Multimodal Referential Acts in a Dialogue Game: From Empirical Investigations to Algorithms

Paul Piwek
ITRI
University of Brighton
BN2 4GJ, Brighton, UK
+44 1273 64 29 16
Paul.Piwek@itri.bton.ac.uk

Robbert-Jan Beun
Computer Science Department
Utrecht University
3508 TB, Utrecht, NL
+31 30 253 3577
rj@cs.uu.nl

1 Abstract

The paper presents an investigation into the question how a specification of the behaviour of Embodied Conversational Agents (ECAs) can be grounded in empirical investigations of human conversational behaviour. We consider various empirical approaches and discuss one particular approach based on Dialogue Games in detail. We identify some pitfalls and problems which one faces when translating the results of such empirical investigations to algorithms for ECAs. Our discussion is illustrated by means of a specific investigation into the use of deictic referential acts in conversations.

1.1 Keywords

Deixis, demonstratives, human-human communication, multimodal dialogue, natural human-machine communication.

2 Introduction

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., [3]).

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 deictic referential acts (i.e., acts of direct reference to objects which are physically presents in the dialogue situation) 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., [15]) and possibly compared with the performance of other ECAs or a condition with no ECA (e.g., [13] compare ECAs which have been given different personality profiles with respect to each other and [5] discuss various studies which investigate whether the presence of an ECA is beneficial). The evaluations involve fully implemented algorithms or, alternatively, Wizard-of-Oz type studies (e.g., [6]).

The focus of this paper lies with the tasks A and B. In task A information about human-human communication is obtained through empirical studies. Such 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 (or more generally, ethnomethodologists), whereas the latter is encountered in experimental studies. Both extremes have their advantages and disadvantages. On the one hand, the situations examined by conversation analysts involve real-world natural conversations but are often difficult to study due to parameters which are hidden from the experimenter. On the other hand, tightly controlled experimental set-ups 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.

3 Empirical Investigations through Dialogue Games

In this paper, we describe an empirical approach which occupies the middle ground between conversation analytical and experimental studies. Our aim is to study fairly controlled situations which allow the subjects enough room to exhibit natural communicative 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., [12, 4]). 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.

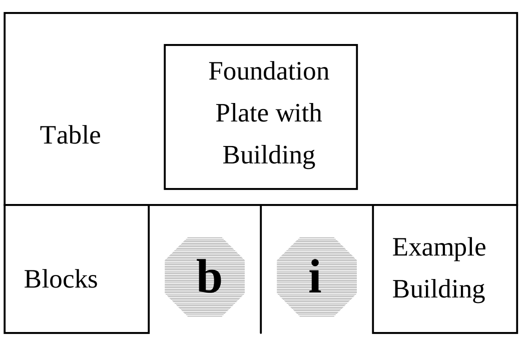

Figure 1: Set-up of the dialogue game

We illustrate how dialogue games can be employed to inform the task of building ECAs by looking at the following instantiation of 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. One of the participants is located next to a second foundation plate with an example building on it and the other is located next to a box containing more blocks (see Figure 1).
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.

4 Results on Deictic Referential Acts

The data obtained from the dialogue games in which the subjects engaged were used to discover regularities in the use of *deictic referential acts*, i.e., acts of direct reference to objects on the shared foundation plate. In total the collected dialogues contained 143 deictic referential acts. We investigated 126 of these, excluding 13 plural referential acts and 4 instances of miscommunication and repair.

A detailed report of our findings can be found in [14]. Here we provide a brief summary of these findings. We are interested in correlations between the form of deictic referential acts and parameters of the dialogue situation.

The identification of the relevant parameters is informed by the notion of *deixis*, i.e., *the force with which the addressee is instructed to find the referent* ([7:65]). According to [10] referential acts can be marked for either low or high deixis. It is hypothesized that speakers draw on three strategies when deciding to mark for either high or low deixis:

1. **Givenness**: Direct the attention of the addressee strongest to the entities which are not given in the addressee's consciousness;
2. **Noteworthiness**: Direct the attention of the addressee with most force to the entities which are of most interest (to the speaker);
3. **Foregrounding**: If you need to signal high deixis, then use multiple means to do so (e.g., in addition to choosing a particular type of referring expression also point to the object in question).

In our dialogue game, the notions of Givenness, Noteworthiness and Foregrounding can be operationalized along the following lines:

- **Givenness** of the intented referent of a referential act. We consider a referent to be +Given if and only if it satisfies at least one of the following two conditions: (**1**) The object is part of the so-called *focus area*. The focus area is either the area to which the speaker explicitly drew the attention of hearer as in ‘Now there completely to the left,…’ or the area made up of the object which the speaker referred to most recently and the objects adjacent to this object. (**2**) The object has a shape which is different from the shape of the majority of other objects on the foundation plate. In other words, the object is visually salient. In our particular set-up most blocks are beam-shaped except for a few blocks concave/convex contours.
- **Noteworthiness** of the intended referent of a referential act. According to [10] the objects which the speaker is most interested in talking about are considered noteworthy. We assume that noteworthiness depends on the task at hand. In our dialogue game this task involves the manipulation of blocks. Therefore, we propose that those blocks that the speaker intends the hearer to manipulate are most noteworthy as opposed to, for instance, blocks which the speaker only refers to in order to identify another block. Consider: ‘The yellow one behind the green cube has to be moved’. Here, the yellow one is noteworthy, whereas the green cube is not.
- **Foregrounding** of an entity. We assume that this is done by pointing to an object.

Our results concerning the relation between the form of referential acts and the aforementioned strategies are summarized in the tables 1 and 2.

Form of Referential Act	**+Given**	**-Given**
Distal N’ + preposed postm. + ↵	1	0
Distal N’ + postmodifier	17	3
Distal N’ + ↵	18	5
Distal N’	24	3
Distal + ↵	2	0
Proximate N’ + ↵	8	11
Proximate + ↵	5	5
Definite article N’	8	1
Indefinite article N’	5	5
N’	4	1
Total	92	34
Percentage	73%	27%

Table 1: The distribution of referential acts over +Given and –Given entities

In table 1. distal stands for distal demonstrative (Dutch: *die, dat*; English: *that*) and proximate stands for proximate demonstrative (Dutch: *dit, deze*; English: *this*). The arrow (↵) is used to denote a pointing act. Furthermore, N’ ranges over the following constructions: an adjective (as in: *die rode; that red*), adjective + noun (as in: *die rode balk; that red bar*) and adjective + adjective + noun.

Demonstrative	**+Noteworthy**	**–Noteworthy**
Distal	40	33
Proximate	23	6

Table 2: The distribution of distals and proximates over +/– Noteworthy entities

Our hypothesis is that *distals signal low deixis* whereas *proximates signal high deixis*. According to this hypothesis we expect proximates to be used in situation in which the strategies for signalling high deixis are triggered, whereas distals should not occur in situations where these strategies are triggered. Our findings confirm the hypothesis in almost all respects:

- *Proximates are preferred to refer to –Given entities*. The distribution of proximates over +Given and –Given entities is 13:16 ($\chi^2_{df=1}$=11.67, P<0.005, where our null hypothesis is based on the distribution of all 126 referential acts over +Given and –Given entities, i.e., 92:34).
- *Distals are preferred to refer to +Given entities*. The distribution of distals over + and –Given entities is 62:11 ($\chi^2_{df=1}$=5.27, P<0.025).
- *Although our data show a preference of proximates for noteworthy entities, this preference is statistically not significant*. The ratio of + and –noteworthy is 23:6 ($\chi^2_{df=1}$=3,69, P<0.1). The result is very close to being statistically significant, but just below the $\chi^2_{df=1}$=3,84, P=0.05 dividing line.
- *Proximates are always used in combination with a pointing act*. The distribution of with and without pointing is 29:0 ($\chi^2_{df=1}$=24,70, P<0.001, where our null hypothesis is based on the distribution of pointing and no pointing over all 102 demonstrative referential acts).
- *Distals are preferred when there is no pointing act present*. The distribution of with and without pointing is 26:47 ($\chi^2_{df=1}$=9,93, P<0.01).

Our operationalization of the notion of Givenness depends on the state of play that is specific to our dialogue game. Similarly, the notion of Noteworthiness depends on the task which in turn is derivable from the definition of the goal state and the role function of the dialogue game. We found that the role function has a further effect on the form of referential acts:

- None of the 10 **builders** used postmodifiers in their referential acts.
- 6 out of the 10 **instructors** used postmodifiers (as in 'the red block *behind the yellow cube*'). For these 6 individuals, the distribution of postmodifier use was: 7, 5, 4, 2, 1, 1. Interestingly, the three most frequent users of postmodifiers never pointed to an object.

5 From Empirical Results to Algorithms

In this section we turn our attention to what we have called task B in section 2: the translation of empirical results into ECA algorithms.

5.1 From the Specific to the General and vice versa

At first sight it might seem that our findings from task A will not be very useful when it comes to building an ECA for a totally different domain/task than the one we studied. It is, however, a mistake to assume that we intend to project findings from one dialogue game *directly* onto completely different ones. Rather, it has to be kept in mind that task A involves *general* hypotheses which are empirically tested through a specific dialogue game. These hypotheses are based on the general notion of *deixis* and the strategies associated with it.

Some evidence for the generality of our findings comes from the fact that the similar results have been obtained for the use of Dutch *anaphoric* demonstratives. [10] found that when Dutch demonstratives are used to refer linguistic antecedents proximates signal high deixis whereas distals signal low deixis. For instance, [10:365] found through the study of a corpus of Dutch texts that proximates are preferred to refer to linguistic antecedents which are further back in the discourse (and whose givenness has therefore 'decayed'). A discussion of the relation between our findings and those for English demonstratives can be found in [14]. There it is suggested that the opposition between distal and proximate demonstratives in English can also be explained along the lines proposed in this paper. However, this involves reinterpreting a number of empirical studies of English demonstratives (e.g., [8]).

Given that the findings which we obtained through our empirical study are sufficiently general, we still face the task of translating these findings to the application domain of a specific ECA. This involves analyzing the interactions in this domain in terms of a dialogue game consisting of participants, an initial state, a goal and a role function. This analysis can then be used to determine which objects in the domain count as given or foregrounded (we have left out the notion of noteworthiness since our results on the effect of noteworthiness did not pass the test of statistical significance). It is undeniable that at the moment there is no ready-made recipe for this task. In the end, the builder of the ECA will have to exercise his or her judgement to isolate those parameters of the dialogue game which determine the givenness and foregrounding of objects. For instance, when dealing with a situation where objects are presented in the usual windows on a computer system, which objects are given to the user might depend on whether the object inhabits an activated (i.e., highlighted) window. This assumption is, for instance, made by the algorithm for the interpretation of referential acts in the DENK system ([9]). Our point is that such decisions should be made on the basis of a systematic analysis of the task at hand in terms of a dialogue game, and theories of human behaviour which are supported by empirical studies.

5.2 Algorithms for Individuals

The theory of deixis which we used to conduct our empirical investigation into deictic referential acts has also some limitations which we need to be aware of when we use it to inform the construction of an algorithm. In particular, the focus of the theory is on assigning meanings or functions to expressions. It does, however, not provide us with any information on what to do when the same function or meaning can be expressed by different expressions.

For instance, in our study pointing and (normal) postmodifiers never co-occur. One possible explanation for this is that they serve the same purpose: an extra means for identifying an object in addition to the core referering expression of the form *determiner N'*. Now, we could use the data (on the distribution of pointing acts and postmodifiers) from table 1 and conclude that our algorithm should generate a postmodifier with the probability of 0.34 and a pointing act with the probability of 0.66 in case this required to identify the intended referent. However, this way we are likely to end up with an algorithm which is not a model of the behaviour of actual human language users. The choice between the two options is not a probabilistic one made by individual users, but rather there are groups of users which consistently choose one or the other. In particular, we found that in our dialogue games builders never use postmodifiers. Instructors, on the other hand come in two categories: those who only use postmodifiers and those who very incidentally (once or twice during a complete dialogue) use a postmodifier.

Thus the choice between pointing or postmodifier should depend on the role of the ECA in the dialogue game not on probabilities derived from the heterogeneous data. In fact, the choice between pointing or no pointing might be due to an even more general principle. [17] have suggested on the basis of the data which are presented in [14,1] that a speaker only points if the object in question is sufficiently close to his or her hand. This might explain why builders, – who are manipulating the blocks– sometimes point, whereas instructors never point. The role of the instructors does not involve the removal of blocks and therefore their hand is likely to be more distant from the blocks.

6 Conclusions

In this paper we have argued that dialogue games are a useful tool for the investigation of human-human conversations. They occupy the middle ground between rigorously controlled empirical studies and open-ended conversation analytical studies. We argued that an empirical study into human-human behaviour should be grounded in a general theoretical framework. This enables us to obtain general findings on human-human conversations which can subsequently be used to inform algorithms for ECAS. This does involve creative input from the builder of the ECA in order to determine exactly how the general findings relate to the situation in which the ECA is deployed, at least as long as no general computational theories of salience (givenness), relevance (noteworthiness) and related notions are available. We argued, however, that the analysis of the situation in terms of a dialogue game gives the ECA builder a handle to systematically undertake this task.

We have also warned for the dangers involved in using the empirical data to motivate probabilistic algorithms. A case in point was the generation of postmodifiers versus pointing acts. In our data these are mutually exclusive and therefore it is tempting to base their generation on the probabilities which can be derived from their distribution in the data. However, a closer look at the data revealed that the choice between the two is correlated with the type of speaker (instructor versus builder and the personal style of the speaker). Hence, we conclude that the attitude that the empirical findings should give rise to one correct algorithm for natural conversational behaviour should be abandoned in favour of the outlook that data obtained from various speakers can harbour a multitude of natural conversational behaviours.

REFERENCES

1. Beun, R.J. & A. Cremers. Object reference in a shared domain of conversation. *Pragmatics and Cognition 6(1/2)* (1998), 121-152.
2. Bunt, H. & R.J. Beun (eds). *Cooperative Multimodal Communication*. Lecture Notes in Artificial Intelligence 2155, Berlin/Heidelberg: Springer, 2001.
3. Cassell, J., J. Sullivan, S. Prevost & E. Churchill (eds). *Embodied Conversational Agents*. Cambridge: The MIT Press, 2000.
4. Clark, H., *Using Language*. Cambridge: Cambridge University Press, 1996.
5. Dehn, D. & S. van Mulken, The impact of animated interface agents: a review of empirical research. *Int. J. Human-Computer Studies 52* (2000), 1-22.
6. Fraser, N. & N. Gilbert. Simulating speech systems. *Computer*, *Speech and Language 5* (1991), 81-99.
7. García, E. *The Role of Theory in Linguistic Analysis: The Spanish Pronoun System*. North Holland, 1975.
8. Gundel, J., N. Hedberg & R. Zacharski. Cognitive status and form of referring expressions in discourse. *Language 69 (2)* (1993), 247-307.
9. Kievit, L., P. Piwek, R.J. Beun, H. Bunt. Multimodal Cooperative Resolution of Referential Expressions in the DENK system, In: [2], 2001.
10. Kirsner, R. Deixis in discourse: An exploratory quantitative study of the modern dutch demonstrative adjectives. In: T. Givón (ed.), *Discourse and Syntax Vol. 3*, Academic Press, New York, 355-375, 1979.
11. Kirsner, R. and V. van Heuven. The significance of demonstrative position in modern Dutch. *Lingua 76* (1988), 209-248.
12. Levinson, S. Activity types and language. In: Drew, P. & J. Heritage (eds.), *Talk at work: Interaction in institutional settings*, Cambridge: Cambridge University Press, 66-100, 1992.
13. Nass, C., K. Isbister & E. Lee. Truth Is Beauty: Researching Embodied Conversational Agents. In: [3], 374-402, 2000.
14. Piwek, P., R.J. Beun & A. Cremers, Demonstratives in Dutch Cooperative Task Dialogues. IPO *manuscript 1134*, 1995. (http://www.itri.bton.ac.uk/~Paul.Piwek)
15. Sanders, G. & J. Scholtz. Measurement and Evaluation of Emboddied Conversational Agents. In: [3], 346-373.
16. Sudnow, D. (ed.). *Studies in Social Interaction*. New York: The Free Press, 1972.
17. Van der Sluis, I. & E. Krahmer, Generating Referring Expressions in a Multimodal Context: An empirically oriented approach. Manuscript (Presented at the CLIN meeting 2000, Tilburg).