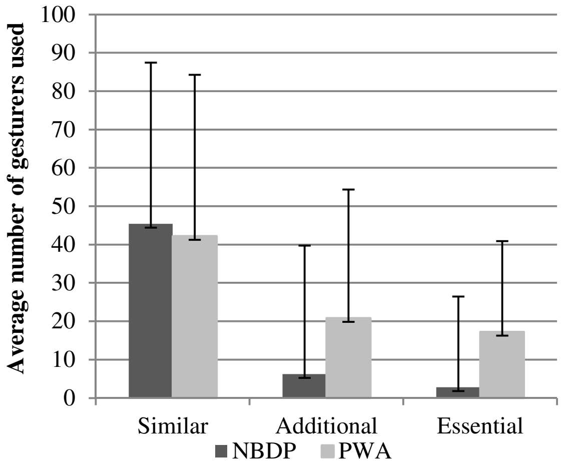


Figure 1. Average numer of gestures used per category; similar, additional and essential for NBDP and PWA (error bars show *SD*).

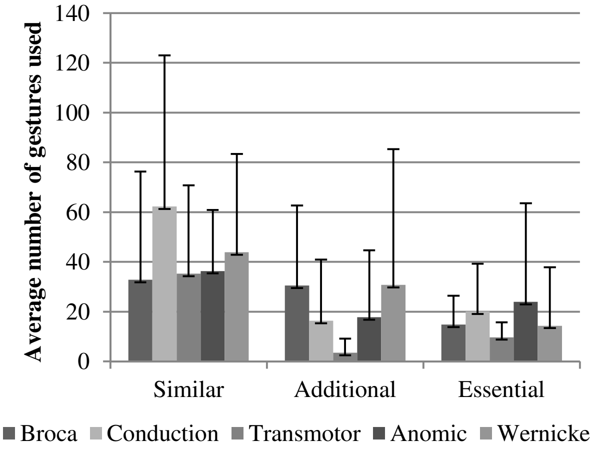


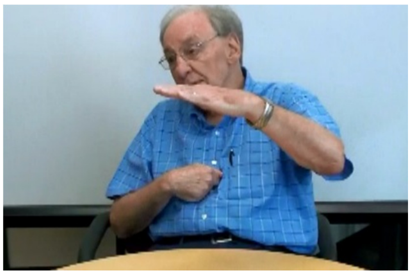
Figure 2. Average numer of gestures used per category; similar, additional and essential by Aphasia type (error bars show *SD*)

3.2. Qualitative analyses

We did not find a difference between NBDP and PWA, or between PWA with different types of aphasia in their use of essential gestures. This does not mean that PWA do not compensate for their speech difficulties in gesture. In this qualitative analysis for two individuals (case ID; Scale 01 and

Kansas 12), we discuss how different essential gestures, occurring with or without speech, compensated for cases of speech break down.

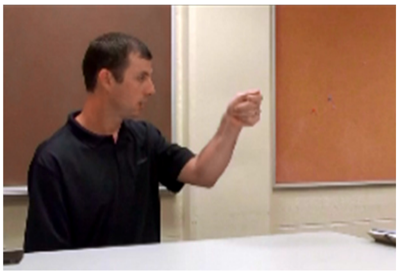
Essential gestures with speech: Both Scale 01 and Kansas 12 use a number of gestures, which are most often produced co-occuring with speech. Their essential gestures also often co-occur with speech (see Figure 4 for an example). The fact that the gesture in this example is used during (semi) fluent speech does not mean that there is no speech break down. The repetition of words ("slowly, slowly"), low speech rate and a short interruption ("uhm") indicate that Scale 01 struggles to find the word ("improving"). The gesture he uses here ensures that this speech break down does not interrupt communication greatly. There are two possible interpretation for the origin of the gesture in this case. Firstly, speech and gesture may have been planned correctly, but the speech was not produced because of difficulties retrieving the correct verb. The gesture may not have been intentionally created for compensation, but is essential under these circumstances nevertheless. A second option is that Scale 01 was aware of the fact that he could not produce the verb "improve" and made this gesture to convey the information instead.



"speech" "Slowly, slowy,uhm,just a tiny bit"
gesture *hand gradually moving upwards*

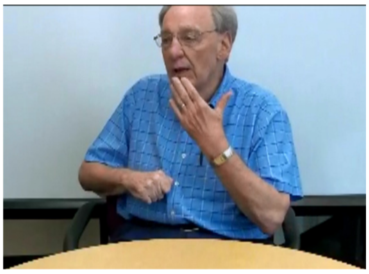
Figure 4. Example of essential gesture co-occuring with speech. The hands illustrate the concept of 'improvement', not conveyed in speech.

Essential gestures occurring without speech: There are some instances of speech break down where information is conveyed in gesture only. In these situations both individuals tried to compensate using gesture. Kansas 12 experienced a speech break down, thought for a moment ("uhm") and switched to using gesture to convey his message (Figure 5). Scale 01 did something similar (Figure 6). Interestingly, after he performed the gesture, he also conveyed the same information in speech "I can't talk". Maybe the extra time given by performing the gesture helped him in retrieving the information needed to give a verbal response. It may also be that his gesture directly facilitated speech production. Considering this context, Scale's gesture is no longer essential. The intention to make this gesture though seems compensatory.



"speech" "hunting and uh ……………."
gesture *swinging the hand as if casting a fishing rod*

Figure 5. Example of essential gesture in absence of speech.



"speech" "Nothing…………………………..I can't talk"
gesture *move lips without sound coming out + moving hand back and forth in front of mouth*

Figure 6. Example of essential gesture in absence of speech.

4. Discussion

4.1. Results

Though PWA seem to use more additional and essential gestures than NBDP, this difference did not reach significance. Neither did we find any significant differences for Aphasia type. Considering the small sample sizes, this might be explained by the low statistical power of our study. We did find correlations between information in speech and the use of similar gestures and the total number of gestures used. Finally, a qualitative analysis showed that PWA use essential gestures during instances of speech break down, which occur both with and without speech.

4.2. Do gestures by PWA add to their communication?

The correlations found between the total number of gestures used and the number of similar gestures used by PWA is in line with the idea that gesture naturally co-occur with the production of speech [20]. Though PWA do not differ from NBDP in their use of additional and essential gestures, these gestures may contribute to the comprehensibility of their communication.

It remains difficult to determine whether gestures are intended compensatorily, or that they are a natural result of planned communication. The observation that essential gestures are used during instances of speech break down suggests that PWA use gesture compensatorily. In this aspect, PWA seem to differ from what NBDP would usually do [2]. Importantly, speech break down often does not result in moments of silence. PWA use various communicative strategies, e.g. speech and gesture, to prevent communication breakdown. These findings support the hypothesis that gestures have a communicative function [4] and can be used compensatorily for information missing in speech [8]

4.3. Future directions

This paper reports on a preliminary results that may contribute to find out whether gestures used by PWA add to the information conveyed in speech and whether gestures are used during instances of speech break down. Our preliminary findings give rise to ideas for future directions.

Firstly, our analyses did not show differences in the use of additional or essential gestures between PWA and NBDP or between different types of aphasia. This suggests a need for both better powered studies and a more detailed analyses in order to determine more precise patient profiles of PWA that do or do not use gestures compensatorily.

Secondly, more analyses are needed to establish whether the coding scheme used is a reliable tool for the analysis of the communicative value of gestures used by PWA. To this aim we will perform inter- and intra-coder reliability testing.

5. Conclusions

PWA use gestures with and without speech, and these gestures can add to the comprehensibility of their communication. During instances of speech break down, PWA seem to make explicit attempts to convey information, which is missing in speech, by gesture. In this aspect they seem to differ from NBDP. More detailed analyses are needed to determine more precise patient profiles of PWA that do or do not use gestures compensatorily.

6. Acknowledgements

We are very grateful to all researchers and participants who contributed to AphasiaBank and who made this research (and many other studies) possible. Furthermore, we like to acknowledge the 'Jo Kolk studiefonds' for providing the first author with travel funds hereby enabling the cooperation between researchers from Australia and The Netherlands.

7. References

[1] Beattie, G. and Shovelton, H., *A critical appraisal of the relationship between speech and gesture and its implications for the treatment of aphasia.* Advances in Speech-Language Pathology, 2006. **8**(2): p. 134-139.

[2] Gullberg, M., *Gesture as a communication strategy in second language discourse: A study of learners of French and Swedish*. Vol. 35. 1998: Lund University.

[3] Beattie, G. and Shovelton, H., *Mapping the Range of Information Contained in the Iconic Hand Gestures that Accompany Spontaneous Speech.* Journal of Language and Social Psychology, 1999. **18**(4): p. 438-462.

[4] Kendon, A., *Gesture: Visible action as utterance*. 2004: Cambridge University Press.

[5] Goldin-Meadow, S. and Beilock, S.L., *Action's Influence on Thought: The Case of Gesture.* Perspectives on Psychological Science, 2010. **5**(6): p. 664-674.

[6] Krauss, R.M., Chen, Y., and Gottesman, R.F., *Lexical gestures and lexical access: A process model.*, in *Language & Gesture*, D. McNeill, Editor. 2000, Cambridge University Press: Cambridge.

[7] Kita, S. and Özyürek, A., *What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking.* Journal of memory and language, 2003. **48**(1): p. 16-32.

[8] De Ruiter, J.P., *The production of gesture and speech*, in *Language & Gesture*, D. McNeill, Editor. 2000, Cambridge University Press: Cambridge. p. 284-311.

[9] Graziano, M. and Gullberg, M. *Gesture production and speech fluency in competent speakers and language learners.* in *Tilburg Gesture Research Meeting (TiGeR) 2013*. 2013. Tilburg University.

[10] Mayberry, R.I., Jaques, J., and DeDe, G., *What stuttering reveals about the development of the gesture-speech relationship.* New Directions for Child and Adolescent Development, 1998. **1998**(79): p. 77-88.

[11] McNeill, D., *Gesture and Thought*. 2005, Chicago and London: University of Chicago Press.

[12] Sekine, K. and Rose, M., *The Relationship of Aphasia Type and Gesture Production in People With Aphasia.* American Journal of Speech-Language Pathology, 2013. **22**(4): p. 662-672.

[13] Sekine, K., Rose, M., Foster, A.M., Attard, M.C., and Lanyon, L.E., *Gesture production patterns in aphasic discourse: In-depth description and preliminary predictions.* Aphasiology, 2013: p. 1-19.

[14] Hogrefe, K., Ziegler, W., Wiesmayer, S., Weidinger, N., and Goldenberg, G., *The actual and potential use of gestures for communication in aphasia.* Aphasiology, 2013. **27**(9): p. 1070-1089.

[15] Mol, L., Krahmer, E., and van de Sandt-Koenderman, M., *Gesturing by Speakers With Aphasia: How Does It Compare?* Journal of Speech Language and Hearing Research, 2013. **56**(4): p. 1224-1236.

[16] Kemmerer, D., Chandrasekaran, B., and Tranel, D., *A case of impaired verbalization but preserved gesticulation of motion events.* Cognitive Neuropsychology, 2007. **24**(1): p. 70-114.

[17] de Beer, C., Carragher, M., van Nispen, K., de Ruiter, J., Hogrefe, K., and Rose, M., *How much information do people with aphasia convey via gesture?* American Journal of Speech-Language Pathology, submitted.

[18] Cocks, N., Dipper, L., Pritchard, M., and Morgan, G., *The impact of impaired semantic knowledge on spontaneous iconic gesture production.* Aphasiology, 2013. **27**(9): p. 1050-1069.

[19] van Nispen, K., van de Sandt-Koenderman, W.M.E., Mol, L., and Krahmer, E., *Pantomiming what you cannot say A study on the influence of a semantic disorder on the ability to compensate for speech loss with the use of pantomimes.*, in *Proceedings of the NeuroPsychoLinguistic Perspectives on Aphasia.* 2012: Toulouse. p. 55.

[20] McNeill, D., *Language and Gesture*. 2000, Cambridge: Cambridge University Press.

[21] van Nispen, K., van de Sandt-Koenderman, M., Mol, L., and Krahmer, E., *Should pantomime and gesticulation be assessed separately for their comprehensibility in aphasia? A case study.* International Journal of Language and Communication Disorders, 2014. **49**(2): p. 265-271.

[22] van Nispen, K., van de Sandt-Koenderman, M., Mol, L., and Krahmer, E., *Pantomime Strategies: On Regularities in How People Translate Mental Representations into the Gesture Modality*, in *Proceedings of the 36th Cognitive Science meeting*. 2014: Quebec.

[23] Colletta, J.-M., Guidetti, M., Capirci, O., Cristilli, C., Demir, O.E., Kunene-Nicolas, R.N., and Levine, S., *Effects of age and language on co-speech gesture production: an investigation of French, American, and Italian children's narratives.* Journal of child language, 2015. **42**(01): p. 122-145.

[24] Colletta, J.-M., Kunene, R., Venouil, A., Kaufmann, V., and Simon, J.-P., *Multi-track Annotation of Child Language and Gestures*, in *Multimodal Corpora*, M. Kipp, et al., Editors. 2009, Springer Berlin Heidelberg. p. 54-72.

[25] MacWhinney, B., Fromm, D., Forbes, M., and Holland, A., *AphasiaBank: Methods for studying discourse* Aphasiology, 2011. **25**: p. 1286-1307.

[26] Shewan, C.M. and Kertesz, A., *Reliability and validity characteristics of the Western Aphasia Battery (WAB).* Journal of Speech and Hearing Disorders, 1980. **45**(3): p. 308-324.

[27] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H., *ELAN : a professional framework for multimodality research.* 2006.

Impact of iconic gestures on naming of taught words in children with specific language impairment

Susanne Vogt [1,2]

[1] University of Applied Sciences Fresenius, Idstein, Germany
[2] Department of German Linguistics, University of Marburg, Marburg, Germany
vogt@hs-fresenus.de

Abstract

The current study tested the hypothesis that learning new words while simultaneously observing iconic gestures benefits word naming in children with specific language impairment (SLI). Children with SLI as well as typically developing children named words learned with iconic gestures better than words learned with a gesture that merely guides children's attention to the taught words when the respective gestures were provided as naming cues during assessment. Naming cues improved naming performance in the iconic condition. Children with SLI showing low abilities in noun naming in particular appear to benefit from learning words paired with iconic gestures.

Index Terms: word, iconic gesture, attention-directing gesture, semantic representation, language impairment

1. Introduction

A popular assumption poses that iconic gesture aids language acquisition in children. The view that the resemblance between the gesture and the referent might ease language learning, and word learning in particular, is attractive also for clinicians in the field if speech and language therapy. It rests on the idea that iconic gestures are intuitively accessible and less arbitrary than spoken symbols. They capture properties of a referent and thus serve as semantic enrichment cues and aid recall (Capone & McGregor, 2005; Capone Singleton, 2012, Hostetter, 2011). This might ease the complex mapping processes required for word learning. During word learning, the child initially creates a preliminary and incomplete representation of the new form-referent link, this is often referred to as fast mapping. Progressively, through slow mapping the child establishes an advanced representation with data from different sources, gesture among them (Alt & Suddarth, 2012, Rohlfing, 2013). Starting by age three, children are able to understand the meaning of iconic co-speech gestures (Stanfield, Williamson & Özçalişkan, 2014) and might strengthen semantic representations of the referents by observing iconic gesture.

Indeed, research suggests that iconic co-speech gestures enhance word learning when a suitable gesture is paired with a word. This has been shown in young typically developing (TD) children (e.g. Capone & McGregor 2005; Capone Singleton, 2012; McGregor, Rohlfing, Bean & Marschner, 2009), in children learning a second language (e.g. Tellier 2008) and also in clinical groups such as children with Down Syndrome (Foreman & Crews 1998; Launonen, 2003). According to Capone & McGregor (2005) iconic gestures serve to enrich semantic representation of words. They showed that toddlers benefit from iconic shape gestures rather than iconic function gestures for naming taught novel nouns. The two iconic gesture conditions were superior to a condition where no gesture was provided (Capone & McGregor, 2005) and similarly, to a condition where a point gesture was provided, that is a gesture that does not exemplify a semantic feature (Capone Singleton, 2012). Concluding from her findings, Capone Singleton (2012) suggests the use of iconic cues in word learning interventions for children with language impairments.

Children with SLI show deficiencies in language development despite otherwise normal development. Their language skills often resemble those of younger children (Alt & Suddarth, 2012; Bishop, 2006). Among other difficulty, vocabulary acquisition and word learning are major problems. Both fast and slow mapping abilities are concerned: word knowledge emerges only slowly and is fragile relative to peers, including semantic representations (Alt & Plante, 2006; Kan & Windsor, 2010). Retrieving words poses a particular challenge. As failure in word retrieval has been related to weak semantic representation (Capone & McGregor, 2005; Sheng & McGregor, 2010), the use of iconic co-speech gesture may prove as useful semantic enrichment cues potentially supporting word learning in children with SLI.

Pioneering work of Ellis Weismer & Hesketh (1993) has shown that in a fast mapping task children with SLI as well as TD children comprehended novel words better when the words were trained with accompanying iconic gestures, compared to a condition where the novel words were trained without gestures. In the group with SLI, children who had demonstrated comprehension deficits in the language profiles tended to benefit more from iconic gestures. For naming the novel words no difference was found.

Lüke & Ritterfeld (in press) extended this work. To go beyond fast mapping, they introduced novel words as names for cartoon characters. Effects of iconic gestures on learning of these names were compared to a no-gesture condition. For fast mapping, no advantages for the iconic gesture condition were found. However, during extended word learning, children showed a gesture benefit for naming the novel names, but not for comprehension.

The current study drew on these findings. A word learning study was conducted, employing a repeated measures within-subjects design under two learning conditions: the new words were trained with iconic gestures or an attention-directing gesture, respectively, in the form of a uniformly raised forefinger as control condition. Such a gesture guides listeners to attend to parts of the accompanying speech and thus serves a metacognitive purpose. This approach allowed me to compare two conditions where both the spoken word and a gesture have to be mapped onto the referent. I was particularly interested in word naming after a period of slow mapping because it is important to be able to retrieve a word when it is needed in daily life. Word naming is usually assessed in a binary way (named or not named). In this study, I graded the naming task by applying the respective gestures as a cue when the child had failed to name the referent accurately, arguing that this scaffolding might enable children with weaker semantic representations to successful naming (Capone & McGregor, 2005; Capone Singleton, 2012).

The study asked a) whether there is an effect of gesture condition on naming of taught words, b) in case of a naming failure, whether there is an effect of gesture cues on naming performance and c) whether there are relations between performance on the naming task and the language profiles (word comprehension and production in particular) of children with SLI and children matched for age and language, respectively. Based on the literature, I expected that learning in the iconic gesture condition would lead to better naming performance and that iconic gestures (but not the attention-directing gesture) as cues for naming would improve performance in case of previous naming failure. As for potential correlations between children's language profiles and performance on the word naming task, no clear hypothesis could be formulated.

2. Methods

2.1 Participants

Participants in the study were monolingual German speaking children who showed normal general development including nonverbal cognition: 18 children with SLI (mean age 4;6 years), 18 TD children matched for language (LM; 3;3 years) and 15 TD children matched for age and gender (AM; 4;5 years). Children with SLI had formerly been diagnosed by the child's speech and language clinicians. The diagnoses and information regarding language and nonverbal cognitive skills were confirmed by standardized measures administered before training. Nonverbal cognitive ability was within normal range for all children. A range of language skills was assessed using norm-referenced tests (table 1: raw score and percentile group means and SD of all measures and between-group comparisons). In all language measures, children with SLI differed from AM (see table 1: raw and percentile scores) and matched LM (see table 1: raw scores). To be included in the SLI group, children had to perform more than one SD below the mean on at least three of the language subtests administered.

Table1. *Participant information: group means (standard deviation) of cognitive and language measures*

		AM (n = 15)		**SLI** (n = 18)		**LM** (n = 18)
Independent Variable	Scores	M (SD)	*p*	M (SD)	*p*	M (SD)
Age	Months	53 (3,2)	ns	54 (7,9)	***	39 (1,4)
Nonverbal Cognition	Percentile	87 (16,3)	ns	73 (24,8)	ns	60 (22,1)
Grammar comprehension	Percentile	44 (25,9)	**	15 (17,1)		56 (26,1)
	Raw score	6 (2,8)		4 (2,1)	ns	4 (1,8)
Noun comprehension	Percentile	56 (32,8)	**	24 (25,5)		67 (28,3)
	Raw score	17 (1,4)		15 (2,1)	ns	16 (2,1)
Verb comprehension	Percentile	64 (31)	***	27 (27)		78 (21,7)
	Raw score	16 (2,3)		13 (3)	ns	14 (2,6)
Noun naming	Percentile	58 (27,8)	***	17 (17)		76 (23,6)
	Raw score	16 (2,2)		10 (4,2)	ns	12 (2,3)
Verb naming	Percentile	60 (39,2)	***	4 (12,2)	no normative data	
	Raw score	12 (3,3)		6 (3,4)	ns	7 (2,2)
Word definition	Percentile	46 (24,3)	***	17 (15,7)		42 (16,1)
	Raw score	10 (2,3)		6 (3,2)	ns	7 (1,6)
Nonword repetition	Percentile	63 (30,2)	***	15 (20,4)		45 (28,2)
	Raw score	10 (4,4)		4 (3,2)	ns	5 (2,8)

Note: AM = age matched children; SLI = children with SLI; LM = language matched children
ns = non siginificant; ** *p* < .01; *** *p* < .001

2.2 Target words and learning conditions

Target words consisted of 12 low-frequency German words, which are unfamiliar for children of this age, nouns and verbs of equal shares. Nouns represented animal species (e.g. *a rail*); verbs were intransitive and represented movement types (e.g. *to stalk*). All items had been tested in a pre-study in terms of appropriateness in a word learning study. Moreover, pilot work had shown that for TD children of this age learning 12 words attained an optimal performance range. For children with SLI and LM children however, learning 12 words led to mental overload and reduced attention during training and to floor effects in learning assessments. For these children, learning 8 new words turned out appropriate. I therefore decided to train unequal numbers of target words and accordingly express data as percentages.

Children learned the words under two conditions: In the iconic gesture condition spoken words were paired by a gesture that mirrored a striking feature in shape or performance of the respective referent. As control condition, an attention-directing gesture in the form of a raised forefinger was used. Such a gesture does not exemplify a semantic feature of the referent but rather serves a metacognitive purpose by guiding listeners' attention to the new words. Thus, children with SLI and LM learned 2 nouns and 2 verbs under both conditions resulting in 8 words in total, whereas AM children learned 3 nouns and 3 verbs under both conditions resulting in 12 words altogether. The words to be learned in the iconic and the attention-directing gesture condition were counterbalanced across children.

2.3 Procedure

Children were seen individually for six sessions. The first two sessions comprised assessment of nonverbal cognition, language abilities and words to be taught during the training. Subsequently, three training sessions were conducted two to three days apart each. The target words were introduced during the first session and repeated in the remaining two sessions, following a standard protocol such that children heard the words repeatedly during bookreading and play. Spoken target words were paired with either the iconic or the attention-directing gesture. We had created a story containing all target words and illustrated the story in a story book. The story served as basis for the training. Children heard the words 17 to 20 times per session, 57 times in total. Naming of taught words was assessed two to three days after training completion. The child was asked “What is this?” and “What is she doing?”, respectively. In case of failing to accurately name the target word, the child was encouraged to think again of the word label. If the child still failed naming, the respective gesture the target word had been paired with during the training (either iconic or attention-directing) was provided as a cue: “Look”: GESTURE. The gesture cue was determined by the respective learning condition. Correct responses scored one point each. A response was rated as correct if the child produced the target word or a morphological variant within a multiple-word response.

2.4 Data analysis

To determine whether iconic gestures served to improve naming auf taught words across groups, we applied a 2 × 2 repeated measure ANOVA with the factors learning condition (iconic = ICON, attention-directing = ATTENT) and naming cue (without cue, with cue) as within-subjects factors and group as between-subjects factor and post hoc Bonferroni tests. The dependent variable was percentage of accurately named taught words. Bonferroni corrections resolved significant main effects and interactions. To explore an iconic gesture benefit on naming performance the benefit (benefitCOND) was defined as the mean difference of words learned with ICON gestures minus words learned with ATTENT gestures (e.g. benefitCOND = 2 implies the child learned two more words in the ICON than in the ATTENT condition). This value was correlated with children’s’ language profiles assessed before the training.

3. Results

3.1 Accuracy of naming responses

There was a significant main effect of cue, $F(1, 48) = 41{,}59$, $p < .001$, $\eta_p^2 = .46$ with children of all groups performing better when a gesture cue was provided during naming than without a cue. There was also a main effect of learning condition $F(1, 48) = 13{,}05$, $p = .001$, $\eta_p^2 = .21$ with learning words in the ICON condition being superior to learning words in the ATTENT condition. The main effects were modified by a significant cue × condition interaction, $F(1, 48) = 23{,}84$, $p < .001$, $\eta_p^2 = .33$, such that children of all groups named words learned in the ICON condition better than words learned in the ATTENT condition (ICON/ATTENT: SLI: 43,2/26,3, LM: 56,9/34,7, AM: 54,4/31,1) when a naming cue was provided during the naming task (significance levels: SLI $p = .018$, LM $p = .004$, AM $p = .001$) but not when no naming cue was provided (ICON/ATTENT: SLI: 26,3/22,2, LM: 38,8/33,3, AM: 35,5/28,8). Naming cues improved naming performance in the ICON condition, but not in the ATTENT condition, except marginally for children with SLI, $t(17) = 1.84$, $p = .083$, $d = .48$. Main and interaction effects are depicted in figure 1.

3.2 Correlations between language profiles and benefit of gesture condition on naming

A moderate negative correlation was found between benefitCOND and noun naming performance in children with SLI ($r_p = -.553$, $p = .017$), explaining 30% of the variance, indicating that children with SLI and lower abilities in naming of nouns on a standardized measure benefit more from learning words with iconic gestures. Note that the gesture benefit in naming was not correlated with additional standardized measures, therefore noun naming performance was an independent predictor and only in children with SLI.

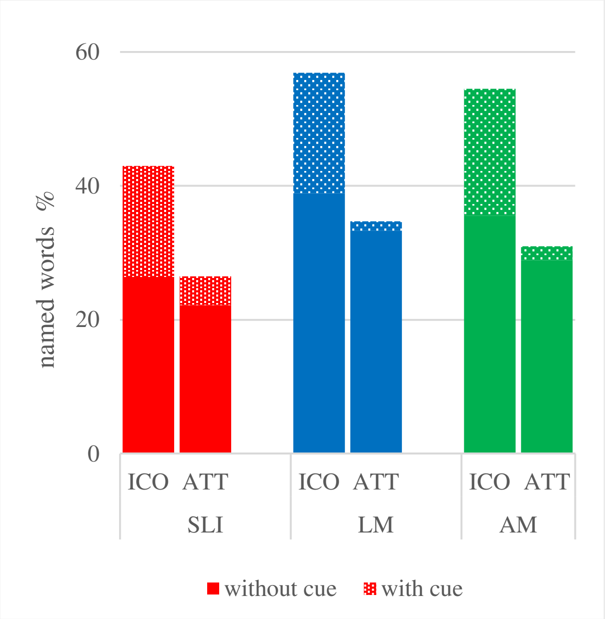


Figure 1: *Percentage of accurately named words in the iconic (ICO) vs. attention-directing (ATT) learning condition across groups (SLI, language matched, age matched children).*
Solid fill colors: naming performance without cue; textured parts: naming performance with gesture cue provided)

4. Discussion

Does learning new words paired with iconic gestures help children name taught words after a period of slow mapping compared to learning words with an attention-directing co-speech gesture that does not exemplify a semantic feature of the referent? To my knowledge, effects of iconic and noniconic (i.e. attention-directing) gestures on naming have not yet been compared in children with SLI and TD children. The hypothesis was that learning words paired with iconic gestures would facilitate naming in children of all groups. Moreover, in case of failure in naming taught words, I expected children to improve naming performance when providing the iconic gesture as a cue during word retrieval.

Despite a descriptive advantage for naming words trained with iconic gestures, the difference to naming words trained with an attention-directing gesture failed to reach significance. A difference between the two gesture conditions only emerged when children were provided the respective gestures (iconic or attention-directing) in case of previous failure in naming the taught word. In that case, children of all groups named words learned with iconic gestures significantly better. Therefore, the prediction that learning new words together with iconic gestures would be superior to learning words with an attention-directing gesture cannot be fully supported. An iconic gesture benefit for naming novel names had been found in the study of Lüke & Ritterfeld (in press) for children with SLI. The authors confirm earlier findings regarding TD children, that iconic gestures enrich semantic representations and thus facilitate naming of words. Why was this not the case for the children in this study? On the one hand, scaffolding effects of iconic gestures on naming taught words in this study might be negligible due to input manipulations differing from those in the abovementioned study, such as word type (novel vs. real words, names vs. nouns and verbs), number of taught words (9 vs. 8) or exposure to target words (46 vs. 57) as well as due to differing participants’ language profiles (e.g. monolingual and bilingual vs. merely monolingual). Alternatively, a gesture that leads children to attend to new words presented may also be supportive. In the study of Lüke & Ritterfeld (in press), the control condition was merely an absence of gesture, whereas here an attention-directing gesture condition was applied allowing to compare word learning under two gesture conditions. Therefore, it appears that children in this study derived some benefit from observing both iconic and attention-directing gestures during word learning. Both types of gesture guide children's attention to the intended target words, thus increasing salience of the words and facilitating learning. The iconic gesture additionally exemplifies a striking feature of the referent; in this way, it aids children to some extent to enrich semantic representation of the referent.

As word learning is an ongoing process, this study assessed children's naming performance not only in terms of named or not named, but instead graded assessment of learning achievement by providing a cue in case of failure in naming. The cues provided were the respective gestures the word had been paired with during the training. Similar to Capone Singleton (2012, 288) cued naming was viewed as providing some scaffolding "to tap word representations that were just on the threshold of activation". Concerning my second hypothesis, as expected, naming performance of all children increased when a gesture cue was provided during the naming task, compared to naming without cue. Providing iconic gesture cues more effectively facilitated word naming than did providing cues to merely direct children's' attention. Here, the characteristic capacity of iconic gestures as to embody semantic features of a referent becomes evident. Observing the respective iconic gesture during word retrieval appears to activate word representations and thus to ease the access to the referent's word form. This enabled children with SLI as well as TD children, who had previously failed to name the taught words, to significantly improve naming. Interestingly, and contrary to my expectation, there was a marginally significant small to medium effect of the attention-directing gesture cue on naming performance in children with SLI, indicating that to some degree these children also took advantage from a gesture that does not exemplify semantic features of the referent. Apparently, the gesture by itself brings implicit word knowledge to the surface, therefore facilitating the access to the word form, even when the gesture does not make any semantic information available.

No clear hypothesis had been formulated regarding the question which children might benefit from learning words paired with iconic gestures for naming. An earlier finding from a fast mapping task (Ellis Weismer & Hesketh, 1993) had pointed towards a possible iconic gesture benefit for comprehension of taught words in children with SLI evidencing low comprehension capacities. In my study, not fast mapping skills, but naming performance after a period of slow mapping was assessed. For TD children, no correlation patterns emerged. However for children with SLI, there was a moderate negative correlation between the iconic gesture benefit and performance in naming nouns in a standardized measure, accounting for 30% of the variance. Although not predicted, this finding does not come as a surprise. It implies that specifically language impaired children with lower word production skills may take advantage of learning words paired with iconic gestures for naming and thus enrich semantic representations. By contrast, children with SLI and better abilities in word production do not. Instead, they rather benefit from increased salience of the referents by merely bringing their attention to the new words. Note that this finding refers to children with SLI, but not to TD children.

Although this was not a question of the study, it was found that children with SLI exhibited patterns in word learning similar to those of TD children when learning requirements were adapted to children's language profiles. The notion that children with SLI might derive a particular benefit from learning words paired with iconic gestures (as compared to TD children) is not supported by the data. Instead, on a group level, the benefit of observing iconic gestures during word learning for naming taught words was quite similar across groups. As mentioned above, performance of children with SLI merely differed from performance of TD children insofar as children with SLI took some advantage of observing the attention-directing gesture during word retrieval to improve naming performance. From a clinical perspective, it can be concluded that both the iconic and the attentional factors of gesture contribute to word learning in children with SLI. Therefore, gestures should necessarily be considered in word learning interventions. On the one hand they function as a visual support to store new words and to enrich representations, on the other hand they assist children retrieve taught words.

5. Conclusions

The notion that the characteristic property of iconic gestures, namely the resemblance between the gesture and the referent, facilitates word learning in children with SLI as well as in TD children is partially supported by this study. For naming newly taught words after a period of training, it appears that observing both iconic gestures and gestures bringing children's attention to a particular referent contribute to word learning. Questions remain as to whether it is the iconicity of the gesture rather than the attention paid to the new word that served to ease the mapping processes, assessed here through naming tasks without and with gesture cues after some period of slow mapping. Children with SLI evidencing low abilities in noun naming in particular appear to benefit from learning words paired with iconic gestures. Observing iconic gestures exemplifying a semantic feature of the referent during word retrieval qualified as effective facilitation in word-naming tasks. To shed further light on how iconic gesture contributes to word learning in children with and without SLI future work will include additional outcome measures above and beyond naming, such as comprehension and word definition tasks.

Acknowledgments

This research was funded by a grant of the German professional association of logopedics (dbl). Thanks go to children and their parents for participating and to Christina Kauschke and Katharina Rohlfing for invaluable support in developing this study.

References

[1] Alt, M., & Suddarth, R. (2012). Learning novel words: Detail and vulnerability of initial representations for children with specific language impairment and typically developing peers. *Journal of Communication Disorders, 45*(2), 84-97

[2] Alt, M., & Plante, E. (2006). Factors that influence lexical and semantic fast mapping of young children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 49*(5), 941-954

[3] Bishop, D. V. M. (2006). What causes specific language impairment in children? *Current Directions in Psychological Science, 15*(5), 217-221.

[4] Capone Singleton, N. (2012). Can Semantic Enrichment Lead to Naming in a Word Extension Task? *American Journal of Speech Language Pathology 21*(4), 279-292

[5] Capone, N. C., & McGregor, K. K. (2005). The Effect of Semantic Representation on Toddler's Word Retrieval. *Journal of Speech, Language and Hearing Research, 48*(6), 1468-148

[6] Ellis Weismer, S., & Hesketh, L. H. (1993). The Influence of Prosodic and Gestural Cues on Novel Word Acquisition by

Children With Specific Language Impairment. *Journal of Speech and Hearing Research, 36*(5), 1013-1025

[7] Foreman, P., & Crews, G. (1998). Using augmentative communication with infants and young children with Down Syndrome. *Down Syndrome Research and Practice, 5*(1), 16-25

[8] Hostetter, A. B. (2011). When Do Gestures Communicate? A Meta-Analysis. *Psychological Bulletin, 137*(2), 297-315

[9] Kan, P. F., & Windsor, J. (2010). Word Learning in Children With Primary Language Impairment: A Meta-Analysis. *Journal of Speech, Language and Hearing Research, 53*(3), 739-756

[10] Launonen, K. (2003). Manual signing as a tool of communicative interaction and language: the development of children with Down syndrome and their parents. In S. v. Tetzchner & N. Grove (Eds.), *Augmentative and alternative communication: developmental issues* (pp. 83-122). London: Whurr Publishers

[11] Lüke, C., & Ritterfeld, U. (in press). The influence of iconic and arbitrary gestures on novel word learning in children with and without SLI. *Gesture, 14*(2)

[12] McGregor, K. K., Rohlfing, K. J., Bean, A., & Marschner, E. (2009). Gesture as a support for word learning: The case of *under*. *Journal of Child Language, 36*(4), 807-882

[13] Rohlfing, K. J. (2013). *Frühkindliche Semantik*. Tübingen: Narr Francke Attempto Verlag

[14] Sheng, L., & McGregor, K. K. (2010). Lexical-Semantic Organization in Children With Specific Language Impairment. Journal of Speech, Language and Hearing Research, 53(1), 146-159

[15] Stanfield, C., Williamson, R., & Özcaliskan, S. (2014). How early do children understand gesture-speech combinations with iconic gestures? Journal of Child Language, 41(2), 462-471

[16] Tellier, M. (2008). The effect of gestures on second languagememorisation by young children. Gesture, 8(2), 219-235

Breaking boundaries:

How gestures reveal conceptualization of boundary-crossing in Italian

Bjørn Wessel-Tolvig

Centre for Language Technology, University of Copenhagen, Denmark

bwt@hum.ku.dk

Abstract

It has long been considered a linguistic constraint for speakers of verb-framed languages to express boundary-crossing events with manner verb + path satellite constructions. Recent theoretical discussions suggest that Italian may overcome this constraint by expressing directed motion by means of manner verbs + complex locative PPs that can be interpreted as boundary-crossing. We ask whether these constructions are produced in natural speech and how co-speech gestures may help disentangle the ambiguous nature of such expressions. Results show that a small number of boundary-crossing events are expressed with manner verbs + complex PPs. Co-expressive gestures support the claim that these constructions are conceptualized as being boundary-crossing.

Index Terms: Motion events, gesture, boundary-crossing, conceptualization, linguistic encoding, Italian

1. Introduction

We typically gesture when we talk about everyday events like motion. These co-speech gestures are often semantically and temporally tightly related to speech and language [1, 2]. When describing how a man walks down a street, co-occurring gestures tend to reflect the same aspects of the event. Several studies have documented how speakers from different languages gesture differently when describing the same events because of differences in the morpho-syntactic and lexico-grammatic properties of their particular language (for an overview see [3]). Languages differ in how meaning is expressed and how this meaning is mapped onto linguistic form. A widely used typological distinction proposed by Talmy [4, 5] divides languages into – at least – two major groups (e.g., *verb-* and *satellite-framed languages*) with respect to how these language types linguistically express MANNER and PATH of motion. Numerous studies across a variety of different languages confirm this typological division showing striking differences in form-meaning mappings when speaking about motion (for a recent overview see [6-8]).

However, other studies suggest that a strict division of language types is not that clear-cut. Most languages straddle more than one of the Talmyan categories [9], and speakers can use a variety of different constructions that fall within both verb- and satellite-framing lexicalization patterns [10]. In fact, modern spoken Italian, which is considered a verb-framed language, shows emerging signs of satellite-framing constructional patterns, e.g., an optional use of manner verbs + directional satellites to express a figure's movement along a path. But one limitation of satellite-framed event construal in Italian, and in verb-framed languages in general, is the *Boundary-crossing constraint* [11, 12]. According to this linguistic constraint, speakers of verb-framed languages cannot construct boundary-crossing expressions using manner verbs and path satellites within a clause, the defining property of satellite-framed languages. Speakers of verb-framed languages must resort to other syntactical measures to express the manner and path components. However, recent theoretical discussions suggest that Italian may overcome this linguistic constraint to express manner verb + locative PP constructions within a clause that can be *interpreted* as a figure's movement across a spatial boundary. However, as these constructions are locative in nature, we look at co-speech gestures as a reflection of linguistic conceptualization to shed light on whether such constructions are in fact conceptualized as being directional and boundary-crossing.

2. Background

Italian is traditionally categorized as a verb-framed language where path of motion is expressed in the verb root and manner of motion, if expressed at all, is subordinated in PPs or adverbial expressions e.g., a gerund like in (1). The main verb *entrò* - 'entered' and the subordinate manner verb *galleggiando* - 'floating' are divided into two separate clauses.

(1) La bottiglia entrò nella grotta galleggiando
'The bottle entered in.the cave floating'

By contrast, satellite-framed languages need only one clause to express the same information. In these languages manner is encoded in the main verb and path in a satellite to the verb with a verb particle or a PP as in (2).

(2) The bottle floated into the cave

However, typologies are not rigidly fixed, and Italian can express motion in satellite-framed ways with manner verbs and directional verb particles [13] as seen in (3) where the directional component (PATH) is expressed in the verb particle *giù* - 'down'.

(3) Il pomodoro rotola giù per la collina
'The tomato rolls down to the hill'

The 'split system' possibility of verb particle constructions is seen as a developing lexicalization pattern in modern spoken Italian [14], but the crucial difference between verb- and satellite-framed construction possibilities lies in the notion of the boundary-crossing constraint [15]. Crossing a spatial boundary is conceived as "*a change of state, and that state changes require an independent predicate in such languages*" [16 pp:441]. Therefore, speakers of verb-framed languages are required to express path in the main verb and subordinate manner in verbs ('descends *rolling*') or in subordinate manner expressions ('descends *in rotation*'). Both these alternatives impose more complex processing demands, resulting in a tendency for speakers to leave out manner of motion [17].

The main problem is the Italian prepositional system, which is inherently locative. Italian prepositions do not encode the directionality needed to express the path of motion in boundary-crossing situations when manner is mapped onto the main verb as in (4).

(4) *La bottiglia galleggiò dentro la grotta
'The bottle floated inside the cave'

Dentro -'in'/'inside' only encodes locative state, so the bottle does not change direction from outside the cave across the spatial boundary into the cave. Motion is self-contained and it moves at a stationary point (inside the cave). Lacking the possibility of encoding goal of motion in prepositions, Italian speakers are inhibited from using satellite-framed patterns in boundary-crossing situations and therefore obey Talmyan generalization [18]. Özçalışkan [19 pp:18] go even further suggesting that "*the boundary-crossing constraint has the potential to serve as a litmus test that can be applied to languages to show that they are verb-framed*".
Current theoretical discussions challenge the absolute categorization of motion constructions for Italian within boundary-crossing expressions.

2.1. Boundary-crossing interpretation

Recently, it has been suggested that motion events can be constructed using manner verbs + complex locative PPs, which can be interpreted as directional and boundary-crossing [20-22]. According to Folli [20], Italian allows for goal of motion constructions with manner verbs in two ways depending on the lexical properties of the verb itself. Some Italian manner verbs (e.g., *correre* – 'to run') allow for directional meaning combined with locative PPs or complex PPs. The verb identifies the notion of movement from one point to another and the complex PP the PATH and the PLACE (*dentro a* – 'inside to') as in (5).

(5) Gianni è *corso* **dentro a**l parco
Gianni is run inside to.the park
'Gianni ran into the park'

Supporting Folli's claim that Italian may allow for constructions that can be read as boundary-crossing, Cardini [22], in a judgment task, asked participants to judge whether expressions containing different manner verb + preposition combinations expressed directional or locative meaning as in (6) and (7).

(6) Il gatto corse dentro la stanza
'The cat ran into/inside the room'

(7) Il gatto corse fuori dalla stanza
'The cat ran out/outside of the room'

Surprisingly, a majority of the participants interpreted the expressions to be boundary-crossing despite the locative PPs contain no clear directional markers. The results only strengthen the claim that the semantic properties of certain Italian manner verbs combined with locative PPs may give rise to not only directional meaning but also boundary-crossing meaning.
In sum, these studies argue that Italian can overcome the boundary-crossing constraint by expressing motion across a spatial boundary by means of manner verbs and locative PPs pragmatically functioning as directional satellites. But as long no directional features are expressed in the locative PPs, we cannot be entirely certain whether these motion constructions are meant to be directional or merely locative.
A more in-depth investigation of such ambiguous expressions is needed to determine whether speakers conceptualize the events as the traversal of a spatial boundary or not, and if the boundary-crossing constraint truly can serve as a litmus test to test whether languages are in fact verb-framed.

2.2. Linguistic conceptualization

According to Slobin [23], cross-linguistic variation in lexicalization patterns lead speakers of different languages to attend to different aspects of experience when constructing events, i.e. what meanings are selected for expressions and how they are linguistically packaged. This process - also known as *thinking-for-speaking*, targets the possible effects of language on thinking that occurs online in the process of speaking (linguistic conceptualization). Lexicalization patterns are often taken as evidence of linguistic conceptualization, but speech analysis alone cannot account for meaning selection in ambiguous expressions as seen in (5) and determine, without pragmatic clues or inference, what meaning is intended to be conveyed [24]. To resolve this problem, we turn to co-speech gestures as a possible window to linguistic conceptualization.

2.3. Why gestures?

Speech, gesture and language are increasingly seen as planned and processed together in production, and co-speech gestures are often semantically and temporally tightly coordinated with speech, expressing closely related meaning. Several cross-linguistic studies have shown that speakers of typologically different languages speak and gesture differently when narrating the same motion events (for an overview see [3, 25]). Gullberg [25], among others, uses co-speech gestures as a tool to investigate how events are conceptualized by speakers across different languages. Depending on how information is syntactically structured in motion expressions, gestures often reflect the linguistic encoding and the linguistic conceptualization. Meaning expressed in a single clause is likely to be represented by one gesture, and the same meaning expressed in a two-clause construction is often accompanied by two separate gestures [26, 27]. Speech-gesture studies of verb-framed languages often show that speakers use a two-clause construction to express path and manner, and that this syntactic allocation is reflected in two separate gestures: one for manner and one for path. Apparently, speakers of verb-framed languages perceive manner as a separate element that can augment directed motion, whereas speakers of satellite-framed languages see manner as an inherent component of directed motion [17].

Previous findings concerning Italian co-speech gestures indicate that Italian speakers may deploy a double strategy for lexicalization - using both satellite-framed and verb-framed constructions - to a greater extent than speakers of other verb-framed languages, and that co-speech gestures reflect the choice of lexicalization [28-30]. When Italian speakers construct motion in a satellite-framed way, they gesture like speakers of satellite-framed languages. This pattern confirms findings by Kita et al. [31] indicating that the linguistic influence on gestural representations is a result of an online interaction between linguistic conceptualization and gestural representations. Co-speech gestures may therefore shed new light on the conceptualization of ambiguous event construal and on what meaning speakers attend to and select for expression in situations calling for constructions atypical of their language's preferred lexicalization pattern.

2.4. Research question

In this paper, we first ask how motion is generally encoded across boundary and non-boundary-crossing events in Italian, and whether Italian speakers show satellite-framed behavior in boundary-crossing situations as proposed in recent literature. To examine whether such possible satellite-framed constructions (manner verb + locative PPs) are in fact conceptualized as a figure's traversal of a spatial boundary, we use co-speech gestures as a reflection of linguistic conceptualization.

3. Methodology

The participants in this study were a group of 25 native Italian speakers (female 15, mean age 25.96, SD 6.45). All participants were students at the University of Roma Tre and of Roman origin. Their English proficiency was generally at an intermediate level (mean 2.73 of 5, SD 1.19) according to a self-rated L2 English test [32].

3.1. Experimental design

Data was collected using two different sets of elicitation material. Four scenes from the Tomato Man Project [33] containing non-boundary-crossing movement, and four scenes from Boundary Ball [34] showing boundary-crossing movement. The figure, a tomato as seen in Figure 1, either rolled or jumped along a path, up and down a hill, or into and out of a small house. All participants narrated the events of the scenes to a confederate listener with the instruction that a third (naïve) listener would be able to understand and re-narrate the details of the storyline based on their descriptions.

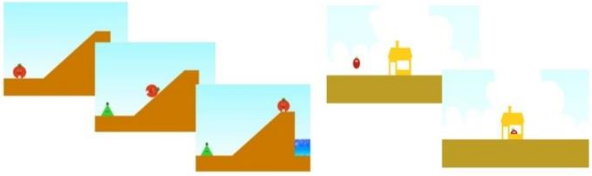


Figure 1: Elicitation material

3.2. Encoding

Speech was tokenized, and the target events labelled as to how they packaged manner and path information syntactically within a clause (8) and or in two clauses (9).

(8) The ball **[bounced down]** the hill
(9) The ball **[descended]** the hill | as it **[bounced]**

Four types (labels) of lexicalization patterns within a target event were defined as seen in Table 1.

Clause type	Example	Labels
One clause	And he **rolls up** the hill	MP
One clause	He **jumps into** the house	MP
One clause	The tomato **rolls**	MO
One clause	He **descends** the hill	PO
One clause	It **goes down** the hill	PO
Two clauses	He **descends** while **rolling**	PO+MO
Two clauses	He **enters** the house **jumping**	PO+MO

Table 1: Speech clause examples and labels

Expressions involving manner verbs + a path denoting satellites were encoded as 'one-clause' manner-path conflated constructions (MP). Constructions only containing manner or path were labelled MO and PO respectively, and expressions containing both mention of manner and path in two separate clauses were defined as a 'two-clause' PO+MO construction.

Gestures were subcategorized into three different types in terms of how information was represented in gesture (see Table 2):

Gesture type	Representation	Labels
Path	Representing only the path of motion with no explicit reference to manner	PG
Manner	Depicting only the manner of motion, that is how the figure moves, with no indication of the path	MG
Manner-path conflating	Conflating both the manner and the path of motion into one single gesture	MPG

Table 2: Gesture examples and labels

3.3. Intercoder Agreement

A second coder (a native Italian speaker) annotated 10% of the corpus and reached a Kappa agreement score of .89 for clause type and .93 for gesture type. The second coder was an experienced speech and gesture coder.

4. Results

The 25 participants produced 209 motion events (198 with gesture) and a total of 275 gestures (1.39 gestures per motion event).

4.1. Clause type results

The overall results for the two event types echo a preference for a verb-framed lexicalization pattern (77.04%), expressing path in the main verb using verbs like *salire*, *scendere*, *entrare*, *uscire* – 'ascend', 'descend', 'enter', 'exit' and subordinating manner, if expressed at all, in adverbial gerunds like *rotolando*, *saltellando* – 'rolling', 'jumping'. Dividing lexicalization patterns based on boundary-crossing (InOut) and non-boundary-crossing (UpDown) events, we observe a more varied lexicalization pattern as illustrated in Figure 2. In the bar plot, we leave out MO (manner only constructions) due to very few occurrences. Although the motion events seem similar, we observe a significant difference in how manner and path are mapped in clauses across the two motion types (X^2 50.8152, df = 3, p-value = < 0.005).

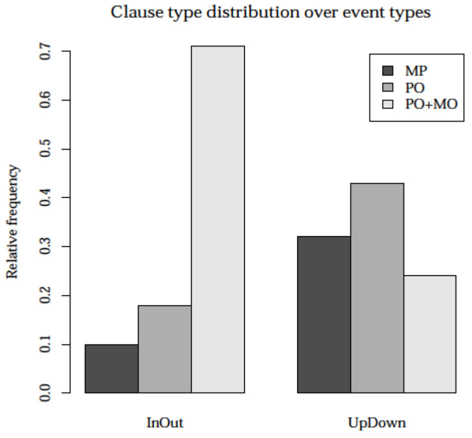


Figure 2: Clause type distribution over event types

For the non-boundary-crossing events (UpDown events), we see a mixed pattern for lexicalizing the event. Path is mainly expressed through the main verb with or without a subordinate manner verb. There is, however, a clear tendency towards expressing motion with a construction typical of satellite-framed languages (30.77%). Here manner is expressed in main verbs like *rotolare, saltellare* – ‘to roll’, ‘to jump’ and path with verb particles like *su, giù* – ‘up’, ‘down’.

The same pattern is not observed in the boundary-crossing situations (InOut events). Manner and path are predominantly separated in two-clause constructions with path verbs and subordinate manner gerunds, e.g., *entra rotolando* – ‘enters rolling’. Although Italian speakers should be limited to verb-framed constructions in the boundary-crossing events, 10.48% of the motion events are expressed in constructions typical of satellite-framed languages.

The speech data results show that Italian speakers do map manner onto main verbs and path onto verb particles or PPs, not only in non-boundary-crossing situations but also in situations where a figure crosses a spatial boundary.

4.2. Clauses types and gestures types

Turning to the gesture data, we examined how the syntactic packaging of manner and path in clauses is reflected in gestures. We observed a clear pattern of co-expressivity between semantic information across modalities. The speech-gesture distribution clearly shows that 1) one-clause constructions are expressed with one gesture and two clauses with two separate gestures, and that 2) co-speech gestures typically express the same information as the information expressed in speech.

For transparency, we visually divide the two event types (UpDown and InOut) into two separate bar charts. Manner-only constructions (MO) and manner-only gestures (MG) are not visually shown in the bar charts due to very few observations. The label 2G is given to gesture constructions in which two separate gestures are expressed within the target event, for example, one for path and one for manner.

Figure 3 shows the absolute frequency of how co-speech gesture types are distributed over clause types in the non-boundary-crossing condition (UpDown).

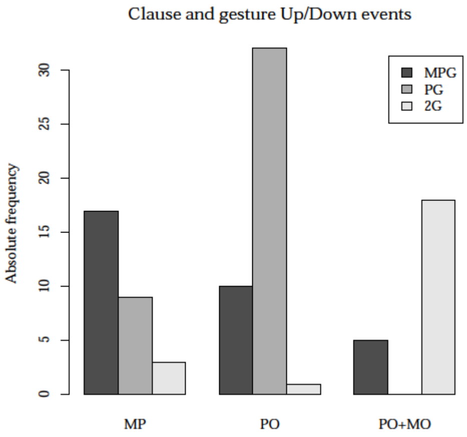


Figure 3: Absolute frequency of gesture constructions over clause type for Up/Down (non-boundary-crossing) events

The data in Figure 3 shows a relationship between clause construction and gesture expression (X^2 = 68.6372, df = 4, p-value = < 0.005). When Italian speakers express only path in speech, e.g., *entra nella casa* – ‘enters the house’ (PO), gestures are typically co-expressive conveying information about path only (PG). Single clause manner + path constructions (MP) are typically reflected in manner-path conflated gestures (MPG) or path gestures (PG).

However, when manner and path occur in two separate clauses, e.g., *entra nella casa rotolando* – ‘enters the house rolling’ (PO+MO), two separate gestures are used (2G), reflecting the conceptual division of the two semantic components. This co-expressive pattern of dividing manner and path in speech and gesture is especially evident in Figure 4, which illustrates the boundary-crossing situations (InOut). When confronted with the boundary-crossing constraint, Italian speakers prefer to express the path of motion in the main verb and subordinate manner most often in the form of an adverbial gerund. The relationship between clause type and gesture types is significant (X^2 = 69.1936, df = 4, p-value = < 0.005). Most interestingly, as observed in the speech data in Figure 2, there are a few manner verb + complex PP constructions (11 of 105 event constructions) amounting to 10.48%.

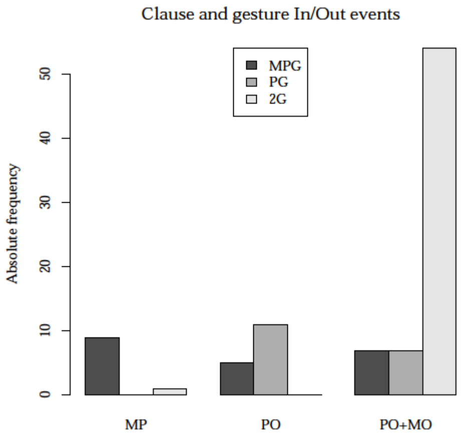


Figure 4: Absolute frequency of gesture constructions over clause type for In/Out (boundary-crossing) events

In these few cases, the tight manner verb and PP constructions (MP) are reflected in manner-path conflated gestures (MPG). Looking more qualitatively at the MP expressions and representation of gesture, we see how the manner-path conflating gestures are typically mapped across the manner verb and the locative PPs as in (10). The stroke of the gesture is indicated with brackets.

(10) Il pomodoro **rotola fuori da**lla casa
[--MPG gesture--]
‘the tomato **rolls out/outside of**.the house’

Although the locative prepositions in (10) do not license directional movement, the co-speech gesture expresses both the manner of rolling and the path of the figure out of the house in one single gesture. With the co-expressivity between speech and gesture in situations where speakers construct boundary-crossing events using manner verbs and locative PPs, we argue that the event is conceptualized as a *figure’s movement across a spatial boundary* and not as locative motion (no change of location). The semantic content of the gestures reinforces the argument that such constructions are valid in Italian for the manner verbs used in this study.

5. Discussion

Because speakers of verb-framed languages usually express path through the main verb, they typically resort to subordinated manner verbs to express manner of motion. Including an additional syntactic element makes the event conceptually more complex to process, which increases the tendency for speakers of verb-framed languages to leave out manner of motion [35]. This is not entirely evident in this study. Manner information is omitted more often in the descriptions of non-boundary-crossing events, but both manner and path are univocally included both in one-clause manner verb + PP constructions and in two-clause path verb + subordinate manner verb constructions across both event types.

The results paint a picture of a language that does not conform exclusively to the Talmyan typology. Speakers of modern spoken Italian widely use an option for linguistic encoding not typical of the verb-framed lexicalization taxonomy. Approximately 30% of the expressions in the non-boundary-crossing events were constructed with manner verbs and directional satellites, a systematic lexicalization pattern typical of satellite-framed languages. This event construction allows Italian speakers to easily include manner of motion in descriptions. This indicates that Italian speakers, or at least the speakers in this study, *do* focus on manner in motion descriptions and *also* resort to lighter syntactical constructions to express motion when available

The findings also support the hypothesis that speakers of Italian can use manner verb + locative PPs to express a figure’s movement across a spatial boundary. This construction is not only hypothetically possible but actually produced in spontaneous speech. This option puts the boundary-crossing constraint into question, especially as a litmus test to ultimately categorize languages in the verb-framed category.

In line with many other studies, we observe co-expressivity between information expressed within the clause and the information represented in gesture. When Italian speakers construct motion atypical of their Talmyan type, gestures reflect the choice of lexicalization. This supports previous findings by Kita et al. [31] which suggest that gestural expressions are determined by the online choice of syntactic packaging of manner and path information rather than by language-specific habitual conceptual schemas. Moreover, the few instances of manner verb + PP constructions in the boundary-crossing events accompanied by manner-path conflating gestures indicate that the construction is conceptualized as being boundary-crossing and not locative motion. The combination of main manner verbs and locative PPs can be, and are, used to express boundary-crossing in Italian, although at very small frequencies.

One question remains: if manner verbs + locative PPs can express spatial crossing, why is this pattern not more widespread in Italian? According to the principles of speech economy, speakers should choose constructions which impose lighter conceptual processing. This is predominantly seen in the non-boundary-crossing situations where single clausal constructions are expressed through both manner verbs + PPs and path-only verb-constructions. But in the boundary-crossing events, we primarily see two-clause constructions. One possible answer is that only some Italian manner verbs contain an internal element of directionality – not to be confused with path verbs – which allows them to combine with locative PPs, giving rise to directional and boundary-crossing interpretation. It is debatable whether pure manner verbs have the same directional property [18]. As the manner verb + PP combination is ambiguous with one group of manner verbs and possibly not allowed for with another group of (pure) manner verbs, speakers of Italian could avoid atypical constructions by pursuing standard verb-framed lexicalization patterns. Furthermore, speakers are trained by their native linguistic experience to structure the elements of motion in a particular way typical of their language [23]. In a sense, native speakers learn to prioritize certain aspects of motion and verbalize them in a certain way.

But the fact that some Italian speakers use manner verb + complex PP constructions to break linguistic boundaries combined with co-expressive gestures reflecting the directional movement, suggests that a typical satellite-framed construction is valid for expressing boundary-crossing meaning in Italian.

These findings naturally call for further investigation into event construction but also emphasize that gesture may serve as a powerful tool to study linguistic conceptualization.

6. Conclusion

We investigated how speakers of Italian express motion events depending on the spatial properties of the elicitation material. We found that Italian speakers prefer typical verb-framed lexicalization patterns, but that the speakers in this study showed signs of an emerging satellite-framed system at least in non-boundary-crossing events, but also in situations involving boundary-crossing. We confirm the hypothesis that goal of motion expressions can be constructed with manner verb + complex locative PPs. We used co-speech gestures as a tool to investigate linguistic conceptualization of event construction, and we found that gestures may help to clarify situations with ambiguous meanings. Co-speech gestures support the claim that manner verb + locative PP constructions are conceptualized as boundary-crossing events. Overall the findings in this paper prove that speakers have a wide range of constructional possibilities at their disposal when constructing meaning in motion events, and that a particular language may use constructions pertaining both to satellite-framed and verb-framed languages. In the end, the question is not what speakers can or cannot do with language, but rather what they *do* with language.

7. Acknowledgements

Thanks to Patrizia Paggio, University of Copenhagen, Isabella Poggi and Laura Vincze, Roma Tre, Emanuela Campisis, MPI Nijmegen and Francesca Gregorino, University of Turin. This research is funded by The Danish Council for Independent Research.

8. References

[1] A. Kendon, "Gesture and speech: two aspects of the process of utterance," in *Nonverbal Communication and Language*, M. R. Key, Ed., ed The Hague: Mouton, 1980, pp. 207-227.

[2] D. McNeill, *Hand and mind: what gestures reveal about thought*. Chicago: University of Chicago Press, 1992.

[3] S. Kita, "Cross-cultural variation of speech-accompanying gesture: A review," *Language and Cognitive Processes*, vol. 24, pp. 145-167, 2009/02/01 2009.

[4] L. Talmy, "Semantics and syntax of motion," in *Language typology and syntactic description, Vol. 3,*

Grammatical categories and the lexicon. vol. 3, T. Shopen, Ed., ed Cambridge: Cambridge UniversityPress, 1985, pp. 57-149.

[5] L. Talmy, *Toward a Cognitive Semantics: Typology and Process in Concept Structuring* vol. II. Cambridge: The MIT Press, 2000.

[6] Z. Han and T. Cadierno, *Linguistic relativity in SLA: Thinking for speaking*. Bristol/Buffalo/Toronto: Multilingual Matters, 2010.

[7] J. Goschler and A. Stefanowitsch, *Variation and change in the encoding of motion events*. Amsterdam: John Benjamins, 2012.

[8] A. Pavlenko, *Thinking and speaking in two languages*. Clevedon, UK: Multilingual Matters, 2011.

[9] J. Beavers, B. Levin, and S. Wei Tham, "The typology of motion expressions revisited," *Journal of Linguistics,* vol. 46, pp. 331-377, 2010.

[10] W. Croft, J. Barðdal, W. Hollmann, V. Sotirova, and C. Taoka, "Revising Talmy's typological classification of complex event constructions," in *Contrastive Studies in Construction Grammar*, H. C. Boas, Ed., ed Amsterdam: John Benjamins, 2010, pp. 201-235.

[11] J. Aske, "Path predicates in English and Spanish: A closer look," presented at the Proceedings of the Berkeley Linguistic Society 15, 1989.

[12] T. Cadierno and L. Ruiz, "Motion events in Spanish l2 acquisition," *Annual Review of Cognitive Linguistics,* vol. 4, pp. 183-216, 2006.

[13] C. Iacobini, "Grammaticalization and innovation in the encoding of motion events," *Folia Linguistica,* vol. 46, pp. 359-385, 2012.

[14] C. Iacobini and F. Masini, "The emergence of verb-particle constructions in Italian: locative and actional meanings," *Morphology,* vol. 16, pp. 155-188, 2006.

[15] D. I. Slobin and N. Hoiting, "Reference to movement in spoken and signed languages: Typological considerations," presented at the Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society, 1994.

[16] D. I. Slobin, "Mind, code, and text," in *Essays on language function and language type: Dedicated to T. Givón*, J. Nybee, J. Haiman, and S. A. Thompson, Eds., ed Amsterdam/Philadelphia: John Benjamins, 1997, pp. 437-467.

[17] D. I. Slobin, "The many ways to search for a frog: Linguistic typology and the expression of motion events," in *Relating events in narrative: Typological and contextual perspectives* S. Strömqvist and L. Verhoeven, Eds., ed Mahwah, NJ: Lawrence Erlbaum Associates, 2004, pp. 219-257.

[18] J. Mateu and G. Rigau, "Verb-particle constructions in Romance: A lexical-syntactic account," *Probus,* vol. 22, pp. 241-269, 2010.

[19] S. Özçalışkan, "Ways of crossing a spatial boundary in typologically distinct languages," *Applied Psycholinguistics,* vol. FirstView, pp. 1-24, 2013.

[20] R. Folli, "Complex PPs in Italian," in *Syntax and Semantics of Spatial P.*, A. Asbury, J. Dotlacil, B. Gehrke, and R. Nouwen, Eds., ed: John Benjamins Publishing Company, 2008, pp. 197-221.

[21] R. Folli and G. Ramchand, "Prepositions and Results in Italian and English: An Analysis from Event Decomposition," in *Perspectives on Aspect*. vol. 32, H. Verkuyl, H. de Swart, and A. van Hout, Eds., ed: Springer Netherlands, 2005, pp. 81-105.

[22] F.-E. Cardini, "Grammatical constraints and verb-framed languages: The case of Italian," vol. 4, ed, 2012, p. 167.

[23] D. I. Slobin, "From ‘‘thought and language’’ to ‘‘thinking for speaking’," in *Rethinking linguistic relativity*, J. J. Gumperz and S. C. Levinson, Eds., ed Cambridge: Cambridge University Press, 1996, pp. 70-96.

[24] P. Athanasopoulos and E. Bylund, "The 'thinking' in thinking-for-speaking: Where is it?," *Language, Interaction & Acquisition,* vol. 4, pp. 91-100, 2013.

[25] M. Gullberg, "Thinking, speaking and gesturing about motion in more than one language," in *Thinking and speaking in two languages*, A. Pavlenko, Ed., ed Bristol: Multilingual Matters, 2011, pp. 143-169.

[26] G. Stam, "Thinking for speaking about motion: L1 and L2 speech and gesture," *International Review of Applied Linguistics,* vol. 44, pp. 143-169, 2006.

[27] A. Özyürek, S. Kita, S. Allen, R. Furman, and A. Brown, "How does linguistic framing of events influence co-speech gestures?: Insights from crosslinguistic variations and similarities," *Gesture,* vol. 5, pp. 219-240, // 2005.

[28] N. Rossini, "Phrasal verbs or words? Towards the analysis of gesture and prosody as indexes of lexicalisation," presented at the On-line Proceedings of the 2nd ISGS Conference "Interacting Bodies", Lyon, France, 2005.

[29] F. Cavicchio and S. Kita, "Gestures Switch in English/Italian Bilinguals," in *MMSYM2014*, Tartu, Estonia, 2014.

[30] B. Wessel-Tolvig, "Up, down, in & out: Following the Path of motion in Danish and Italian," presented at the Proceedings of the 1st European Symposium on Multimodal Communication, Valletta, Malta, 2014.

[31] S. Kita, A. Özyürek, S. Allen, A. Brown, R. Furman, and T. Ishizuka, "Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production," *Language and Cognitive Processes,* vol. 22, pp. 1212-1236, 2007.

[32] M. Gullberg and P. Indefrey, "Language Background Questionnaire. Developed in The Dynamics of Multilingual Processing," Max Planck Institute for Psycholinguistics, Nijmegen, 2003.

[33] A. Özyürek, S. Kita, and S. Allen, "Tomato Man movies: Stimulus kit designed to elicit manner, path and causal constructions in motion events with regard to speech and gestures.," ed. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics, Language and Cognition group, 2001.

[34] B. Wessel-Tolvig, "Boundary Ball: An animated stimulus designed to elicit motion with boundary crossing situations.," ed. University of Copenhagen, 2013.

[35] D. I. Slobin, "Language and thought online: cognitive consequences of linguistic relativity," in *Language in mind: Advances in the study of language and thought,* D. G. S. Goldin-Meadow, Ed., ed Cambridge, MA: MIT Press, 2003, pp. 157-192.

Author index

GESPIN 2015

NANTES

GESPIN 4