Towards a Balanced Corpus of Multimodal Referring Expressions in Dialogue

Ielka van der Sluis [1] and **Paul Piwek** [2] and **Albert Gatt** [3] and **Adrian Bangerter** [4]

Abstract. This paper describes an experiment in which dialogues are elicited through an identification task. Currently we are transcribing the collected data. The primary purpose of the experiment is to test a number of hypotheses regarding both the production and perception of multimodal referring expressions. To achieve this, the experiment was designed such that a number of factors (prior reference, focus of attention, visual attributes and cardinality) were systematically manipulated. We anticipate that the results of the experiment will yield information that can inform the construction of algorithms for the automatic generation of natural and easy-to-understand referring expressions. Moreover, the balanced corpus of multimodal referring expressions that was collected will hopefully become a resource for answering further, as yet unanticipated, questions on the nature of multimodal referring expressions.

1 Introduction

One of the fundamental tasks of Natural Language Generation (NLG) systems is the Generation of Referring Expressions (GRE). Over the past couple of decades, this has been the subject of intensive research [2, 12, 11], and is typically defined as an *identification* problem: given a domain representing entities and their properties, construct a referring expression for a target referent which singles it out from its distractors. While several recent proposals have generalised this problem definition, to deal for example with relations [10, 17], plural referents [26, 13, 14], and vague predicates [27], there has been comparatively little work on the generation of *multimodal* referring acts (but see [18, 23, 28]). Moreover, the majority of contributions have focused on monologue, with interaction between user and NLG system assumed to be absent or limited. Meanwhile, psycholinguistic work is increasingly focusing attention on the conditions governing the use of pointing gestures as part of referring acts in dialogue. Of particular relevance to the questions addressed in this paper is the interaction between the two modalities of *pointing* and *describing* [6, 4, 8, 23, 24].

This paper describes the design of an ongoing experiment on multimodal reference in two-party dialogue. Our aim is to harness the empirical evidence for the design of multimodal GRE algorithms, by studying the corpus of interactions collected in the experiment. The resulting corpus is balanced, in the sense put forward by [15], because the conditions under which references were elicited correspond to experimental variables that are counter-balanced. Moreover, the focus on dialogue permits the investigation to take both a speaker/generator's and a hearer/reader's point of view, with potentially useful data on such factors as alignment and entrainment [7], and the nature of collaboration or negotiation that is a feature of interactive referential communication [9], currently a hotly debated topic in the psycholinguistic literature [22].

Describing vs. pointing Following the influential work in [11], GRE algorithms often take into account the finding that speakers manifest *attribute preferences*, which cause them to overspecify their descriptions. For example, in experiments on reference in visual domains, colour tends to feature in speakers' descriptions irrespective of its discriminatory value, while vague properties like size are relatively dispreferred [21, 5, 3]. On the other hand, recent work on modality choice in reference suggests a potential trade-off between the use of pointing and the amount of information given in a description [28], though the use of pointing also depends on the potential ambiguity of a reference [8] and whether a change of focus is taking place [23]. Our experiment seeks to further this research in four principal directions. First, we look at modality choice as a function of the properties which are available to verbally describe a referent. Thus, if attribute preferences play a role, the possibility of describing a referent using properties like colour may reduce the likelihood of a pointing gesture. Second, we also manipulate the extent to which a referent is in (discourse) focus, that is, whether it was recently mentioned in the dialogue or not. Typically, verbal references to previously mentioned entities tend to be reduced. Does this affect the likelihood of pointing? Third, we look at both singular and plural references, the latter being references to groups of 5 entities. This may increase the visual salience of a referent, which in turn may interact with the other two factors. Finally, we examine to what extent a change of the domain focus (i.e., when the current target is distant from the previous target) affects use of pointing gestures.

Data on these questions will inform the design of multimodal GRE algorithms whose output is 1) *natural*, that is, corresponds closely to what human speakers do in comparable situations, and 2) *easy-to-understand*, i.e., allows the addressee to quickly identify the intended referent without the need for prolonged clarificatory exchanges.

2 The Experiment

2.1 Task and Setup

Figure 1 presents a bird's eye view of the experimental setup in which a director and a follower are talking about a map that is situated on the wall in front of them, henceforth the *shared map*. Both can interact freely using speech and gesture, without touching the shared map or standing up. Each also has a private copy of the map; the

[1] Computing Science, University of Aberdeen, UK
[2] Centre for Research in Computing, The Open University, UK
[3] Computing Science, University of Aberdeen, UK
[4] Institut de Psychologie du Travail et des Organisations, University of Neuchâtel, Switzerland.

director's copy has an itinerary on it, and her task is to communicate the itinerary to the follower. The follower needs to reproduce the itinerary on his private copy. The rules of the experiment were as follows:

- Since this is a conversation, the follower is free to interrupt the director and ask for any clarification s/he thinks is necessary.
- Both participants are free to indicate landmarks or parts of the map to their partner in any way they like.
- Both participants are not permitted to show their partner their private map at any point. They can only discuss the shared map.
- Both participants must remain seated throughout the experiment.

While this task resembles the MapTask experiments ([1]), the latter manipulated mismatches between features on the director and follower map, phonological properties of feature labels on maps, familiarity of participants with each other and eye contact between participants[5]. The current experiment systematically manipulates target size, colour, cardinality, prior reference and domain focus, in a balanced design. Though this arguably leads to a certain degree of artificiality in the conversational setting, the balance would not be easy to obtain in an uncontrolled setting or with off-the-shelf materials like real maps. Further properties of our experiment that distinguish it from the MapTask are: (1) objects in the visual domains are not named, so that participants need to produce their own referring expressions, (2) the participants are always able to see each other; (3) the participants are allowed to include pointing gestures in their referring expressions (a MapTask type experiment that does include non-verbal behaviour, in particular, eye gaze, is reported in [20]).

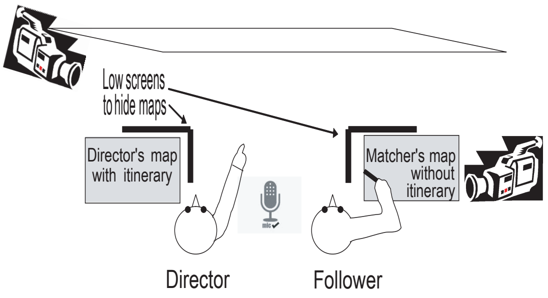

Figure 1. Bird's-eye view of the experiment setup.

2.2 Materials and Independent Variables

Four maps were constructed, consisting of simple geometrical landmarks (ovals or squares). Two of the maps (one each for ovals and squares) have *group* landmarks, whereas the other two have singletons. Objects differ in their size (large, medium, small) and colour (red, blue, green). Each dyad in the experiment discusses all four maps. Per dyad, the participants switch director/follower roles after each map. The order in which dyads discuss maps is counter balanced acrosss dyads. There are four independent variables in this experiment:

- **Cardinality** The target destinations in the itineraries are either singleton sets or sets of 5 objects that have the same attributes (e.g., all green squares)
- **Visual Attributes:** Targets on the itinerary differ from their distractors – the objects in their immediate vicinity (the 'focus area') – in colour, or in size, or in both colour and size. The focus area is defined as the set of objects immediately surrounding a target.
- **Prior reference:** Some of the targets are visited twice in the itinerary.
- **Shift of domain focus:** Targets are located near to or far away from the previous target. If two targets t_1 and t_2 are in the *near* condition, then t_1 is one of the distractors of t_2 and vice versa.

The overall set up of this experiment is illustrated in Table 2.2:

I	near	new	size
II	near	new	colour
III	near	new	both
IV	near	old	size
V	near	old	colour
VI	near	old	both
VII	far	new	size
VIII	far	new	colour
IX	far	new	both
X	far	old	size
XI	far	old	colour
XII	far	old	both

Table 1. Overview of the experimental design. Each different type of target corresponds to a Roman numeral. The types are a function of *Focus* × *Prior* reference × *Attributes*, yielding a total of $2 \times 2 \times 3 = 12$ types of targets that appear on each map. Moreover, half of the maps is populated only with singleton targets, whereas the other half is populated with target sets whose cardinality is 5. Taking this into account, overall we have $2 \times 12 = 24$ types of target in our experiment.

2.3 Current Status and Annotation Plans

Using the maps described in Section 2.2, a pilot of the experiment was carried out in Aberdeen (see Figure 3 for an impression). The pilot led to a few minor adjustments in the setup (e.g., we moved from a projected to a printed shared map), and subsequently data was collected from 24 dyads with the validated setup. Currently, the data is being transcribed. See Figure 2.3 for an example.

128	D	Uh and if you *go straight up* from that you've got five blue ones	D points at the map and moves his finger upwards
129	F	Yeah [*there*?]	D is still pointing F points
130	D	[There] yeah	D is still pointing F is still pointing
131	F	one two three four five	D is still pointing M is still pointing
132	D	Yeah. They're all number three	D is still pointing
133	F	Right. Right.	
134	D	And the five reds just *to the right over*	D points and moves his finger to the right
135	F	And like a kind of *downwards* arrow	D is still pointing F moves his hand upwards
136	D	Arrow yeah they're all number four. Number five. Uh and five is paired with one *with these ones*.	D stops pointing D points
137	F	All right.	

Figure 2. Excerpt from dialogue O17-S33-S34, where *D* = director, *F* = follower and where the brackets indicate overlapping speech and the text in italics indicates approximately the co-duration of gesture and speech

Our next task will be to annotate the data. For this, we will build on existing guidelines and best practice, e.g., use of stand-off XML, for annotation of multimodal data (see [19, 16]). Our main annotation tasks will be: identification of multimodal referring expressions, linking of referring expressions with domain objects (i.e., intended referents) and segmentation of dialogue into episodes spanning the point in time from initiation to successful completion of a target identification.

[5] http://www.hcrc.ed.ac.uk/maptask/maptask-description.html

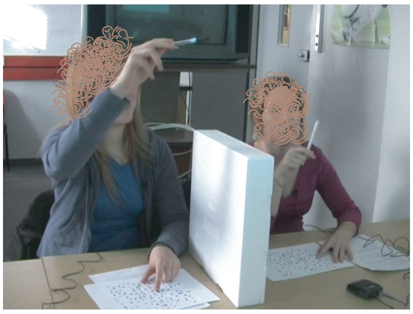

Figure 3. Two participants at work in the pilot of the experiment.

3 Research Questions and Hypotheses

3.1 Production: The Director

The main distinctive feature of the current experiment is that we rigorously controlled for a significant number of features of the referents. The experimental design allows us to both address new questions, and validate existing findings from previous observational studies that were made in more natural and less controlled settings. For example, in an observational study [23] found that some speakers used a lot of gestural information, while others did not at all. The current study will help us to answer the question whether such different styles and strategies are tied to particular features of the communicative situation, or are really a result of individual differences. Other findings include whether speakers use extensive pointing gestures or keep their gestures close to their body depends on the communicative function of the message they want to get across (c.f. [8]). Also, the linguistic information that speakers use varies considerably depending on how difficult it is to describe an object as a function of the number of relevant attributes [25]. In addition, speakers display different approaches in conveying the distinguishing properties of an object to the addressee. For instance, in the world map study discussed in [28] speakers used different strategies to indicate a country. Some used prominent objects on the map and others used the map itself as a point of reference, some used the visible properties of the objects (e.g. size, color, shape) and others traveled through landscapes, politics and economics. In the current experiment, the following **research questions**, some new and some closely related to the aforementioned conjectures and findings, will be addressed:

- **Use of Pointing Gestures:** When are pointing gestures used and in which cases are they omitted. How do the duration of the pointing gesture and the extension of the pointing device (in this case a human arm) relate to the object that is indicated?
- **Use of Linguistic Material:** Are the linguistic descriptions minimal, underspecified or overspecified? What information (e.g. preferred, absolute or relative attributes) is included in the description?
- **Interaction of Pointing Gestures and Linguistic Material** How is linguistic and gestural information combined in multimodal referring expressions? What linguistic information is left out or added if a pointing gesture is included in the referring expression?
- **Speaker's Strategies:** Which strategies for referring to objects are used (e.g. describing targets by their global position on the map, or in relation to other salient targets etc.)? How does the speaker relate a target description to the dialogue context? Are speakers consistent in their use and composition of referring expressions throughout the dialogue (e.g. entrainment)? Do they adapt their strategies to the addressee?

The experiment will directly address the following **hypotheses on production**, where we denote a target referent as t_n, where n represents the order in which the targets are referred to:

- If t_1 and t_2 are far away from each other, a reference to t_2 is more likely to include a pointing gesture, compared to the case where t_1 and t_2 are near.
- If t_1 and t_2 are far away from each other, a description of t_2 is expected to include more linguistic information compared to the case where they are near.
- If t is discourse-old, then there is less likelihood of a pointing gesture, compared to the case where t has not been referred to earlier. The amplitude of such pointing gestures is expected to be smaller.
- If t is discourse-old, then a description is expected to include less linguistic information compared to a discourse-new reference.
- If t is distinguishable only by *size* (a dispreferred property), then the descriptions is likely to include more linguistic and gestural material than descriptions of targets that are distinguishable by their colour.
- A referring expression for identifying a singleton is expected to include more linguistic and gestural material than a referring expression for identifying a target group.

3.2 Perception: The Follower

In addition to the production perspective, our experiment will also shed new light on the interpretation of multimodal referring expressions. We are particularly interested in the conditions that influence whether and how quickly an addressee successfully interpreted a referring expression. One way to measure successful reference is to take as indicative the point when the interlocutors move on from one target to the next in an itinerary. This allows one to count the number of turns or measure the time it takes from the first reference to a target to the first reference to the next target in the itinerary; the shorter the time, or the number of turns needed for identification, the easier the identification. There is, however, a danger that such a way of measuring success overestimates the time it takes to arrive at an identification, since this identification will always take place prior to moving to the next target. Moreover, how can we know that the addressee has actually identified the correct target? In our experiment this problem is addressed because we ask the follower to indicate the itinerary on his private map. Thus, the use of a camera that tracks the status of the follower's map (see Figure 1), might enable us to get a better estimate of when identification of the target takes place. In summary, our experiment will help us explore features that facilitate easy identification of targets,[6] and this will involve the following **research questions**:

- **Use of Pointing Gestures:** Are targets more easily identified, when a referring expression includes a pointing gesture? What effects does the amplitude (e.g. duration, extension) of the pointing gesture have on identification of the target by the addressee?

[6] Note that the value of these features may differ per person

- **Use of Linguistic Material:** Are targets more easily identified when the linguistic descriptions are minimal, underspecified or overspecified? Does it matter which information (e.g. preferred, absolute or relative attributes) is included or left out in the description?
- **Interaction of Pointing Gestures and Linguistic Material** What linguistic information is best combined with pointing gestures to facilitate identification? What linguistic information can be left out when a pointing gesture is included in the referring expression?
- **Addressee's Strategies:** How does the addressee check for success? By repeating or rephrasing the information that is provided by the speaker (alignment of speech, gesture or both?), or by adding extra material (e.g. relata, properties), or otherwise?

Hypotheses on perception that will be tested with this experiment:

- Target groups (consisting of 5 objects with the same features) are more easy to identify than single targets (need less time and less extensive identification by the director).
- Targets that have a prior reference in the dialogue are more easy to identify.
- Targets that are located near to the previous target are more easy to identify than targets that are located far away from the previous target.
- Ease of recognition is expected to be related to the visual attributes of the targets: Targets that differ in color and size $\leq$ Targets that differ only in color $\leq$ targets that differ only in size from their distractors.

4 Conclusion

In order to build language generation systems that produce natural and effective multimodal behaviour, a deep understanding is needed of the way human speakers choose what to say and gesture, and the impact of their choices on the hearer's ability to understand the message. This requires corpora of human–human dialogue which are annotated not just with information on the linguistic and non-linguistic realization of the speakers' utterances and non-verbal behaviour, but which also lay bare the underlying communicative situation, including the attributes of the objects that speakers refer to, and provide information on success or failure of communicative acts. The current paper reports on an effort to produce such a corpus, focussing on multimodal referring expressions. Though it is intended primarily to address a number of specific hypotheses on production and perception of multimodal referring expressions, we are also taking care to package it as a resource that might prove useful for the exploration of yet unanticipated research questions on multimodal behaviour.

Acknowledgements This work was carried out under the EPSRC platform grant 'Affecting people with natural language' (EP/E011764/1).

REFERENCES

[1] A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, 'The HCRC Map Task corpus.', *Language and Speech*, **34**, 351–366, (1991).

[2] D. Appelt, 'Planning english referring expressions.', *Artificial Intelligence*, **26**(1), 1–33, (1985).

[3] A. Arts, *Overspecification in Instructive Texts*, Ph.D. dissertation, Univiersity of Tilburg, 2004.

[4] A. Bangerter, 'Using pointing and describing to achieve joint focus of attention in dialogue.', *Psychological Science*, **15**, 415–419, (2004).

[5] E. Belke and A. Meyer, 'Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during same-different decisions', *European Journal of Cognitive Psychology*, **14**(2), 237–266, (2002).

[6] R.J. Beun and A. Cremers, 'Multimodal reference to objects: An empirical approach', in *Proceedings of the Conference on Cooperative Multimodal Communication (CLC 1998)*, pp. 64–88, (1998).

[7] S. Brennan and H.H. Clark, 'Conceptual pacts and lexical choice in conversation.', *Journal of Experimental Psychology*, **22**(6), 1482–1493, (1996).

[8] A. Bangerter & E. Chevalley, 'Pointing and describing in referential communication: When are pointing gestures used to communicate?', in *Proceedings of the workshop on multimodal output generation (MOG 2007)*, (2007).

[9] H.H. Clark and D. Wilkes-Gibbs, 'Referring as a collaborative process.', *Cognition*, **22**, 1–39, (1986).

[10] R. Dale and N. Haddock, 'Generating referring expressions containing relations.', in *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics.*, (1991).

[11] R. Dale and E. Reiter, 'Computational interpretation of the Gricean maxims in the generation of referring expressions', *Cognitive Science*, **19**(8), 233–263, (1995).

[12] Robert Dale, 'Cooking up referring expressions.', in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL-89.*, (1989).

[13] C. Gardent, 'Generating minimal definite descriptions', in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, ACL-02*, (2002).

[14] A. Gatt and K. van Deemter, 'Incremental generation of plural descriptions: Similarity and partitioning.', in *Proceedings of EMNLP'07*, (2007).

[15] A. Gatt, I. van der Sluis, and K. van Deemter, 'Evaluating algorithms for the generation of referring expressions using a balanced corpus', in *Proceedings of ENLG'07*, (2007).

[16] ISLE Natural Interactivity and Multimodality Working Group, 'Guidelines for the Creation of NIMM Data Resources', Technical Report D8.2, IST-1999-10647 Project, (2003).

[17] J. D. Kelleher and G-J Kruijff, 'Incremental generation of spatial referring expressions in situated dialog.', in *Proceedings of the joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, ACL/COLING-06*, (2006).

[18] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth, 'Deictic object reference in task-oriented dialogue', in *Situated Communication*, eds., G. Rickheit and I. Wachsmuth, 155–208, Mouton de Gruiter, (2006).

[19] P. Kühnlein and J. Stegmann, 'Empirical Issues in Deictic Gestures: Referring to Objects in Simple Identification Tasks', Technical Report 2003/3, SFB 360, Univ. Bielefeld, (2003).

[20] M. Louwerse, P. Jeuniaux, M. Hoqueand J. Wu, and G. Lewis, 'Multimodal communication in computer-mediated map task scenarios', in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, (2006).

[21] Thomas Pechmann, 'Incremental speech production and referential overspecification', *Linguistics*, **27**, 89–110, (1989).

[22] M. Pickering and S. Garrod, 'Toward a Mechanistic Psychology of Dialogue', *Behavioural and Brain Sciences*, **27**(2), 169–226, (2004).

[23] P. Piwek, 'Modality choice for generation of referring acts: Pointing versus describing.', in *Proceedings of the Workshop on Multimodal Output Generation (MOG 2007).*, (2007).

[24] P. Piwek, R.J. Beun, and A. Cremers, ''Proximal' and 'Distal' in language and cognition: evidence from deictic demonstratives in Dutch', *Journal of Pragmatics*, (June 2007). doi: 10.1016/j.pragma.2007.05.001.

[25] I. van der Sluis, A. Gatt, and K. van Deemter, 'Evaluating algorithms for the generation of referring expressions: Going beyond toy domains', in *Proceedings of RANLP'07*, (2007).

[26] K. van Deemter, 'Generating referring expressions: Boolean extensions of the incremental algorithm', *Computational Linguistics*, **28**(1), 37–52, (2002).

[27] K. van Deemter, 'Generating referring expressions that involve gradable properties.', *Computational Linguistics*, **32**(2), 195–222, (2006).

[28] I. van der Sluis and E. Krahmer, 'Generating multimodal referring expressions.', *Discourse Processes*, **44**(3), 145–174, (2007).