Multimodal Referential Acts in a Dialogue Game
FROM EMPIRICAL INVESTIGATIONS TO ALGORITHMS

Paul Piwek
ITRI – Univ. of Brighton – UK

Robbert-Jan Beun
Comp. Science – Utrecht Univ. – NL

Keywords: deixis, demonstratives, human-human communication, multimodal dialogue, natural human-machine communication.

Human-human conversation has often been heralded as a model for human-computer interaction, the rationale being that human-computer interaction can be improved significantly if it relies on those skills and abilities which come most natural to humans. The most rigorous application of this idea can be found in recent work on embodied conversational agents (ECAs; e.g., Cassell et al., 2000).

The aim of this paper is to describe how a specification of the behaviour of ECAs can be grounded in empirical investigations into human conversational behaviour. We consider various empirical approaches and identify some pitfalls and problems which one faces when translating empirical results to algorithms for ECAs. As an illustration of the aforementioned issues, we discuss an empirical study into multimodal referential acts and the considerations which are involved in deriving algorithms for the interpretation and generation of referential acts from this study.

We would like to propose that three tasks can be distinguished in the construction of an ECA (this model is used as a starting point for our discussion, not necessarily as a description of current practice).

- In task **A** principles or regularities which are involved in human-human communication are sought on the basis of empirical studies.
- In task **B** the findings collected in task **A** are translated into one or more possible ECA algorithms. At this point, the gap has to be bridged between possibly abstract principles/regularities and algorithms which are suited for a specific application domain.
- In task **C** the performance of the ECA algorithms is evaluated with respect to a set of metrics (e.g., Sanders & Scholtz, 2000) or with respect to each other (e.g., Nass et al., 2000 compare ECAs which have been given different personality profiles with respect to each other). The evaluations involve fully implemented algorithms or, alternatively, Wizard-of-Oz type studies (e.g., Fraser & Gilbert, 1991).

The focus of this paper lies with the tasks **A** and **B**. Let us start with task **A**. Empirical approaches can be thought of as occupying a scale from situations where the experimenter has no control over the situation which s/he observes to situations where as many features of the situation as possible are under his or her control. The former situation is typical for the kind of studies which are carried out by conversation analysts whereas the latter are encountered in experimental studies. Both extremes have their advantages and disadvantages. On the one hand, conversation analytical studies involve real-world natural conversations but are often difficult to study due to parameters which are hidden from the experimenter. On the other hand, experimental studies provide the experimenter with an extensive insight into the parameters of the situation but can also lead to the study of artificial situations or situations which hardly ever occur in the real world.

In this paper, we describe an approach which occupies the middle ground. Our aim is to study fairly controlled situations which allow the subjects enough room to exhibit natural communicative

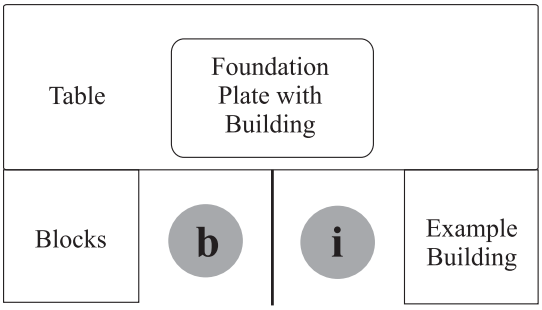

Figure 1: Set-up of the dialogue game

behaviour. We build on the insight that language use has to be understood with reference to the activity in which it takes place (e.g., Levinson, 1992; Clark, 1996). Our aim is to make sure that the parameters of this activity are known to the experimenter. This means that s/he designs such an activity, henceforth a *dialogue game*, and gets his or her subjects to communicate within the bounds of this game. We propose to define such a dialogue game in terms of four components.

A DIALOGUE GAME consists of:

1. A set of *participants*;
2. An *initial state of play*;
3. A *joint public goal state* which the participants are supposed to achieve;
4. A *role function* which assigns to each of the participants its entitlements, prohibitions and abilities to access various types of information and perform various types of action during the game.

The paper describes a concrete instantiation of such a dialogue game: (1) The set of participants consists of two subjects. (2) In the initial state the participants are separated by a non-transparent screen and facing a foundation plate (38x38cm) which is occupied by a building made of LEGO blocks of the DUPLO series (see Figure 1). One of the participants is located next to another foundation plate with an example building on it and the other is located next to a box containing more blocks. (3) The goal state is achieved when the building on the shared foundation plate is identical to the example building. (4) The leftmost participant is assigned the role of builder (**b**) and the one on the right the role of instructor (**i**). Both **b** and **i** can point at and observe all objects present on the foundation plate and they are allowed to talk with each other. Whereas only **b** is allowed to *move* the objects with his or her hands, only **i** has visual access to the example building.

Ten pairs of Dutch subjects engaged in dialogue games of the described type. Their interactions were recorded on video tape and subsequently transcribed. These data were then employed to discover regularities in the use of *deictic referential acts*, i.e., acts of direct reference to objects on the shared foundation plate. Two questions were central to our investigation:

- Under which circumstances do participants opt for multimodal communicative acts?
- How do the components of multimodal acts –in particular, linguistic and pointing acts– influence each other? That is, do pointing acts affect the descriptive content and determiner of the accompanying referring expression, or perhaps vice versa?

In order to answer the first question we needed to find a correlation between the occurrence of referential acts and the state of the dialogue at the point in time at which the act occurred. Here we benefitted from the fact that the ingredients of the dialogue game were transparent to the observer. It

turned out that, for instance, the salience of objects played an significant role. Generally speaking, subjects typically pointed to objects which were not salient at the time immediately prior to the referential act. Furthermore, we found that there was a statistically significant correlation between the use of proximate demonstratives (Dutch: *dit*, *deze*; English: *this*) and reference to non-salient objects (a finding, which at first sight seems to contradict existing analyses of English demonstratives; e.g., Gundel et al., 1993). Pointing and proximate demonstratives also occurred more often in tandem than distal demonstratives (Dutch: *die*, *dat*; English: *that*) and pointing. In the full paper we provide a detailed description of our findings. Our main point, however, is that notions such as salience could only operationalized with sufficient precision due to the fact that the experimenters knew the parameters of the dialogue game and could follow the state of the game through the video recordings. In a typical conversational analytical study (see, e.g., Sudnow, 1972 for a collection of such studies) such information often has to be guessed at, since much information is implicit in the interlocutors background knowledge.

In task **B**, the findings obtained in task **A** need to be translated to concrete algorithms. In the paper, we discuss a number of complications which can arise such as:

- It might not be possible to extract one single algorithm from the findings: different human individuals might have different communicative strategies.
- The discrepancy between algorithms consisting of definite rules for behaviour and our findings which are in terms of statistically significant correlations between variables.
- There is the problem of translating notions such as salience to measurable or identifiable phenomena in the application domain.

We will discuss these issues from two perspectives. Firstly, we consider how the results of our empirical study were used to inform an algorithm for the interpretation of referential acts in the multimodal DENK dialogue system (see Kievit et al., forthcoming). Secondly, we will consider the implications of our findings for extensions of Dale & Reiter's (1995) incremental algorithm to multimodal referential acts. Recently, an algorithm based on the data which are discussed here (see also Beun & Cremers, 1998 and Piwek et al., ms) has been put forward (Van der Sluis & Krahmer, ms).

References

Beun, R.J. & A. Cremers (1998), 'Object reference in a shared domain of conversation'. *Pragmatics and Cognition* **6**(1/2): 121–152.

Bunt, H. & R.J. Beun (forthcoming) (eds), *Advances in Multimodal Human-Computer Communication*, Lecture Notes in Artificial Intelligence 2155, Berlin: Springer.

Cassell, J., J. Sullivan, S. Prevost & E. Churchill (eds) (2000), *Embodied Conversational Agents*. Cambridge: The MIT Press.

Clark, H. (1996), *Using Language*. Cambridge: Cambridge University Press.

Dale, R. & E. Reiter (1995), 'Computional interpretations of the gricean maxims in the generation of referring expressions'. *Cognitive Science* **18**: 233–263.

Fraser, N. & N. Gilbert (1991), 'Simulating speech systems'. *Computer, Speech and Language*, **5**, 81–99.

Gundel, J., N. Hedberg & R. Zacharski (1993), 'Cognitive status and form of referring expressions in discourse'. *Language*, **69**(2): 247–307.

Kievit, L., P. Piwek, R.J. Beun, H. Bunt (forthcoming), 'Multimodal Cooperative Resolution of Referential Expressions in the DENK system', In: Bunt, H. & R.J. Beun (forthcoming).

Levinson, S. (1992), 'Activity types and language'. In: Drew, P. & J. Heritage (eds.), *Talk at work: Interaction in institutional settings*, Cambridge: Cambridge University Press, 66–100.

Nass, C., K. Isbister & E. Lee (2000), 'Truth Is Beauty: Researching Embodied Conversational Agents', In: Cassell et al. (2000), 374–402.

Piwek, P., R.J. Beun & A. Cremers (ms), 'Demonstratives in Dutch Cooperative Task Dialogues', IPO manuscript 1134.

Sanders, G. & J. Scholtz (2000), 'Measurement and Evaluation of Embodied Conversational Agents', In: Cassell et al. (2000), 346–373.

Sudnow, D. (ed.) (1972), *Studies in Social Interaction*. New York: The Free Press.

Van der Sluis, I. & E. Krahmer (ms), 'Generating Referring Expressions in a Multimodal Context: An empirically oriented approach'. To be presented at the CLIN meeting 2001.