Multimodal Reference in Dialogue: Towards a Balanced Corpus

Paul Piwek
Centre for Research in Computing
The Open University, UK
p.piwek@open.ac.uk

Ielka van der Sluis
Computer Science
Trinity College Dublin, Ireland
ielka.vandersluis@cs.tcd.ie

Albert Gatt
Computing Science
University of Aberdeen, UK
a.gatt@abdn.ac.uk

Adrian Bangerter
Institut de Psychologie du Travail et des Organisations
University of Neuchâtel, Switzerland
adrian.bangerter@unine.ch

Introduction

Generation of Referring Expressions (GRE), e.g., Dale and Reiter (1995), is one of the core tasks of Natural Language Generation (NLG) systems. Usually it is formulated as an identification problem: given a domain representing entities and their properties, construct a referring expression for a target referent or set of target referents which singles it out from its distractors. Recently, researchers in this area have turned their attention to *multimodal referring acts*, in particular, the interaction between the two modalities of *pointing* and *describing* – e.g., Kranstedt et al. (2006), Piwek (2007), and Van der Sluis and Krahmer (2007). Additionally, psycholinguistic work is increasingly investigating the conditions governing the use of pointing gestures as part of referring acts in *dialogue*, opposed to *monologue*. Here, we present the design of an experiment on multimodal reference in two-party dialogue. The purpose of the experiment is to create a corpus that can inform the development of multimodal GRE algorithms.

Collecting a Balanced Corpus

We have paid specific attention to balancing the corpus: the conditions under which references were elicited correspond to experimental variables that are counter-balanced. The use of a dialogue setting will allow us to investigate both the speaker/generator's and hearer/reader's point of view, with potentially useful data on such factors as alignment and entrainment, and the nature of collaboration or negotiation, topics of much debate in the psycholinguistic literature (Pickering and Garrod, 2004).

In our setup for collecting dialogues, 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 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 for the interaction 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 shared 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 (Anderson et al., 1991), 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. 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.

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 across 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.

Current Status and Further Work

After a pilot of the experiment, data was collected from 22 dyads with the validated setup. Currently, the data is being transcribed, see Figure 1 for an example. Our next task is to annotate the data, focussing on 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. Elsewhere (van der Sluis et al., 2008), we provide information on the hypotheses that we intend to test on the annotated corpus.

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 F 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	D stops pointing
		with one *with these ones*.	D points
137	F	All right.	

Figure 1: 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

Acknowledgements This work was partly funded by the EPSRC platform grant 'Affecting people with natural language' (EP/E011764/1).

References

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. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34:351–366.

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

A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth. 2006. Deictic object reference in task-oriented dialogue. In G. Rickheit and I. Wachsmuth, editors, *Situated Communication*, pages 155–208. Mouton de Gruiter.

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

P. Piwek. 2007. Modality choice for generation of referring acts: Pointing versus describing. In *Procs of Workshop on Multimodal Output Generation (MOG 2007)*, Aberdeen, January.

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

I. van der Sluis, P. Piwek, A. Gatt, and A. Bangerter. 2008. Towards a balanced corpus of multimodal referring expressions in dialogue. In *Procs of Symposium on Multimodal Output Generation (MOG 2008)*, Aberdeen, Scotland, April.