Structuring Information through Gesture and Intonation[*]

Stefanie Jannedy & Norma Mendoza-Denton

Humboldt Universität zu Berlin & University of Arizona

Face-to-face communication is multimodal. In unscripted spoken discourse we can observe the interaction of several "semiotic layers", modalities of information such as syntax, discourse structure, gesture, and intonation. We explore the role of gesture and intonation in structuring and aligning information in spoken discourse through a study of the co-occurrence of pitch accents and gestural apices. Metaphorical spatialization through gesture also plays a role in conveying the contextual relationships between the speaker, the government and other external forces in a naturally-occurring political speech setting.

Keywords: Gesture, Intonation, Spoken Discourse, Narrative Structure, Political Speech, Affect.

1 Introduction

It has been widely accepted that the gesture and intonation systems correlate, both aiding in the structuring of verbally rendered discourse (Cassell, 2000;

[*] Many thanks to Emily Kidder, Ashley Stinnett, Mourad Mjahed, Bernd Pompino-Marschall & Ulrike Kölsch for valuable comments. Video fieldwork assistance was provided by Ashley Stinnett and Aomar Boum. Labelers at the University of Arizona included Ruby Basham, Jason Haugen, Hannah Jones, Emily Kidder and Heidi McLaughlin for ToBI Labelling; Natasha Gibson, Kerry Murray, and Cara O'Connor for gesture labelling. Bryce Coughlin generated the lexical transcription. This research would not have been possible without the consent and cooperation of Congressman Jim Kolbe, his staff in Tucson, Arizona and in Washington, D.C., and anonymous constituents at various Town Hall Meetings from 2000-2001. This research was supported by the University of Arizona Vice Provost Faculty Research Grant (to Mendoza-Denton), the University of Arizona Department of Anthropology Riecker Grant Program (to Mendoza-Denton), and the University of Arizona Undergraduate Honors Program Grant (to Stinnett for fieldwork). Further support comes from SFB632-D3 (to Jannedy) and from the German Research Foundation/Humboldt Universität zu Berlin.

Interdisciplinary Studies on Information Structure 03 (2005): 199–244
Ishihara, S., M. Schmitz, and A. Schwarz (eds.):

Loehr, 2004). Yet most of the studies on co-speech gesturing analyzed in the literature result from experimental elicitation and give rise to gestures that are narrative and/or descriptive in nature. That is, participants in laboratory settings are asked to describe something rather concrete such as fish-trapping constructions (Enfield, 2003), or to narrate preselected cartoon strips or films (McNeill, 1992). In this study, we analyze a video recorded sample from a corpus of spontaneous, naturally-occurring data gathered at public Congressional Town Hall Meetings (THMs) in Tucson, Arizona, in which a speaker engages in political discourse, a task much more abstract and goal-directed than elicited narrative, and one in which the speaker is trying to put a political viewpoint across. The speaker (whom we call by the pseudonym *Mary-Jane*) is a woman from Arizona who appears to be in her early forties; her primary interlocutor (U.S. House of Representatives Congressman Jim Kolbe, Republican, Arizona 5th district) is a middle-aged man in his late fifties. They stand in a constituent-representative relationship to each other in the United States political system. In this THM, the co-speech gesturing deployed by Mary-Jane is used not only to describe the political landscape as she sees it, but to persuade, cajole, shame, provide evidence and otherwise convince the interlocutor and the audience to adopt her point of view. Our findings indicate that gesturing not only correlates with grammar and supports the structuring of the information that is acoustically rendered; at the same time, use of the visuo-spatial field conveys information on the relationship posited by the speaker between the government, its constituents and outside forces.

Kendon (1996) thinks of gestures as a ‘spill-over’ effect from the effort of speaking. For him, gesture is a separate and distinct mode of expression with its own properties which can be brought into a cooperative relationship with spoken utterances and used in a complementary way. He suggests that a study on how phrases of gesture and phrases of speech are related would throw useful light on how information is structured. Bolinger (1986, p. 199) proposes that “gesture and speech/intonation are a single form in two guises, one visible and the other audible”, and argues that gesture and speech stem from the same semantic intent. McNeill et al. (2001, p. 23) argue that “the organization of discourse is inseparable from gesture and prosody” and that “[they are] different sides of a single mental communicative process” (cf. Enfield, 2003, p. 45-46). The results of McNeill’s experimental studies indicate that motion, prosody and discourse structure are integrated at each moment of speaking.

Languages such as English or German, unlike for example Swahili (McNeill, 1992), use few morphosyntactic cues to structure discourse. Instead, information is structured primarily by syntactic and prosodic means. Vallduví and Engdahl (1996) discuss how information is packaged into prime units, and give evidence for the difference between English (which uses primarily intonation) and Catalan (using primarily syntax) in the packaging of information.

The role of gesturing during speaking and its role in the structuring of discourse has until recently been largely unexplored. Yet the relationship of gesture to speech increasingly captures the interest of human-computer interface designers aiming to model the movements of animated agents, since it has been shown that gestures facilitate information comprehension (Beattie & Shovelton 1999). Gesture researchers have further shown that subjects can recall information that was selectively presented in the gestural but not in the verbal channel (Kelly et al., 1999; Cassell et al., 1999; McNeill, 1992), all without being able to remember what channel the information was presented in. These results are provocative because they imply that as far as the design of animated agents is concerned, greater information density can be achieved by presenting part of the information acoustically and part of it (possibly complementary or reinforcing information) visually, bringing us closer to modelling the online workings of face-to-face conversations. A thorough description of gestural-speech co-dynamics is thus necessary for an algorithmic approach to be formulated, and essential in the design of automated agents and in achieving some naturalness in their movements (Cassell, 2000 and references therein).

The purpose of this research is to describe and understand the timelines involved in a case study of the complementary presentation of information. Part of the information is presented through the speaker's verbal stream, while the broader and complementary political setting and the assumptions on which it rests are presented gesturally. We have independently verified that both of these meanings come across by asking audiences of graduate and undergraduate students in English-speaking universities (at the University of Arizona, the Ohio State University, and at University College Dublin in Ireland) to describe what they think the speaker was trying to get across. Three different modes of presentation of our data have been judged by the student populations:

1) Acoustic and visual information: Audiences who have seen the video and listened to the audio have repeatedly described the complementary information presented on both channels.

2) Acoustic information only: Audiences who have been presented with the audio portion describe the main assertions presented in the spoken discourse with no reference to the larger political landscape.

3) Visual information only: Audiences who have only seen the video without the audio merely impute anger and emotionality to the speaker.

These results highlight one well-known fact about the relationship of gesture when it is ancillary to spoken discourse: while spoken discourse has a high referential resolution, that is, it is able to pick out referents with relatively little ambiguity, gesture has a low referential resolution, so most of the information presented gesturally is supportive to speech and not recoverable solely from the gestural channel (an obvious exception to this being gestural systems that are full-fledged languages, such as American Sign Language). Gestures and spoken discourse are thus complementary in nature, and have different affordances and limits to their ability to present information. Harper et al. (2000) have found that in a study of 3D multiplayer wargame interactions, gesture was used for much more than just simple deictic functions. They explain:

> …language facilitates complex queries with the ability to express quantification, attribute and object relations, negation, counterfactuals, categorization, ordering, and aggregate operations. Gesture is more natural for manipulating spatial properties of objects (size, shape, and placement) in graphical environments. (Harper et al. 2000:3)

In Harper et al.'s study, subjects participated in a wargame simulation which involved some players working around a scaled, 3D model of a battlefield, while other players were consulted remotely. In post-game discussions, participants remarked on the difficulty of communicating with those not working around the 3D model being used (2000:5). Because of their heavy reliance on definite descriptions in combination with gestures in the face-to-face game, players encountered difficulties when they were restricted to using only the verbal channel.

We hold there to be a parallelism between gesture and speech, both of them carrying meaning on at least three different planes: structure, content, and social meaning.

The acoustic signal is an enormously rich source of structured information. from a phonetic-phonological point of view, there are positional restrictions or co-occurrence restrictions of phonemes, the phonotactics that make up a specific language variety. There is a specific prosodic or rhythmic structure, such that some parts of a word are uttered louder and longer lending a specific syllable more perceptual prominence either to indicate lexical stress or informational salience. At the same time, obviously, the acoustic signal also contains other grammatical information. That is, lexical or semantic content is being transported by the choice or words spoken, and pragmatic content is chosen by the context the words are uttered in. Further, the acoustic signal transmits social content, that is, it also carries social information pertaining to the speaker (age, gender, ethnicity, etc.) and their interlocutors.

Co-speech gestures correlate with grammar, that is, they correlate with grammatical structure by indicating beats (correlated with pitch accents), or gestural phrases (correlated with syntactic phrases). These same gestures also carry semantic and pragmatic content by being deictic or emblematic or just emphatic beats on parts of the information emphasized by the speaker. Simultaneously, these same gestures also transmit social content in terms of how the speaker recounts and relates to the world or abstract universe surrounding them.

The gestural information stream is able to work online and incorporate context to take advantage of props and of the spatial location of the audience. We argue that the communicative constraints of the sociolinguistic situation in our case study maximized the need for the simultaneous presentation of information, since the speaker was only grudgingly given the floor in the first place, was constrained in the time she was allotted for her turn at talk; while the interlocutor constantly threatened to interrupt her, attempted to shift his attention away from her and to cut her off. Thus, the sociolinguistic pressure was great for the speaker to pack as much information as possible into her short turn at talk. For our purposes, this results in a naturally-occurring situation where time and interactional constraints push the limits of both intonational and gestural information packaging.

2 Background

2.1 What do hand, arm, and mouth movements have in common?

Both the motor theory of speech production/perception (Liberman and Mattingly 1985) and the articulatory phonology account of speech as a stream of articulatory gestures (Browman and Goldstein, 1990, 1992) place articulatory movements produced during speech and "paralinguistic" gestures on the same cognitive plane, as they are both thought to be *coordinated patterns of goal-directed articulatory movements* (Tatham 1996). These approaches require that speech movements and gestural movements be accounted for by the same mechanisms, since it is assumed (contra Chomsky 2000, Fodor 1983, Marantz to appear) that the structures of language are not modular or unique in comprising a "language organ", but rather that these structures were derived and modelled on the pre-existing neural systems which had evolved for the control of body movement. Rizzolatti et al. (1988, 1999) posited a class of premotor neurons, the "mirror" neurons based on their experiments with monkeys. They explain: "…with this term [mirror neurons] we define neurons that discharge both when the monkey makes a particular action and when it observes another individual (monkey or human) making a similar action. […] Transcranial magnetic stimulation and positron emission tomography (PET) experiments suggest that a mirror system for gesture recognition also exists in humans and includes Broca's area (1988:92)".

The results of these earlier studies of motor action have been replicated, and links shown between the sensorimotor system and the acoustic system both in monkeys and in humans (Berthoz, 1997: Fadiga & Craighero, 2003). The relationship between speech and limb movements is thus taken for granted, since both derive from elementary motor programs or *articulatory gestures*. According to Berthoz (1997:176) "Le mouvement est donc organisé à partir d'un répertoire de synergies qui compose autant d'actes possibles…une bibliothèque de mouvements facilement déclenchables. [*Movement is then organized beginning with a synergistic repertoire which makes up possible actions…making up a library of movements which are easily carried out.]"*

Both speech gestures (Browman & Goldstein, 1990) and hand/arm gestures then are finely coordinated and stem from higher level cognitive processes through which information is structured. While there is no one-to-one correspondence between form and meaning in gesturing, gestural organization maintains a tight link with the semantics of speech. That is, gestures strongly correlate with grammar and grammatical structures: it was found that the stroke (peak of effort) of a gesture occurs with the intonationally most prominent syllable of the aligned speech segment (Kendon, 1980; McNeill, 1992; Cassell, 2000; Loehr 2004).

The significance of the motor theory of action for our study is the following: we believe that when a speaker executes a particular gesture, the embodied nature of the gesture causes the interlocutor to generate an internal representation of the movements in order to decode and interpret the spatial field (concrete or abstract, see our discussion of Krifka 2005 below) depicted by the speaker. This is essentially a parallel to what the motor theory holds for speech (Liberman and Blumstein, 1990:147), where we practice "analysis by synthesis"

(Halle and Stevens 1959) in generating an internalized representation of the incoming signal.

Thus, given an interlocutor that shares spatial conventions, speakers may gesture and have their concrete and abstract meaning understood without uniqueness of referents. Spatial conventions exhibit cross-cultural variability: Haviland (1993), for example, has found that narrative retellings involving co-speech gesturing in the Gugu Yimidhirrh aboriginal language spoken in Queensland, Australia, exhibit absolute directionality, that is to say, when a speaker retells an incident their deictic pointing will refer to the same cardinal direction in which the relevant incident took place. Similar findings have been reported for Amerindian Languages such as Tzotzil (Brown and Levinson 1993) and Assiniboine (Farnell 1995). Other speakers, especially those from Western cultures, exhibit relative directionality, so that their gestures tend to be located relative to the ego (though some flexibility also exists here, with speakers able to present gestural situations from different points of view (McNeill 1992:190). There may additionally be abstract topic/comment structuring in the directions of pointing or in the handedness of a gesture (Krifka 2005). We hold that topic/comment structures will be present in preferential hand-gesturing, but will be forced to interact with culturally-conventionalized gesturing directionality. This would predict that, given the same story stimuli, speakers from absolute-directionality gesture traditions would encode topic/comment gestures differently from those of ego-centric gesture traditions. We exhort the next generation of linguists and linguistic anthropologists to conduct these experiments.

2.2 A typology of gestures

To gain a better understanding of the classification of gestures, a short digression into gesture theory seems necessary. From a kinetic point of view, gestural movements are described to have obligatorily three and at most five phases (McNeill, 1992): The *preparation phase* (optional) marks the beginning of the motion in which the parts of the body involved in the gesture leave the neutral position and move to the position necessary for the upcoming gesture. The *pre-stroke hold* (optional) is the position of the hand/arm at the end of the preparation phase before the beginning of the stroke. The *stroke* (obligatory) is the climax or peak of effort of the gesture. It is one of the most recognizable components of a gesture, it is synchronized with the linguistic forms, such as accented syllables it is co-expressive with. The *post-stroke hold* is the position the hand/arm remains in when the co-expressive spoken utterance is delayed. The *retraction* is the end of the motion in which the parts of the body involved in the gesture return to neutral positions. The *beat* gesture is smaller and is described to only have 2 phases which are typically flicks with the wrists or fingers. In the case of Mary Jane, we will see that her impassioned argument motivates her marking beats with her entire upper torso (Panel 4.2).

From a semantic or pragmatic point of view, gestures can have different attribute values:

1) *Deictic* gestures point towards a concrete or abstract referent;
2) *Beats*, sometimes called *batonic gestures* because they resemble a conductor keeping an orchestra in time, are rhythmic gestures that have no specific form but which are synchronized with the speaker's utterances.
3) *Iconic* gestures depict a concrete object or event in a narrative;
4) *Emblematic* gestures replace speech (for instance, shaking one's head to mean "no").

5) *Metaphoric* gestures represent an abstract idea.

McNeill (1992) calls the recurring combination of the same gestures with prosody and discourse organization *catchments*. These catchments are recognized from recurrences of gesture form features over a stretch of discourse (two or more gestures with partially or fully recurring features of shape, movement, space, orientation, dynamics etc.) and serve to offer clues to a cohesive linkage in the text in which it occurs.

Rather than assuming that gesturing only serves the interlocutor in structuring acoustically rendered information, we take it that gesturing aids both the speaker and the interlocutor: recall Harper et al.'s (2000) wargame simulations referenced above. It has also been observed that gestures occur no less frequently when talking over the telephone and the speaker cannot be seen by the listener (Cosnier 1982; Rimé 1982).

We also assume that the very same gestures that are so tightly aligned with grammar and prosody serve as a device to mediate between the speaker and the world. What does this mean? By having gestures that are ego-centered, and placing herself in the middle of a depiction of the political landscape, Mary-Jane sketches *through gestures* her view of an idealized public sphere, and of the relationship between a constituent and the broader powers of government.

2.3 Intonational Grammar

The grammar of intonation we are assuming goes back to Pierrehumbert (1980) and was formalized as a transcription system for prosodic annotation summarized in Beckman & Ayers (1994). According to this grammar of intonation there are pitch accents, intermediate phrases and intonation phrases. The grammar of intonation can be summarized as follows: each intonation phrase (marked with ‘%’ at the right edge) must contain at least one intermediate phrase (marked with a ‘-‘ at the right edge), which in turn must contain at least one pitch accent (marked with a ‘*’).

Pitch accents are associated with the stressed syllables of words. These accents mark local prominences above the level of the word in an utterance. There is a fixed inventory of pitch accents, they can have different tonal shapes that are marked with labels such as H*, L* or L+H* etc. The ‘*’ indicates that there is a pitch accent which is defined as a tonal target. The tonal events between pitch targets are accounted for by interpolation between these pitch targets. For example, one should observe a falling fundamental frequency contour when a H* tone target is followed by a L* tone target. Downstep is a phonologically triggered process by which a H* accent is downstepped (!H*) when it occurs subsequent to the downstep trigger. Often this trigger is a bitonal pitch accent type such as L+H* or L*+H accent. By definition though, the domain of downstep is the intermeditate phrase, thus, the trigger must occur in the same intermediate phrase as the downstepped accent. Downstepping is a compression of the pitch range lowering the F0 targets for the H* accents following the downstep trigger.

An intermediate phrase is a minor phrase and consists of one or more pitch accents plus a phrase accent associated with the right edge. This phrase accent can be either L or H. An intonational phrase consists of at least one or more intermediate phrases. The right edge of the intonational phrase has a phrase tone, taking the shape of L or H. These edge tones determine the shape of the F0 contour between the last pitch accent and the end of the phrase.

3 Data and Methods

We utilize video data collected in 2000-2001 for an ethnographic study of Congressional Town Hall Meetings (THMs) in Tucson, Arizona. The data forms part of a fieldwork project on political discourse, language and power. A total of approximately 20 hours of data was collected at locations around Southern Arizona and Washington, D.C. THMs are public fora announced in the media, on websites and through fliers to homes in the neighborhoods that are represented by Kolbe. A THM in this district normally lasts one and a half hours to two hours and is led by Congressman Kolbe, typically taking the form of an initial period of question-writing by the audience on slips of paper circulated by Kolbe staffers. This is followed by a rehearsed monologue from Rep. Kolbe in which he states as his aim the updating of his constituents on important happenings in Washington, D.C. Then Kolbe selects some of the questions that constituents have written out as questions to be answered without calling on anybody specifically. After this he might take a couple of spontaneous questions, run to the end of the allotted time and then invite those who wanted to talk to him further to stay and discuss matters after the official THM ends.

The video data of Rep. Kolbe and Tucson citizens that we analyze for this paper was recorded simultaneously with 2 video cameras, both located to the left of the audience (from their perspective), one pointed toward Rep. Kolbe and the other more aimed more generally toward the audience. Rep. Kolbe was miked with a lavaliere microphone plugged into camera 1, and the audience sounds was captured by a microphone mounted on camera 2. By aligning the sound tracks of the two videos, we were able to synchronize the videos exactly so as to gain accurate descriptions of the hand and arm movements from two different angles. Both camera 1 and camera 2 capture the audience from different vantage points.

On this particular occasion, on a Saturday morning in February 2001, the THM was held in the cafeteria of a midtown school. The congressman introduced the researchers and advised participants in the THM that they were being video taped for a research project and that their participation was strictly voluntary. If any of the constituents objected to being taped, they were encouraged to approach the researchers after the THM. None did.

The data we selected for microanalysis for this study lasts exactly 130 seconds (a reasonable amount of data within the gestural analysis literature (See Loehr 2004 for a discussion)), and was transcribed according to the ToBI intonational transcription framework in addition to being subjected to a modified McNeill-style gestural transcription. Eleven tiers of body movement and gestural transcription were coded.

3.1 ToBI and Gesture Transcriptions

Trained linguistics and anthropology students at the University of Arizona worked on transcribing the data on several different levels by using the program Praat. Orthographic transcriptions were done and then checked by another student. Prosodic transcriptions were made in a team of three students of linguistics by majority vote: They were trained to transcribe intonational events within the ToBI (tone and break index) framework based on the Pierrehumbert (1980) system of English intonation. In cases of uncertainty or disagreement, they discussed the issue until a consensus was reached. Then, all prosodic transcriptions were checked by an experienced labeler of US-English (one of the authors). It must be noted though that the data was very difficult to annotate due to the fact that the recordings were made in a naturalistic setting. In addition to room noises (random background noise) which shows up in spectrograms as energy in all frequency regions, the audio data we analyzed contains claps and coughs from members of the audience, partial interruptions by the congressman and other noises that are unidentifiable. In figure 1 below, taken from our Praat display, we show the sound pressure wave (waveform) in the upper display, and a spectrogram (individual frequency bands) which was overlain with the fundamental frequency trace (F0) necessary for the tonal analysis. This is how we have coded lines 23-29 of the full transcription included as an appendix to this paper.

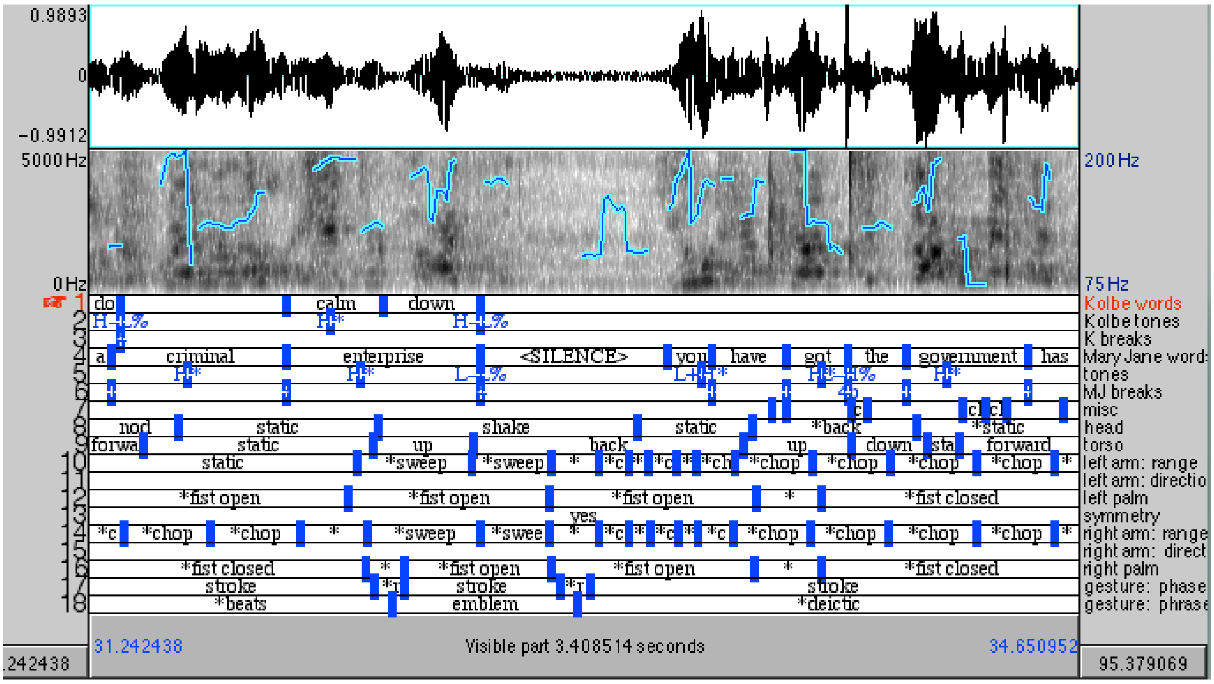


Figure 1: Praat display of multitier transcription of intonation and gesture.

Though there is a growing body of literature on gesture, no standard transcription system for manual gestures has been agreed upon and described. There are a couple of systems which are in relatively wide use; within language studies, the most prominent is McNeill's system, developed expressly for gestures. Based on gestural primes proposed by McNeill (1992), we transcribed movement of the left and right arm and the left and right hands on 6 tiers (range of movement, direction of movement, palm configuration for each side). In addition, head and torso movements were transcribed as well as whether or not there was a symmetry in the movements of both hands and arms. We also annotated the gestural phases which have kinetic properties (preparation vs. stroke, for example) as well as the gestural phrases which are more semantic or pragmatic in nature (deictic vs. metaphoric etc.).

In order to have an exact alignment of audio and video, the data was looked at in ANVIL, a program that allows video and audio to be time aligned for transcription purposes. For the purpose of our study it became crucial not only to mark intervals during which gestural movements were performed, but to mark points as well. We were faced with the issue of data reduction since it seemed impossible to gain any insights from the amount of data available to us. As an initial step, we wanted to investigate if and when and how often pitch accents would co-occur with gestural movements. Since we had mainly coded intervals, and pitch accents are by default point events in time, we added another transcription level in which we coded what Loehr (2004) has called the *apex*.

The apex is the point in the hand or arm gesture during which (in Task Dynamics terms) the equilibrium of a particular gestural movement is reached (Browman & Goldstein, 1990). Task Dynamics holds that gestures are defined in terms of tasks (extend arm, point with finger etc.) which are the coordinated movements of several limbs. Further, the gestures are defined in terms of the dynamics that specify the motions in terms of a mass-spring model applied to articulation. Gestures are defined in terms of three specifications: 1. the stiffness of the gesture, the target of the movement and the phasing of the gestures with regard to each other. The stiffness roughly corresponds to the speed of the gesture, that is, speed is a reflection of the stiffness of the gesture. The stiffer the gesture, the faster it is being executed. When the target is being approached relatively slowly then there is less stiffness. The amplitude of the gesture is the amount of displacement from the resting position. The phasing of gestures then is the intergestural timing, that is, the timing between gestures, specifying when one gesture begins and when the next gesture ends.

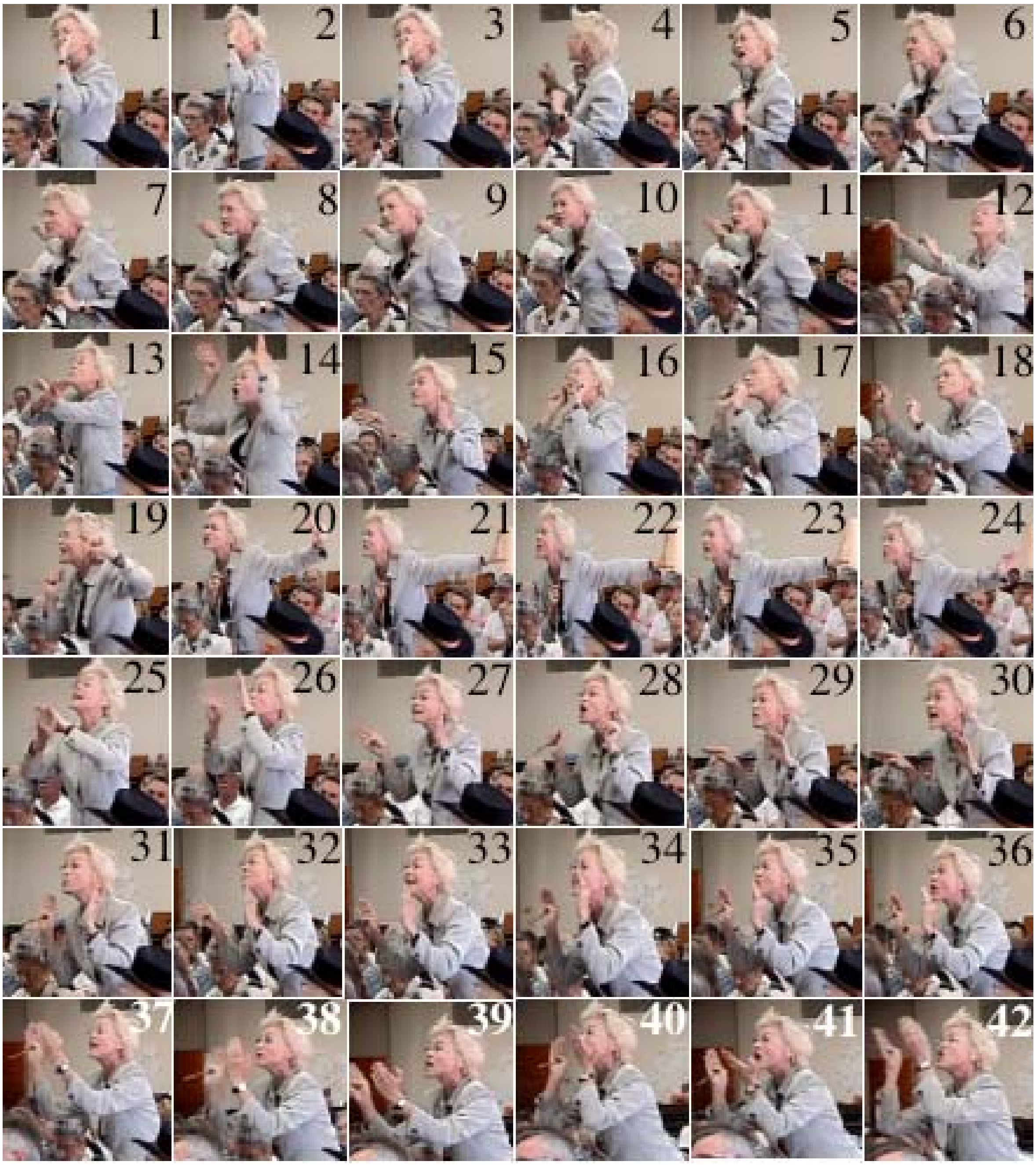
In terms of our specification of the apex, we adapt Loehr's (2004:89) definition. He describes the apex as the "peak of the peak" or as the "kinetic goal of the stroke". In task dynamics terms then this is the target of the gesture. The reason why this is described here in some detail is that we noticed that in more rapid, highly emotional speech, the co-speech gestures are not fully carried out, there often is just a succession of apices, tightly coordinated with pitch accents without the hands or arms returning to a resting position. The gestures overlap and cut each other off, only leaving the peaks of the gestures being observed. (In our example we find that the speaker nods her head for emphasis as well. We have however excluded this from our analysis and annotated hand, arm and body movements only).

4 Analysis

The total length of the sequence we selected for analysis is 2 minutes and 10 seconds. We have provided a full transcription of the whole sequence of Mary-Jane's monologue as Appendix A at the end of this paper. For microanalysis we have selected several segments which most clearly show the various gestural and intonational effects that we aim to illustrate. These segments are subdivisions created by us in our data and are not meant to be compared against each other, but to illustrate sequences that we felt had a unity of topic, sequencing and purpose within the total data set. They are presented in the order in which they occurred, as clusters of frame grabs in panels 1-4. Each panel has been divided into individual frame grabs in order to show the exact gestural sequences. Thus, P2:12-14 means Panel2, frames 12 through 14. We will follow the following format in the presentation of this data: First we will present the transcript along with associated annotations, then the corresponding panels and finally our discussion and any illustrative figures related to them. Let’s begin with the first sequence:

4.1 Panel 1 analysis: The drug war will be over.

M-J	the drug war will be over. **P1:1-3**
(0:20)	we won't have to ship guns and military to Colombia, **P1:4-11**
	it will be over. **P1:12-15**
	Bush won't have to talk to Fox about the drug war, **P1:16-24**
	it will be **P1:25-27**
	over. **P1:28-30**
Audience	[XXclappingXX]
	[there will not be a drug war] **P1:31-36**
	if you legalize drugs. **P1:37-42**



Panel 1: “It will be over”

In P1:5-11, Mary-Jane begins with a deictic gesture with the origin close to her chest. As she utters the phrase, “we won’t have to ship guns and military to Colombia,” she extends her right arm out and points out to the right, metaphorically placing Colombia to one side and away from the origin at ego.

This gesture has a clear parallel with later gestures, most notably the one in P1:16-24, where M-J does a bimanual pointing gesture immediately in front of her and then extends both arms to the left while saying, “Bush won’t have to talk to Fox about the drug war”, producing apices and pitch accents aligned with “Bush” and “Fox.” Interestingly, though both Colombia and current Mexican president Vicente Fox (and by extension, Mexico) are located south of her, she has placed them to the right and left respectively, while Bush, who is north-east of her, has been placed directly in front of her (she herself is facing absolute south). Although we would not expect a speaker of English to exhibit absolute cardinal directionality, people in Tucson often do, since the city is two hours north of the border with Mexico and has dramatic mountain ranges which orient residents to absolute directions. This apparent disparity in the directions of well-known places was our first clue that something beyond simple deictic direction was being represented through Mary-Jane’s gestures. A bird’s eye representation of her gestures would show Bush, guns and military directly facing M-J, while Fox and Mexico would be to the left side and Colombia to the right (Figure 2). :

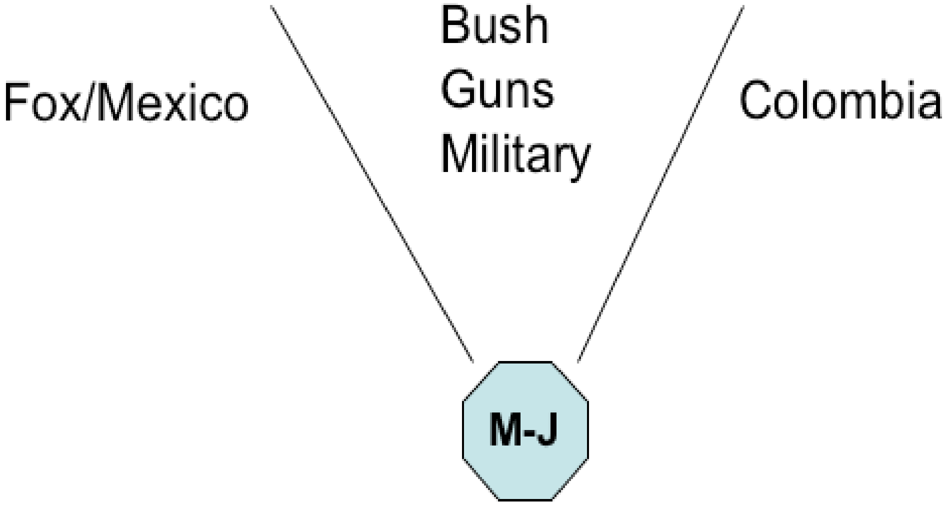


Figure 2: Birds’ eye representation of spatialization of entities by M-J, Panel 1

An interesting catchment of gradually exaggerated gestures also occurs in this stretch. M-J repeats the phrase "it will be over" three times, each time with successively downstepped intonation and more emphatic gesturing. The first time, she brings her left hand (which is in ASL handshape G/X1 (McNeill 1992:88) in one downward stroke from eye level to shoulder level (this is visible in P1:1-4). For the second iteration of the phrase "it will be over," she draws a three-sided, open-bottom box with both hands, starting at the center top (P1:12-15; P1:14 shows both palms facing each other as she sweeps them down to make the sides of the box).

In the final iteration of the "it will be over" phrase (P1:25-30), she draws the complete box and slows down her speech, exhibiting an intonational phrase boundary and a gestural hold in the exact spot where she prepares to bring her hands in to close the box (P1:29), right before the last word, "over." The nested structural parallelisms appear thus:

1) lexical, intonational and syntactic parallelisms with coindexed pronouns at ACE;
2) intonational, gestural, and syntactic parallelisms at BD,
3) and repetition along with gestural parallelism and expansion at CE.

A The drug war will be over
B We won't have to send guns and military to Colombia
C It will be over
D Bush won't have to talk to Fox about the drug war
E It will be over.

Tannen (1989) has argued that repetition in discourse has a cohesion and focussing function, serving to highlight information and structure listener expectations. We believe the communicative constraints that we mentioned above motivate the use of four different semiotic levels of repetition (lexical, syntactic, phonological (intonation), and paralinguistic (gesture)), all within a space of nine seconds of spontaneous speech. The fine-grained coordination of moment-to-moment speech has been discussed in the conversation analytic literature, and to some extent in the literature that is called "emergent grammar," which is concerned with finding the emergent structure of speech as it happens in naturalistic interaction. Our description of the incorporation into emergent structure of affect-laden gesture and intonational phenomena contributes to this literature.

4.2 Panel 2 Analysis: Excuse me, hello?

	L+H* H* L-H%
M-J	because they're going to ex**pe**riment with **drugs**, (P2:1-3)
	L* L* L-H%
	like **ev**ery one has **done**, (P2:4-9)
	L* L-H% H* L-L%
	since **day** **one**. (P2:10-14)
Kolbe	ok,
	well I--[
	H* L-L%
M-J	] we ex**pe**riment. (P2:15-16)
	H* H* H* L-L%
	alcohol and to**ba**cco **kill** people (P2:17-23)

H* H* H* L-L%

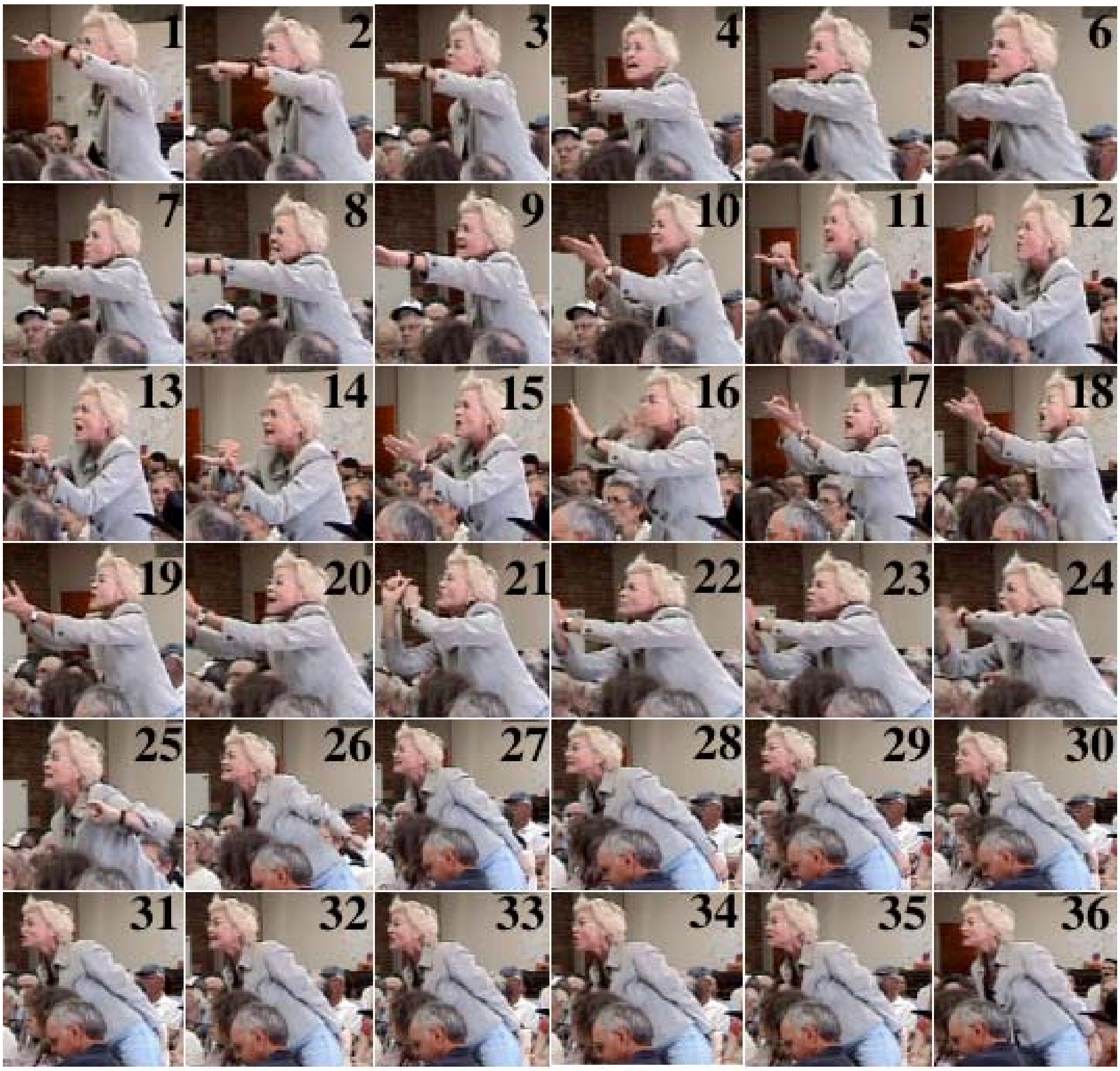
marijuan**a**na doesn't **kill** **a**nybody. (P2:24-28)

L*+H L-H%

ex**cuse** me (P2:29-32)

L+H* !H-L%

hel**lo**: (P2:33-36)



Panel 2: “Excuse me, hello?”

This segment is annotated with the ToBI tones marking pitch accents (*) as well as phrase tones (-) and boundary tones (%). The bolded words indicate that there is a simultaneous occurrence of a pitch accent with a gestural apex. It is particularly noteworthy that Mary-Jane deliberately separates the phrase 'day one' into two intonational phrases. Both words are lent prominence by accenting them and by making pointing gestures where her right index finger lands on the upwards-open palm of her left hand (day: P2:10-11 and one P2:12-14). Note also that Mary-Jane produces the last gestural apex in this segment with her entire upper body as the articulator, which she abruptly stops in mid motion at a precarious 45-degree angle as she says to the Congressman "excuse me, hello".

The intonation contour L*+H L-H% occurring on 'excuse me' has been described as carrying a holistic pragmatic meaning ranging from uncertainty to incredulity (Ward and Hirschberg 1985, Hirschberg & Ward 1992). Hirschberg and Ward argue that "when speakers use the contour to express incredulity, they generally express that incredulity about a value already evoked in the discourse" (p. 243). A striking fact about this use of the uncertainty/incredulity contour is that it is not aligned with the utterance that one might expect it to, given the partially-ordered set relationship (poset) framework described in Ward and Hirschberg (1985). Consider the relevant part of the utterance once more:

H* H* H* L-L%

alcohol and to**ba**cco **kill** people (P2:17-23)

H* H* H* L-L%

mariju**a**na doesn't **kill** **a**nybody. (P2:24-28)

L+H* !H-L%

ex**cuse** me (P2:29-32)

In denying that marijuana kills people, Mary-Jane invokes the possible set of substances that do kill people, and discursively selects for the interlocutor (Kolbe) and overhearers (other constituents) deadly substances that are *legal*. The incredulity being expressed here, signalled by the wide pitch range that the contour carries and which upgrades it from uncertainty (Hirschberg and Ward 1992), is in the assertion that it is the noxious substances that are legal while the innocuous substance is illegal. These discourse facts all line up with the licensed uses of this contour according to Ward and Hirschberg (1985), with one exception: the contour doesn't fall on the expected phrase (marijuana…), but instead on the *following* one (excuse me...). We believe that this finding has two interpretations with important implications:

1) It is possible that this is the point where a speaker under strong communicative demands finally reaches the limits of the processing capacity for online information alignment. In this short segment, Mary-Jane simultaneously aligns pitch accents with syllables contained in informationally prominent words, gestural apices with those pitch accents, while making a complex argument and taking parts of her body to the limits of their physical space. Could the canonical alignment of the incredulity contour have been given up in order to meet the extra processing demands of this task?

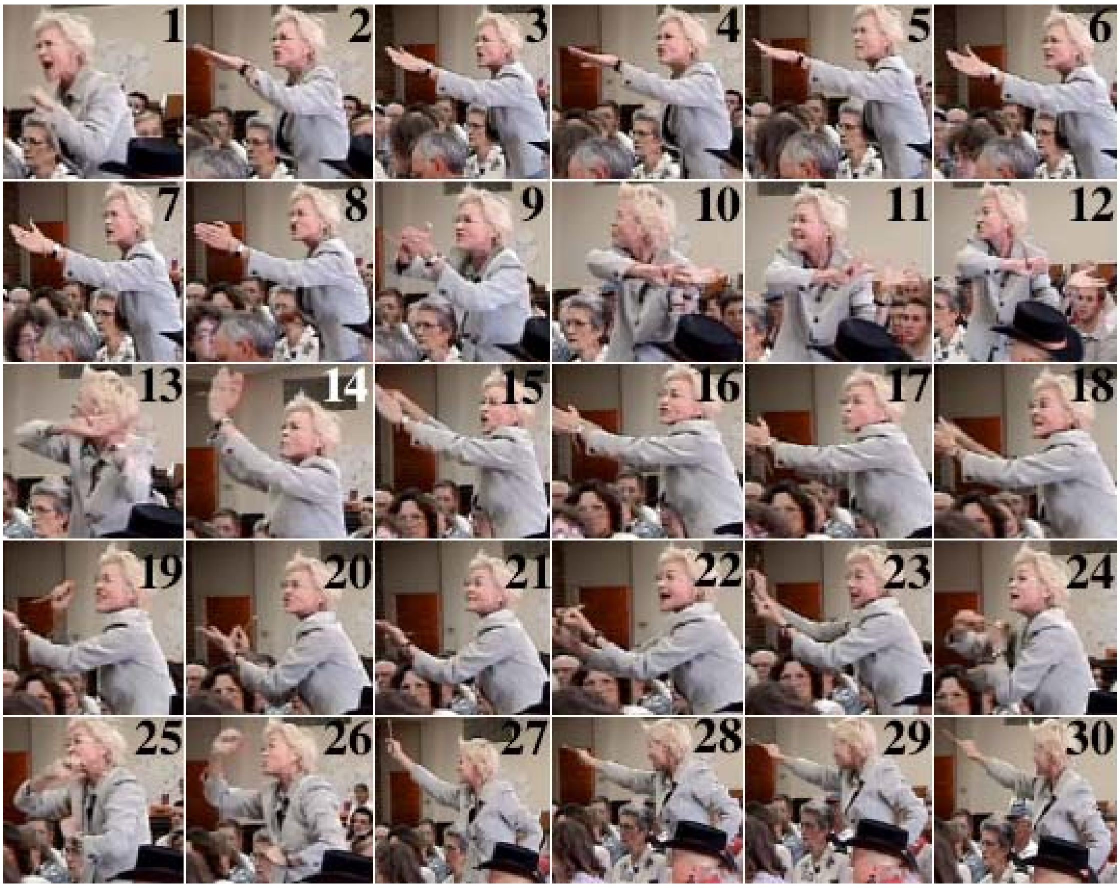
2) A second interpretation is that the speaker, knowing that she was going to use a conventionalized expression with its own L*+H L-H% contour ("excuse me, hello?" often is uttered this way colloquially in the United States), chose to suppress the first contour so as to avoid two identical contours together, following a kind of prosodic OCP, or presumably

> because the meaning is already accessible from a single utterance of the contour.

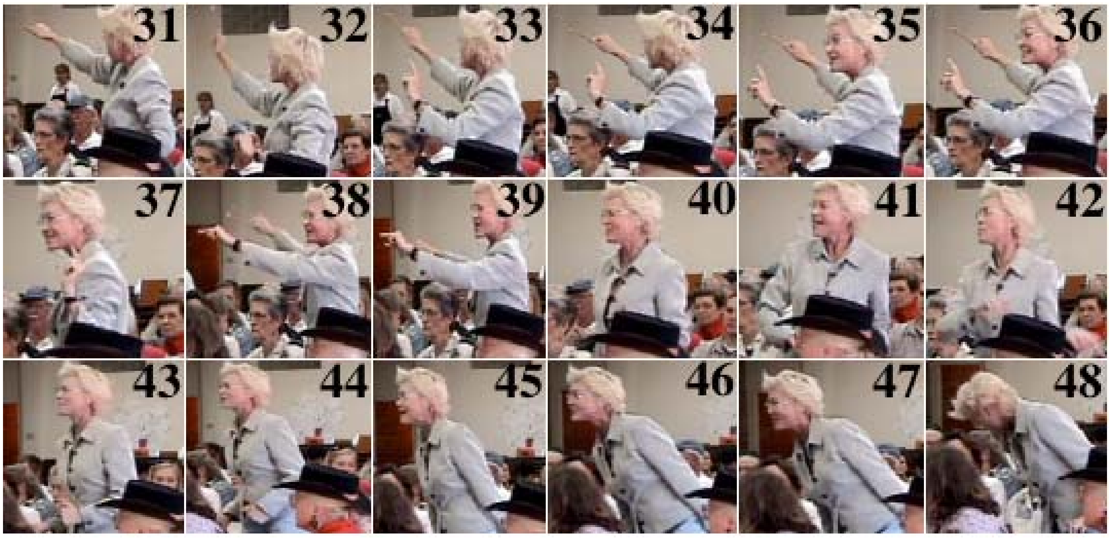
On “hello”, Mary-Jane produces a pitch contour that we described with the tonal sequence L+H* !H-L%. This contour has also been described as a ‘calling contour’ (Beckman & Ayers, 1994). It is the contour often used when shouting a name during the process of looking for somebody (for example to call somebody for dinner). It is our impression that Mary-Jane uses this contour to call Congressman Kolbe metaphorically to do his job, to reproach him.

4.3 Panel 3 Analysis: Alcohol and Tobacco

M-J	]why aren't you talking about alcohol and cigarettes? (P3.1:1-7)
	what you've done is you've taken that industry, (P3.1:8-12)
	and you are putting it on third world. (P3.1:13-17)
Kolbe	what--
M-J	they are consuming the tobacco now. (P3.1:18-23)
	that's where all the tobacco money is coming from. (P3.1:24-30)
Kolbe	[we're spending]
M-J	[they're still making] a profit on cigarettes (P3.2:31-37)
	cause they're selling them to all the [people] (P3.2:38-40)
	that we're dumping our chemicals on (P3.2:41-48)



Panel 3.1: “Alcohol and Tobacco.”



Panel 3.2: “Alcohol and Tobacco.”

The segment represented by panels 3.1 and 3.2 returns to the earlier metaphorical spatialization described in Figure 2 despite the intervening segment in panel 2. Also before panel 3, Mary-Jane has been referring to “kids you’ve been talking about today” who will go to prison because of their alcohol and drug use. Interestingly, the gesture in “kids” starts out directly in front of Mary-Jane, metaphorically in a space shared between the government and the people, and after drug use the “kids” get moved off to “prison,” which also occupies the peripheral position earlier used for external entities (other countries) outside the relationship between the government and the people. Another movement from “us” to “them” takes place in Panels 3.1 and 3.2: as Mary-Jane discusses the practice of chemical dumping by the United States, she makes a gesture that starts out with both hands right behind her left shoulder (as though she were holding a ball and beginning to toss it, P3.1:10), and winds up gesturally in what is by now in her gesture system a stable place for the “third world,” to the far right in her spatial system P3.1:18.

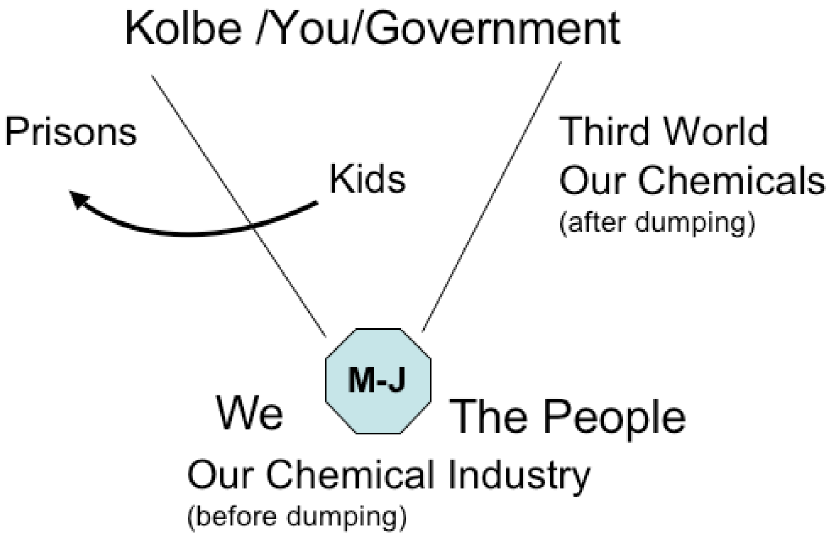


Figure 3: Bird’s eye view of spatialized entities represented by M-J, Panels 3.1-3.2

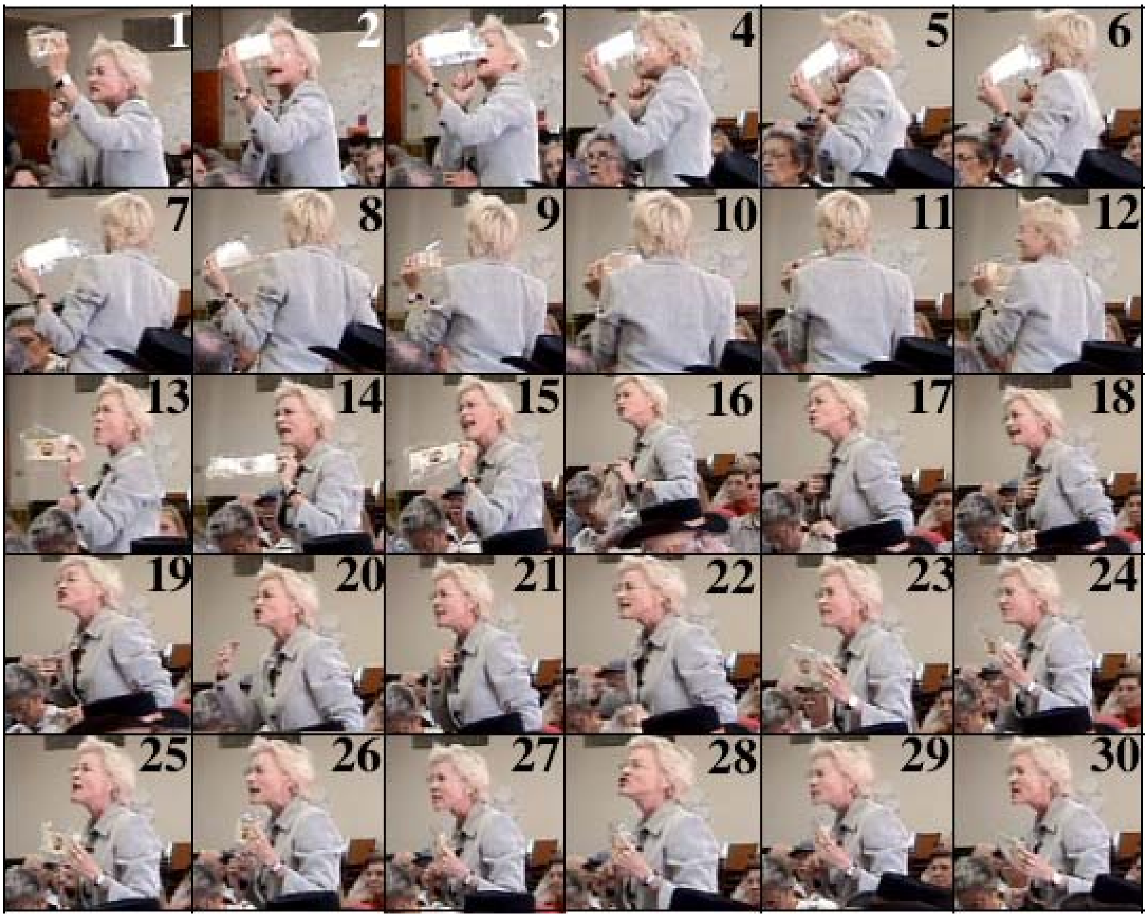
Although a detailed discourse analytic account of the shifting pronominal referents in this data is beyond the aims and scope of this paper, we will note that the pronoun “we” variously refers to a) the people of the United States as a collective that produces chemicals that need to be dumped someplace (“people that we’re dumping our chemicals on”); b) the United States as an entity that ships guns and military overseas (“we won’t have to ship guns and military”); c) constituents that engage in collective behavior and do not share in the government’s definition of what is legal and illegal (“we experiment”); and d) individual citizens who have rights to control their own bodies (“yes we do.”) The pronoun “you” likewise shifts in reference to the addressee in general, to representative Kolbe, to current president G.W. Bush, and to the government in general. This variability in the instantiation of pronominal reference has been observed in political discourse cross-linguistically, as political figures (Van Dijk 2003, Wodak 1989) act on the political stage. In this case tracking the shifts in pronominal reference helps us to anchor and interpret the gestural data, and to make explicit the implicit political model that Mary-Jane sketches with her gestures.

4.4 Analysis Panel 4: My rights as a citizen

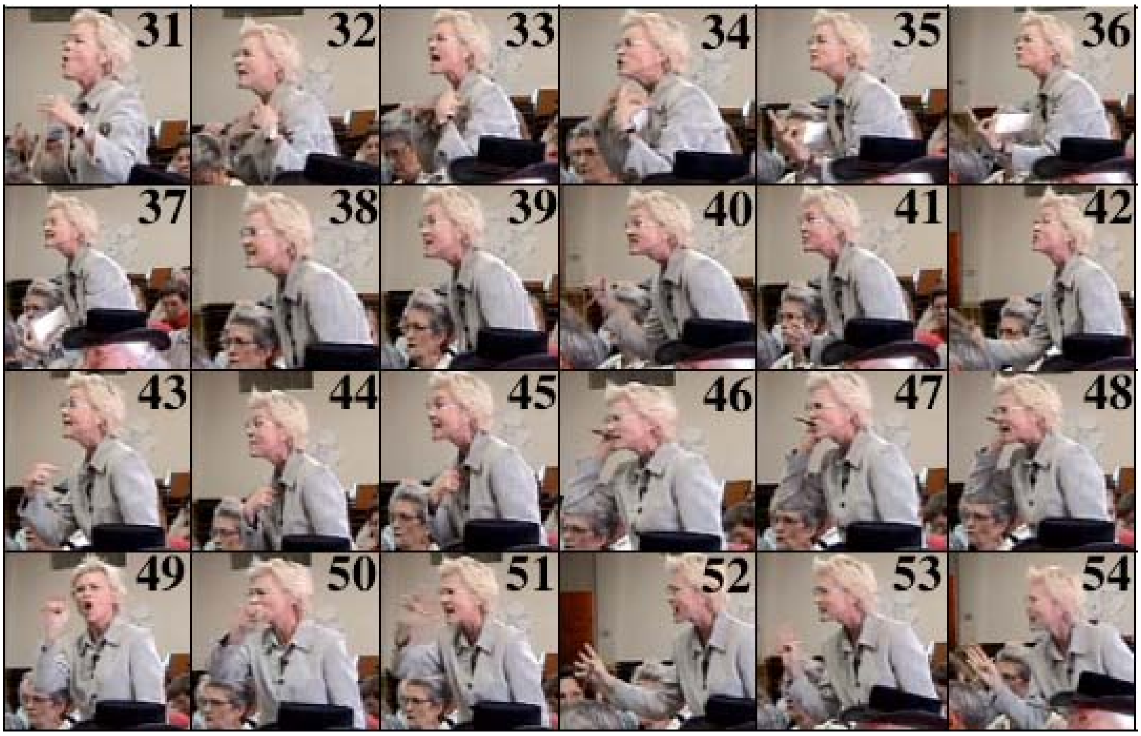
M-J they want to make (P4.1:1)
(.)
hemp seed (P4.1:2-4)
a schedule one narcotic (P4.1:5-11)
(.)
when did I lose the right,(P1:12-17)
(.)

to eat, (P1:18-22)
(.)
the food, (P1:23-24)
(.)
I want, (P1:25-27)
(.)
to eat. (P1:28-30)
when did I lose my rights
as a citizen of this country, (P2:31-41)
to put into my body, (P2:42-45)
and listen to whatever, (P2:46-48)
or watch whatever I want. (P2:49-54)

In the sequence "to eat, the food, I want, to eat" is composed of 4 separate intonational phrases. Each of these phrases contains one intermediate phrase which in turn contains one pitch accent. There are accents on "eat", "food", "want" and "eat". We have observed that perceptually, these accents appear to occur in a downstepping relationship to each other (Liberman & Pierrehumbert, 1984). According to the tonal labeling conventions (ToBI) though, this would be infelicitous. Unfortunately, measurements of the fundamental frequency contour during the time point of the F0 maximum did not generate any insights because the signal was too perturbed. While the quality of the recording allows us to understand the message and hearing the pitch, the pitch tracking algorithms in different analyses software packages (Wavesurfer, Praat) logged values that were unreliable. We would like to point out though that we believe that there is a relationship between the pitch accents in this sequence and that the trigger of downstep can transcend the limits of an intermediate phrase.



Panel 4.1: "My rights as a citizen"



Panel 4.2: "My rights as a citizen."

The final panels for our gestural/intonational microanalysis are panels 4.1 and 4.2. In this segment Mary-Jane has taken advantage of a prop that she brought to the Town Hall meeting to dramatize her point: a hemp-seed chocolate chip cookie in a ziploc bag. She holds up the cookie and addresses the audience, showing them the offending item which is made with hemp seed (the seed of the marijuana plant). She attempts to highlight the absurdity of the government's decree that marijuana is illegal by claiming that she does not have the right to eat a chocolate chip cookie. She ends her performance (one of the issues for us is that we cannot be sure that this is not a rehearsed, prepared speech) with a dramatic flourish, by uttering two parallel constructions, with the same syntactic structure, in a slow and dramatically delivered style. They are presented below, in column format to highlight the parallelism. In conversation analytic transcription conventions, the periods in parenthesis mean that there were significant pauses in the speech stream.

1	**2**
when did I lose the right	when did I lose my rights
(.)	(.)
to eat	as a citizen of this country
(.)	(.)
the food	to put into my body
(.)	(.)
I want	and listen to whatever
(.)	(.)
to eat.	or watch whatever I want.

In this segment, not only is the syntactic/constructional parallelism evident above, but we further note that there occur to sequences of pitch accents in these two utterances with downstep quality, despite the fact that in standard Tones and Breaks Indices theory it is thought to be the case that intonational downstep cannot happen across intonational phrase boundaries (Beckman & Ayers 1994). A downstepped H* accent is a type of high star accent, except that the tonal realization is being influenced by the accent preceding the high tone. Often a L+H* accent (a bitonal pitch accent) triggers downstep, so it is thought not to be able to occur across an intermediate phrase, because the trigger is displaced to the preceding intonational phrase. In this case, we would like to claim that for reasons of parallelism and emphasis, downstepped H* (notation: !H*) accents were used in places that have not been previously documented. These !H* accents coincide in our data with the apices of the gestures as well, since Mary-Jane rocks her entire body back and forth approximately 30 degrees to land at the apex of these movements on the !H* accents in both of these utterances.

This brings us to a more general discussion of the correlation between pitch accents and apices.

4.5 Pitch accents & Apices

In order to establish some kind of a measure of what the relation is between the occurrence of an accent and the occurrence of an apex, we counted the number of times Mary-Jane produced accents, that is, prosodic prominences, as well as the number of times where her gesturing included an apex, the peak of the gesture. The numbers are shown in the table below.

Segment	Dur in Sec	Apices	PA	Co-occurring Apices & PA	PA but no Apices	Apices but no PA
1. over	14.23	16 (100%)		16 (100%)		0 (100%)
			24 (100%)	16 (66.7%)	8 (33.3%)	
2. excuse me, hello	11.72	11 (100%)		11 (100%)		0 (100%)
			15 (100%)	11 (73.3%)	4 (26.7%)	
3. tobacco	14.26	16 (100%)		15 (93.8%)		1 (6.3%)
			27 (100%)	15 (55.6%)	12 (44.4%)	
4. lose the right	19.15	28 (100%)		26 (92.8%)		2 (7.1%)
			32 (100%)	26 (81.3%)	6 (18.8%)	
Total	59.36	71 (100%)		68 (95.7%)		3 (4.2%)
			98 (100%)	68 (69.4%)	30 (30.6%)	

Table 1: "(Co-)occurrence of Pitch Accents and Gestural Apices"

The general tendency apparent from this table comparing columns 'Apices' and 'PA' is that there are always more pitch accents than apices in any given segment. Note that we are not comparing the segments with each other, we have merely stated the numbers separately as to make the counting more transparent and easier. As we can see from column 'PA but no Apices', in verbally rendered speech, information is highlighted by acoustic prominences that obviously must not have visually co-occurring gestural apices. On the other hand, it only rarely happens that there is an apex occurring without a pitch accent (see column 'Apices but no PA'). This suggests to us either that some prosodic prominences are just rhythmic in nature or that not all prominences are semantically 'worthy' of being marked by co-speech gestures. It is possible also that the information given on these two different planes is complementary in nature, highlighting *different parts* of the message.

In column, 'Co-occurring Apices & PA', the numbers in the unshaded box (e.g. 16 (100%)) indicates that whenever there is a gestural apex, there also is a pitch accent. The number in the shaded box indicates that only 66.7% of all pitch accents were accompanied by an apex, a beat gesture. In terms of the total numbers, we find that 95.7% of all apices were accompanied by a pitch accent whereas only 69.4% of all pitch accents were additionally marked by a gestural apex.

We have looked for the co-occurrence of accent location and gestural apex. However, it may be promising also to differentiate between the accent types, which we have not done so far. It is possible that certain (less prominent) accent types such as a downstepped !H* or a L* may be reinforced by gestural means. However we can make no claims regarding this hypothesis.

5 Discussion

We found that speech and gesturing are two different channels/modes of information transfer which allow for different content to be transmitted. If we assume the validity of Bolinger's (1986) claim that "gesture and speech stem from the same semantic intent […]" then we commit ourselves to the notion that some degree of preplanning is involved in generating not only speech output but also gestural output in order to convey information on different planes. How information is structured and divided up across the two channels is not understood at this point. From our data it appears that complimentary and contextual information is transmitted via gestures while concrete assertions are made explicit via speech. We also do not know what constraints exist on (pre-) planning complex gestures that we know are time aligned with linguistic structure in the final output.

Gestures are not just involuntary movements but finely coordinated structures of motion, aligned with semantic content. Since speech can be understood over the telephone or in the dark (with a complete absence of visual cues), we must assume that gestures facilitate the information transfer from the speaker to the listener/viewer but are not necessary for successful transmittal of content. Gestures play an important role for the naturalness of speech and for cuing speaker stance. On the other hand, it is also known that speakers gesture while speaking when nobody is there to see them. Therefore, it appears that gesturing is not just a facilitation device for the listener/viewer but a mode of self expression for the speaker.

Gesturing exhibits cross-cultural differences, and it is, just as other communicative actions, learned behavior. The mechanisms of acquiring co-speech gesturing consists of coordinating the individual tasks of the complex gestures with each other (for example, lift right arm, rotate palm upward, release arm in this constellation) and then coordinating these motions with speech so that points of informational prominence in speech are accompanied for example by apices in gesture.

An infant as young as a few hours displays the ability to mimic the sticking out of one's tongue when prompted (Meltzoff & Moore 1977, p. 78). This suggests that our cognitive systems provides for learning by example and imitation. The task is a formidable one, as inverse mapping has to take place: the infant observes tongue movement via the visual channel, and then has to map the observed movements to own motor patterns of the articulators (opening lips, lower jaw, extending tongue) in order to perform the task of sticking out the tongue.

It appears that the same type of mechanisms should be involved to learn other motor skills such as finely coordinating hand, arm and body movements to be timed with speech: based on our casual observation, often the offset of a complex gesture co-occurs with the end of a prosodic phrase (intermediate or intonational phrase) or as we have shown in this paper, apices, the peaks of the gestures, co-occur with pitch accents. An interesting test case would be an investigation of the type and timing of co-speech gesturing of people who have been blind since birth.

If the interpretation of intonation contours, that is, if the interpretation of prosodic focus is partly determined by context, we have to take into account that this context is not just provided verbally or relates back to knowledge the interlocutors possess already. Rather, in face-to-face communication the gestural channel is able to provide the speaker’s stance, or part of the context in which to interpret the utterance. It appears though that the verbal and gestural channels are interpreted simultaneously and holistically so that the semantic content can be recalled but the presentation and structure of information is more fleeting in nature and thus cannot easily be teased out by the receiver. That is, we tend to remember meaning rather than form but also make inferences which let listeners arrive at an interpretation (Bransford, Barclay, and Franks 1972).

6 Conclusion

In this work we have carried out a case study of the coordination of spontaneous speech with gesture, focusing on intonational alignment with pitch peaks (after Loehr 2004), and have found that wherever there is a gestural apex in our data, there is also a pitch accent. The reverse, however, is not true, because pitch accents often occur without gestures or the apices are phase-shifted from the gestural peaks. Our results concur with those of Loehr (2004) for laboratory data, and indicate that gestural phenomena are in robust co-occurrence with pitch accents in both laboratory and spontaneous speech. Our findings also include the discovery of a downstep relationship across intonational phrase boundaries, as well as the pervasive use of lexical, syntactic, gestural, and intonational parallelism in the performance of a speaker under high pressure in an affect-laden, spontaneous communicative situation. While the intonational and gestural alignments were observed within and across intonational phrases, metaphorical gestural spatialization had a larger domain, it took longer to unfold, and sketched out complementary information on this speaker's notion of the relationship between a political representative, the government, and outside entities.

7 Appendix: Transcription of Town Hall Meeting

Town Hall meeting at St. Cyril's, midtown Tucson school, February 2001. Filmed with Sony DVTR8 mini-DV camera, sound aligned with audience microphone, quicktime clip name: kolbestcyril2(2/18,DVTR8).mov

Kolbe and since you wanna comment on [this
[I'm sorry]
We're gonna] get your comment[
M-J]my--
my only problem with--
(0:04) n- n- n- not legalizing all drugs?
Green right.
M-J you take the criminal element out of it,
when you end prohibition,
(.)
this is what we saw in the thirties.
the drug war will be over.
(0:20) we won't have to ship guns and military to Colombia,
it will be over.
Bush won't have to talk to Fox about the drug war,
it will be
(.)
over.

Aud [xxclappingxx]
[there will not BE a drug war]
if you legalize drugs.
Kolbe ok [[calm down,
(0:30) calm down,
(.)
calm down,]]
M-J [[you are making it a criminal]] enterprise,
Aud: boooo]
[you have got-]
Aud: [boooo]
the government has made it a criminal enterprise,
the government is making money,
you are making money hand over fist,
(0:40) you are building prisons so fast it's disgusting,
you're not putting any of that money into education,
you're locking those-
you're building prisons to lock those kids up that you're talking about today,
they're going to go to prison.
(0:50) because they're going to experiment with drugs,
like every one has done
Since day one.
Kolbe ok,
well I--[
M-J]we experiment.
alcohol and tobacco kill people.
marijuana doesn't kill anybody.

excuse me,
hello-o:
(.)
Kolbe ok.
(.)
well--[
M-J]why aren't you talking about alcohol and cigarettes?
what you've done is you've taken that industry,
and you are putting it on third world.
Kolbe what--
M-J they are consuming the tobacco now.
that's where all the tobacco money is coming from.
Kolbe [we're spending]
M-J [they're still making] a profit on cigarettes cause they're selling them to all the [people]
that we're dumping our chemicals on
Kolbe [we're--]
we're spending a lot of money on a--
(.)
on education,
which is what I think—
(.)
on tobacco,
which is what I think we need to be doing,
(.)
[spending a lot on that].
M-J [proper name] wants to make this
a schedule one narcotic.

A chocolate chip cookie,
ok?
a chocolate chip cookie,
this is what they tried to run through on October the thirtieth,
they want to make
(.)
hemp seed
a schedule one narcotic.
when did I lose the right,
(.)
to eat,
(.)
the food,
(.)
I want,
(.)
to eat.
(1:40) when did I lose my rights
(.)
as a citizen of this country
(.)
to put in my body
(.)
and listen to whatever
(.)
or watch whatever I want.
Kolbe well,
(.)

uh,

there--

there always have been limits on doing--

on some things.

you do—

we do--

you do not have[

M-J]I limit myself,

[you don't limit me thank you].

Kolbe [y:--

y:--

you] don't have absolute rights to do everything.

M-J yes we do.

Kolbe and we do--

(.)

we never have had.

References

Beattie, G., & Shovelton, H. (1999) An Experimental Investigation of the Role of Iconic Gestures in Lexical Access Using the Tip of the Tongue Phenomenon. *British Journal of Social Psychology*, 90, pp. 35-56.

Beckman, M. E. & Ayers, G. E. (1994) *Guidelines to ToBI Labelling (version 2.0)*. The Ohio State University.

Berthoz, A. (1997) Le Sens du Mouvement. Paris: Editions Odile Jacob.

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972) Sentence Memory: a Constructive Versus Interpretive Approach. *Cognitive Psychology* 3, pp. 193-209.

Bohman, J. (1996) *Public Deliberation.* Cambridge, MA: MIT Press.

Bollinger, D. D. (1986) Intonation and Gesture. *American Speech*, 58.2, pp. 156-174.

Brown, P., & Levinson, S. C. (1993) “Uphill” and “Downhill” in Tzeltal. *Journal of Linguistic Anthropology* 3(1), pp. 46-74.

Browman, C. P., & Goldstein, L. (1990) Tiers in articulatory phonology, with some implications for casual speech. In T. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pp. 341-376. Cambridge University Press.

Browman, C. P., & Goldstein, L. (1992) Articulatory Phonology: An overview. *Phonetica,* 49, pp. 155-180.

Cassell, J., McNeill, D., & McCollough, K. E. (1999) Speech-gesture mismatches: Evidence for One Underlying Representation of Linguistic and Nonlinguistic Information. *Pragmatics and Cognition*, 7(1), pp. 1-33.

Cassell, J. (2000) A framework for Gesture Generation and Interpretation. In R. Cipolla and A. Pentland (eds.) Computer Vision in Human-Machine Interaction. Cambridge: Cambridge University Press. http://www.soc.northwestern.edu/justine/publications/gesture_wkshop.pdf

Chomsky, N. (2000) Minimalist Inquiries: The Framework. In *Step by Step*, R. Martin, D. Michaels, and J. Uriagereka (eds.), pp. 89-155. Cambridge, MA: MIT Press.

Cosnier, J. (1982) Communications et langages gestuels. In J. Cosnier, A. Berrendonner, J. Coulon, and C. Orecchioni (eds.) *Les voies du langage: Communications verbales, gestuelles, et animales*, pp. 255-304. Paris: Dunod.

Enfield, N. (2003) Producing and Editing Diagrams Using Co-Speech Gesture: Spatializing Non-spatial Relations in Explanations of Kinship in Laos. *Journal of Linguistic Anthropology* 13, pp. 7-50.

Ekman, P. & W. Friesen (1972) Hand Movements. *The Journal of Communication* 22, pp. 353-374.

Farnell, B. (1995) Human Action Signs in Cultural Context: the Visible and the Invisible in Movement and Dance. Metuchen, NJ: Scarecrow Press

Fadiga, L. & Craighero, L. (2003) New Insights on Sensorimotor Integration: From hand action to speech perception. *Brain and Cognition* 53, pp. 514-534.

Fodor, J. (1983) *The Modularity of Mind.* Cambridge, MA: MIT Press.

Halle, M. & Stevens, K. (1959) Analysis by Synthesis. In W. Wathen-Dunn and L.E. Woods (eds.) *Proceedings of the Seminar on Speech Compression and Processing.* AFCRC-TR-59-198, December 1959, Vol. II, paper D7.

Harper, L., Loehr, D. & Bigbee, A. (2000) Gesture is not just Pointing. Presented at and published by the International Conference on Natural Language Generation.

Haviland, J. B. (1993) Anchoring, Iconicity, and Orientation in Guugu Yimidhirr Pointing Gestures. *Journal of Linguistic Anthropology*, Vol. III(1), pp. 3-45.

Hirschberg, J. & Ward, G. (1992) The Influence of Pitch Range, Duration, Amplitude and Spectral Features on the Interpretation of the rise-fall-rise Intonation Contour in English. *Journal of Phonetics*, 20, pp. 241-251.

Goodwin, C. (1981) *Conversational Organization. Interaction between Speakers and Hearers.* New York: Academic Press.

Kelly, S., Barr, D., Church, R., & Lynch, K. (1999) Offering a Hand to Pragmatic Understanding: The Role of Speech and Gesture in Comprehension and Memory. *Journal of Memory and Language,* 40, pp. 577-592.

Kendon, A. (1972) Some Relationships between Body Motion and Speech. In Siegman, A. & Pope, B., (eds.) *Studies in Dyadic Communication*, pp. 177-210. New York: Pergamon Press.

Kendon, A. (1980) Features of the Structural Analysis of Human Communicational Behaviour. In von Raffler-Engel, W. (ed.), *Apsects of Nonverbal Communication*. Swets & Zeitlinger, Lisse, Netherlands, pp. 29.43.

Kendon, A. (1996) Gesture in Language Acquisition. *Multilingua* 15, pp. 201-214

Krifka, M. (2005) A Functional Similarity between Bimanual Coordination and Topic/Comment Structure. Paper presented at the Blankensee Colloquium, Berlin Schmöckwitz: Language Evolution: Cognitive and Cultural Factors, July 14-16, 2005.

Liberman, P. & Blumstein, S. (1990) Speech Physiology, Speech Perception, and Acoustic Phonetics. *Cambridge Studies in Speech Science and Communication*. Cambridge: Cambridge University Press.

Liberman, A. & Mattingly, I. (1985) The Motor Theory Revised. *Cognition* 21, pp. 1-36.

Liberman, M. & Pierrehumbert, J. (1984) Intonational Invariants under Changes in Pitch Range and Length. In M. Aronoff & R. Oehrle (eds.) Language Sound Structure. MIT Press, Cambridge, MA.

Loehr, D. (2004) *Gesture and Intonation.* Doctoral Dissertation, Georgetown University, Washington, DC.

Marantz, A. (to appear) Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review.* Accessible from author's website: http://web.mit.edu/linguistics/www/marantz/index.html

McNeill, D. (1992) *Hand and mind.* The University of Chicago Press.

McNeill, D. (2000) *Language and gesture.* Cambridge: Cambridge University Press.

McNeill, D., Quek, F., McCullough, K.-E., Duncan, S., Furuyama, N., Bryll, R., Ma, X.-F. & Ansari, R. (2001) Catchments, Prosody and Discourse. *Gesture* 1:1, pp. 9-33.

Meltzoff, A. N., & Moore, M. K. (1977) Imitation of Facial and Manual Gestures by Human Neonates. Science, 198, pp. 75-78.

Pierrehumbert, J. (1980) *The Phonology and Phonetics of English Intonation.* Doctoral Dissertation, MIT.

Pierrehumbert, J. & Hirschberg, J. (1990) "The Meaning of Intonation Contours in the Interpretation of Discourse." In Cohen, Morgan, and Pollack, (eds.) *Plans and Intentions in Communication and Discourse.* MIT Press, pp. 271-311.

Rimé, B. (1982) The Elimination of Visible Behavior from Social Interactions: Effects on Verbal, Nonverbal, and Interpersonal Variables. *European Journal of Social Psychology* 12, pp. 113-29.

Rizzolatti G & M. Gentilucci (1988) Motor and Visual-Motor Functions of the Premotor Cortex. In Rakic, P. and W. Singer, eds. *Neurobiology of the Neo Cortex*. Dahlem Workshop Reports. Chichester: John Wiley.

Rizzolatti, G., Fadiga, L., Fogassi, L., Gallese, V. (1999) Resonance Behaviors and Mirror Neurons. *Arch Ital Biol.* May 137(2-3), pp. 85-100.

Tannen, D. (1989) Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse. Cambridge: Cambridge University Press.

Tatham, M. (1996) Articulatory Phonology, Task Dynamics and Computational Adequacy. *Proceedings. Institute of Acoustics.* 18.

Vallduví, E. & E. Engdahl (1996) The Linguistic Realization of Information Packaging. *Linguistics,* 34, 3(343), pp. 459-519.

Van Dijk, T. A. (2003) Text and Context of Parliamentary Debates. In Paul Bayley (ed.), Cross-Cultural Perspectives on Parliamentary Discourse, pp. 339-372. Amsterdam: Benjamins.

Ward, G. & J. Hirschberg (1985) Implicating Uncertainty: the Pragmatics of Fall-Rise Intonation. *Language,* 61, pp. 747-776.

Wodak, R. (ed.) (1989) Language, Power, and Ideology: Studies in Political Discourse. Amsterdam Philadelphia: J. Benjamins Co.

Stefanie Jannedy
Humboldt Universität zu Berlin
SFB 632 „Informationsstruktur“
Mohrenstr.
14415 Berlin
Germany
jannedy@ling.ohio-state.edu