In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). CTIT Workshop Proceedings.
I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds), University of Aberdeen, 25-26 January 2007. pp.129-139.
ISSN 1574-0846

Modality Choice for Generation of Referring Acts* Pointing versus Describing

Paul L.A. Piwek
NLG group, Centre for Research in Computing
The Open University, Walton Hall, Milton Keynes, UK
p.piwek@open.ac.uk

Abstract

The main aim of this paper is to challenge two commonly held assumptions regarding modality selection in the generation of referring acts: the assumption that non-verbal means of referring are secondary to verbal ones, and the assumption that there is a single strategy that speakers follow for generating referring acts. Our evidence is drawn from a corpus of task-oriented dialogues that was obtained through an observational study. We propose two alternative strategies for modality selection based on correlation data from the observational study. Speakers that follow the first strategy simply abstain from pointing. Speakers that follow the other strategy make the decision whether to point dependent on whether the intended referent is in focus and/or important. This decision precedes the selection of verbal means (i.e., words) for referring.

Keywords: generation of referring expressions, choice and integration of output modalities, cognitive aspects, pointing

1 Introduction

In the field of Natural Language Generation, *referring expressions* are defined as '[...] phrases that identify particular domain entities to the human recipient of the generation system's output' (Dale & Reiter, 1995). More precisely, referring expressions are used in *referring acts* to identify domain entities. *Multimodal referring acts* combine verbal means for referring (the aforementioned phrases) with non-verbal means such as pointing. The main aim of this paper is to challenge two assumptions that are commonly held in the literature (see Section 2.1) on generation of multimodal referring expressions:

> (A1) Non-verbal means of referring are secondary to verbal means and only to be resorted to when verbal means are judged inadequate.
>
> (A2) There is a *single* strategy for the choice of output modality when generating of referring acts.

We proceed as follows. Firstly, in Section 2, we discuss and compare several existing approaches to the generation of multimodal referring acts. We also consider some arguments for and against developing algorithms that model human production of multimodal referring acts. In the next section, we examine the existing approaches in the light of data from an observational study, focusing on assumptions A1 and A2. Our own method for modality choice is presented in Section 4. Based on the observational study, we will distinguish between two strategies for choice of output modality. We will also argue that both *domain focus* and *importance* of the intended referent are principal factors that influence modality choice. We conclude this paper with Section 5, where we present our conclusions and some issues for further research.

*I would like to thank the three anonymous reviewers of the MOG workshop for their extremely helpful comments and suggestions.

In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). CTIT Workshop Proceedings.
I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds), University of Aberdeen, 25-26 January 2007. pp.129-139.
ISSN 1574-0846

2 Background

Before we proceed, let us lay out the basic assumptions underlying the current study. We focus on referring acts that are produced by a speaker and intended to be understood by an addressee. A referring act is understood if the addressee identifies the object that the speaker had in mind. We focus on situations where several objects are present in the domain and visible to both speaker and addressee. Only one of these objects is the *intended referent* (i.e., we concentrate on singular reference); the other objects are known as *distractors*. The speaker can use spoken language and/or pointing to identify objects in the domain. For instance, when referring to a particular object she might utter 'it', 'that tower with the rounded corners' or 'this tower' accompanied by a pointing act. The main concern of this paper is modality choice, in particular, the decision whether to include a pointing act.

2.1 Related Work

Most work on the generation of multimodal referring acts is based on the assumption that pointing acts are only used if 'proper' verbal means of referring do not suffice. For example, Lester et al. (1999) only include a pointing act, *if* a pronoun cannot be used to refer to the object, and Claassen (1992) goes even further and only resorts to pointing acts when no purely verbal means of identification can be found. Van der Sluis and Krahmer (2001) also see pointing as a last course of action; they use a pointing act only if the object is sufficiently close and a purely verbal referring act would be too complex.

A common thread underlying all these approaches is that the proposed algorithm first tries to formulate an exclusively verbal referring expression, and only if that fails, or the resulting expression becomes to complex, does it start anew and produce a multimodal referring act that includes pointing. This seems like a rather inefficient way to create referring acts, especially given that speakers do often point (see Section 3), and precise pointing is rather effective (since precise pointing rules out *all* distractors at once).

The aforementioned algorithms all focus on precise pointing: the object that is pointed at is uniquely identified by the pointing act. Typically, this is achieved by the (computer-animated) pointing hand touching the object or being very close to it. In contrast, an imprecise pointing act does not single out a unique object, but rather identifies a set of objects (*cf.* Wilkins 1999 and Clark & Bangerter 2004 on close and distant pointing). More recently, some computational work has also considered imprecise pointing. Kranstedt and Wachsmut (2005) speak of object-pointing versus region-pointing. They propose a way – based on a pointing cone originating in the demonstrating hand – for computing whether a particular pointing gesture constitutes precise or imprecise pointing. This then influences the set of distractors that the generation algorithm uses for determining the linguistic content of the referring act. The decision whether or not to point is based on whether the intended referent is visible to both interlocutors: if it is visible, then a pointing act is included. We will, however, see (Section 3) that speakers do not always point even when the intented referent is visible to all interlocutors.

Krahmer and Van der Sluis (2003) propose an account that not only covers different levels of precision of pointing, but also aims to make the decision on whether to point without a priori favouring verbal or non-verbal means of referring. This is achieved by assigning costs to both verbal and non-verbal components of referring expressions. The cost function is set up such that the cost of a non-verbal pointing act is somewhat higher than that of a single verbal component (inclusion of a property such as type or colour). Overall cost is calculated by summing the component costs. Additionally, they introduce three different degrees of precision for pointing: precise, imprecise and very imprecise pointing. The idea is that pointing is a bit like highlighting objects with flashlight: when one is close to the objects it is easy to single out one specific object, whereas as one moves further away, a bigger area is illuminated by the flashlight, thus making it more likely that several objects are highlighted rather than just one. They propose that as precision decreases cost of pointing also decreases.

This approach hinges very much on the values that are assigned by the cost function. Currently, the cost of pointing acts is derived from Fitts's (1954) Index of Difficulty, according to which the

In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). CTIT Workshop Proceedings.
I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds), University of Aberdeen, 25-26 January 2007. pp.129-139.
ISSN 1574-0846

difficulty to reach a target is a function of the size of and the distance to a target. This empirically validated index is adapted to the current application by replacing *distance to a target* with *distance from the current position of the hand to the position at which pointing takes place*. Whether this substitution preserves the correctness of Fitts's Index is an open question. More importantly, in addition to the cost for pointing acts, comparable costs for components of referring expressions need to be assigned. Krahmer and Van der Sluis propose without further justification that '[...] type edges (block) are for free, color edges cost 0.75, size edges cost 1.50 and relational edges 2.25.'. In short, a weak aspect of the proposal is that it requires quite a few parameter settings which might be difficult to obtain empirically.[1]

2.2 Automated Generation versus Human Production

All the work we have discussed so far deals with the automated generation of referring acts. In the next section, we examine these algorithms in the light of data from an observational study on multimodal reference by humans. The focus will be on the work by Krahmer and Van der Sluis (2003) which, in our view, represents the most advanced proposal so far. This proposal shares with those by Kranstedt and Wachsmut (2005) and Van der Sluis and Krahmer (2001) the desire to present an algorithm that models human production. The other proposals that we have mentioned are different in this respect.

Lester et al. (1999) introduce the notion of *deictic believability*. Deictic believability is ascribed to a lifelike agent if it simultaneously achieves the following three goals: 1) its spatial references are non-ambiguous, 2) it refers to objects whilst being immersed in the environment (just like human beings can, for example, refer by pointing, speaking and walking at the same time, thus combining gesture, speech and locomotion), 3) its references are pedagogically sound. Both 1) and 3) derive from the learning context for which Lester et al. developed their COSMO system. They are independent of the aim to model human production. Arguably, 2) does suggest that modeling of human production is desirable when developing an algorithm for multimodal generation. Given that human behaviour is generally considered believable (i.e., it generally creates the impression of a sentient being with its own personality and mental states, see Bates 1994), emulating such behaviour could be a good strategy for achieving believability. For Lester et al., believability is, however, not a goal in itself. Rather they argue for it on the basis of the benefits it brings. In particular, they refer to a study (Lester et al., 1997) which showed that believable pedagogical agents are able to produce the *persona effect*: the lifelike character in a learning environment might not have a direct measurable effect on the learning of the students, but it can improve their perception of the learning experience.

The work by Claassen (1992) was done in the context of the EDWARD system. This system was conceived as a prototype multimodal user interface for studying the use and usefulness of multimodal inferfaces. The aim was 'making interaction between a user and a computer more like normal day to day interaction' (Huls & Bos, 1998: 315). Although emulation of human behaviour does not necessarily lead to the most useful systems, it certainly is a good starting point when one is trying to make computer-human interaction more like everyday human-human interaction.

In summary, we have seen that the requirements for certain systems that generate multimodal referring acts are, at least at first sight, independent of the issue of whether or not to model human production. These requirements include *reduction of ambiguity*, *pedagogical soundness* and *interface usability*. It should also be noted that the possibilities afforded by multimodal interfaces can lead to adoption of ways to realize referring acts that do not correspond with human realization. An early example is the CUBRICON system (Neal et al., 1989). This system has the capability to 'point' to the same object simultaneously in different ways: if the object appears in several windows on the computer screen, the strategy is to produce a strong pointing gesture (blinking icon and pointing text box) to the object in the activated window and weak pointing (only highlighting) in all other windows in which the object appears.

[1] Cost functions might, however, have wider applications in the field of referring expressions generation. See, for example, Khan et al. (2006) for further phenomena that may yield to analyses in terms of cost functions.

In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). CTIT Workshop Proceedings.
I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds), University of Aberdeen, 25-26 January 2007. pp.129-139.
ISSN 1574-0846

We have also seen that requirements such as *believability* and *naturalness* do suggest that models of human production of multimodal refential acts are a good starting point for building algorithms, given that human behaviour is our main yardstick for what is considered natural and believable.[2] In this connection, it is also worthwhile to consider an argument that has been put forward by Dale & Reiter (1995:253). This argument suggests that algorithms for generating referring expressions based on psycholinguistic data might be superior to those based on abstract principles (such as the Gricean maxims of conversation). Let us assume that Grice's notion of implicature (Grice, 1975) is correct and can be paraphrased as saying that if a speaker produces an utterance that is unexpected, then the addressee is likely to attempt to infer a reason for why the speaker did not use the expected utterances. Our second premiss is that a system that emulates human behaviour would be more likely to produce expected utterance. If we take these two premisses for granted, then the conclusion follows that systems that are based on a model of real human behaviour are less likely to cause the addressee to erroneously infer unintended reasons for the choice of expression. In contrast, a system based on abstract principles (e.g., avoid ambiguity) might cause the addressee to make inferences purely as result of the unexpected choice of words in the situation at hand (unless, of course, 'avoid ambiguity' is a principle that human speakers tacitly follow anyway).

Last, but not least, let us not forget that even someone who is not persuaded by any of these arguments will hopefully nevertheless concede that computational modeling of human production of referring acts is a valid topic of scientific study in its own right.

3 Empirical Findings from an Observational Study

In this section, we present a number of findings that were derived from a corpus of video-recorded task-oriented spoken dialogues (Cremers, 1996; Beun & Cremers, 1998) obtained in the setting illustrated by Figure 1.[3] The corpus consists of a total of 20 dialogues between Dutch speaking interlocutors. The dialogues arose during a game that the interlocutors were asked to play. In this game, there are two roles, that of the Builder (B), on the right in Figure 1, and that of the Instructor (I). In front of B and I, there is a workspace. The aim of the game is for B to build a structure in the workspace that is a copy of the example structure next to I (on the left in Figure 1). Only I can see the example structure. I and B are, however, allowed to talk with each other and they can also point at the (LEGO) blocks in the workspace. Only B is allowed to move blocks.

For the current study, we used 10 out of the 20 dialogues. At the start of these dialogues, several blocks already occupied the visually shared foundation plate. We examined the initial references to these objects. A total of 121 singular initial references was found after discounting 14 initial plural referring acts, 2 cases of misunderstanding, and 2 cases of self-correction.

Finding 1: *Out of a total 121 singular referring acts in the corpus, 53 included a pointing gesture (Figure 2).* In other words, almost half of the referring acts involved pointing. Such frequent use of pointing suggests that it is more than simply a fall-back strategy for situations where purely verbal referring acts are not adequate.

Finding 2: *The average number of linguistically realized properties in purely verbal referring acts was 1.7, whereas in referring acts that included pointing the average number of properties was 0.98. (two-tailed highly significant at $P \leq 0.0001$, $t = 4.9790$, $df = 119$). See Figure 3 for an overview of the distribution of properties.* This finding is compatible with the Krahmer and Van der Sluis (2003) algorithm; it seems like there is indeed a trade-off between pointing and verbal referring. Additionally, Krahmer and Van der Sluis argue that imprecise pointing is less costly, but also less discriminative. This suggest that it will co-occur with more descriptive content. This could not be verified in the current study, but other work in which distance to target was

[2] The recently emerging user interface paradigm of embodied conversational agents/lifelike characters (Cassell et al. 2004; Prendinger & Ishizuka, 2004) is partly based on the idea that human-computer interaction can be improved by making it more like human-human interaction.

[3] See Cremers (1993) for a written transcript of the corpus.

In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). CTIT Workshop Proceedings.
I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds), University of Aberdeen, 25-26 January 2007. pp.129-139.
ISSN 1574-0846

Figure 1: Set-up for collection of task-oriented spoken dialogue corpus involving two roles: Instructor (I) on the left and Builder (B) on the right

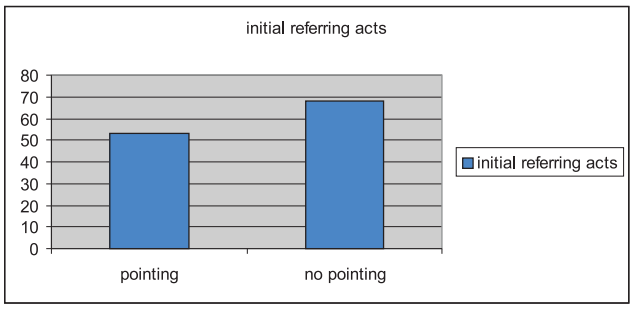

Figure 2: Number of initial referring acts that included and did not include a pointing act.

In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). CTIT Workshop Proceedings.
I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds), University of Aberdeen, 25-26 January 2007. pp.129-139.
ISSN 1574-0846

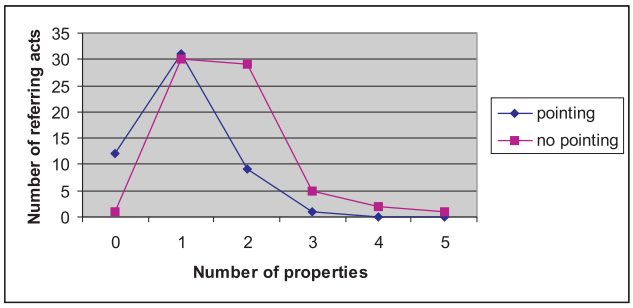

Figure 3: Number of initial referring acts with and without a pointing act mapped against the number of properties expressed by the referring act.

experimentally controlled does seem to confirm this idea (e.g., Bangerter, 2004; Krandstedt et al., 2006).

Finding 3: *Speakers tend to point more frequently when the referent is not in domain focus. When the referent is not in domain focus they point 66% (23/35) of the time, whereas when the referent is in focus they point only 35% (30/86) of the time (significant at* $P \leq 0.01$, $\chi^2 = 9.60$, $df = 1$*). See Figure 4.* We discern two types of focus of attention: *discourse focus* (Grosz, 1977) and *domain focus* and concentrate on the latter notion. This notion applies because we are dealing with *initial* identification of objects in a visually accessible domain of discourse. An object is part of the domain focus if it satisfies one of the following two criteria (*cf.* Cremers, 1994):[4]

1. The object was referred to in the preceding utterance or is adjacent to an object that was referred to in the preceding utterance (note that for the initial referring acts that we are considering, it is always the second of these two conditions that is met when an object qualifies as being part of the domain focus);

2. The object lies in an area to which the speaker explicitly directed the attention of the addressee. This is marked by what we will call *focussing expressions* as in 'Wat nou helemaal naar voren zit, daar zit die rode dwars' (literally: *What now entirely to the front is, there is that red one diagonally.* Loose translation: If you look at the bit in the front, you will find a red diagonally placed block).

Finding 3 is also compatible with Krahmer and Van der Sluis (2003). When an object is not one of the objects that are in focus (i.e., –F) there is typically a larger set of distractors than when the object is in focus. As a result, reference to a –F object is more likely to involve a pointing act because the cost of the pointing act is balanced by the fact that with a large number of distractors a large number of linguistic properties is needed. In contrast, if the referent is +F, there are few distractors and consequently few properties are required to identify the object. In that situation, pointing is too expensive to replace verbal identification.

Finding 4: *Speakers tend to point more frequently when the referent is important. When the referent is important they point 55% (42/76) of the time, whereas when the referent is not important they point only 24% (11/45) of the time (highly significant at* $P \leq 0.001$, $\chi^2 = 10.90$, $df = 1$*). See*

[4]Cremers's (1994) notion of domain focus seems to bear some relation to the notion of an active object as proposed recently in Gergle (2006). Active objects are objects that were recently moved in the shared workspace. Roughly speaking, Cremers's notion of a domain focus can be viewed as often coinciding with the active object together with the objects in its immediate surroundings (in the current study, if an object was referred to in the preceding utterance, it was typically manipulated immediately after that, and would therefore count as an active object).

In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). CTIT Workshop Proceedings.
I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds), University of Aberdeen, 25-26 January 2007. pp.129-139.
ISSN 1574-0846

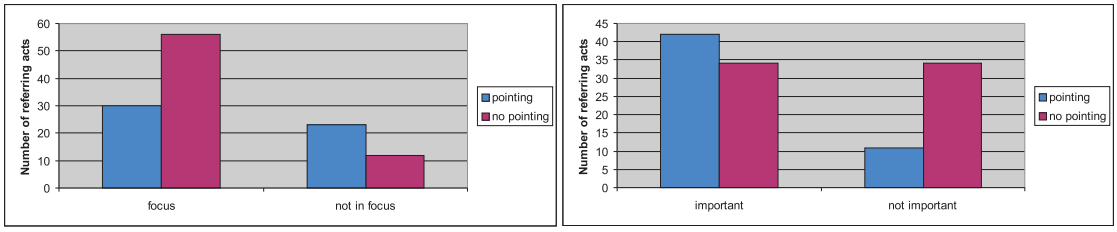

Figure 4: Number of initial referring acts with and without a pointing act mapped against whether the intended referent is in or out of domain focus, and important or not important.

Figure 4. For the purpose of the current study, an object was classified as + Important (+I) at time t if the speaker instructed the addressee to manipulate the object at t. All other objects, at the same point in time t, were labeled – Important (–I). Our finding on the influence of importance on modality choice does not seem to be reflected in any of the extant algorithms.

Finding 5: *Speakers appear to follow at least two different strategies when referring to objects. In particular, there seem to be two types of speaker: those that* never *use pointing and those that do use pointing (see Tables 1 and 2).* Approximately 29% of the speakers never use pointing (2I,12B,16I and 20I). For the remaining speakers, both –F and +I are good predictors for pointing: (a) for –F, 6 speakers point in the majority of cases, 2 speakers point in half of the cases (10I and 18B) and for the remaining 2 speakers we have no referring acts to –F objects; (b) for +I, 6 speakers point in the majority of cases, 2 do not (6I and 10I) and for the remaining 2 we have no referring acts to +I objects.

When we consider only the data for subjects that do at least point once, then we get the following statistics: (A) for referring acts to +F and –I objects, we have 8 referring acts involving pointing and 13 without pointing. (B) for referring acts to –F and +I objects, we have 20 referring acts involving pointing and 4 without pointing.

Participant		+F		–F	
Dialogue	Role	+P	–P	+P	–P
2	B	1	2	0	0
2	I	0	3	0	2
4	I	7	0	2	1
6	I	2	4	1	0
8	I	5	1	5	0
10	I	4	6	1	1
12	B	0	2	0	0
12	I	3	2	4	0
14	I	3	0	4	0
16	I	0	15	0	1
18	B	0	3	1	1
18	I	4	5	5	2
20	B	1	0	0	0
20	I	0	13	0	4

Table 1: Per dialogue participant (identified by dialogue number and role), the number of referring acts with and without pointing (+/– P) is listed for objects in and out of Focus (+/– F)

In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). CTIT Workshop Proceedings.
I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds), University of Aberdeen, 25-26 January 2007. pp.129-139.
ISSN 1574-0846

Participant		–I		+I	
Dialogue	Role	+P	–P	+P	–P
2	B	1	2	0	0
2	I	0	2	0	0
4	I	1	0	8	1
6	I	0	0	3	4
8	I	1	1	9	0
10	I	4	3	1	4
12	B	0	1	0	1
12	I	2	1	5	1
14	I	0	0	7	0
16	I	0	8	0	8
18	B	1	4	0	0
18	I	1	2	8	5
20	B	0	0	1	0
20	I	0	10	0	7

Table 2: Per dialogue participant (identified by dialogue number and role), the number of referring acts with and without pointing (+/– P) is listed for not Important and Important objects (–/+ I)

4 Strategies for Modality Choice

Based on the findings that are listed in the previous section we suggest that there are at least two (mutually exclusive) strategies for modality choice.

> Strategy 1: Never use pointing. Formulate the referring expression using only verbal means of expression.

> Strategy 2: Prefer pointing, if the intended referent is not in focus (–F) *or* important (+I).

Strategy 1 is directly based on the data from our observational study which had a substantial proportion of speakers that never pointed. For these speakers, the factors +/– F and +/– I consistently seemed to have no impact. We have not been able to find a factor that might explain why these speakers never pointed. Note that such a factor (or collection of factors) can take two very different forms: 1) there is something to the specific situations in which these speakers referred which made them refrain from pointing (and which would have made all the speakers involved in the study refrain from pointing). This would mean that if this factor is taken into account, we can merge the strategies 1 and 2 into a single strategy for all speakers. 2) The reason for not pointing lies not so much with the situation, but rather with the speakers themselves. This would mean that we have two types of speakers: those following strategy 1 and those following strategy 2. This is not so different from, for example, saying that there are speakers with different personalities (and consequently ways of expressing themselves) or, at the extreme end, individual styles. Here, we want to offer this second option as hypothesis that needs further investigation. The assumption that under the same circumstances different people have identical inclinations to refer by pointing is called into question by the current study and is in need of further investigation through controlled experiments. Additionally, note that observational reports from sociologist suggest that the inclination to point might vary between subcultures of the same language community (e.g., Scheflen, 1972).

Strategy 2 models speakers that do occasionally point. It covers in particular situations where the intended referent is –F or +I. We assume that speakers decide on pointing *before* they choose the verbal means that will accompany the pointing act.

The strategy 2 does not tells us what to do when the referent +F and –I. Our data, however, tell us that speakers most of the time do not point in under those circumstances. This suggests that in such situations speakers might very well apply an algorithm akin to the one advocated by

In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). CTIT Workshop Proceedings.
I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds), University of Aberdeen, 25-26 January 2007. pp.129-139.
ISSN 1574-0846

Krahmer and Theune (2003). Normally, this will lead to a referring act which contains no pointing, but if the number of distractors is sufficiently large, pointing might be included. Alternatively, it might be that also for the +F and –I situation there are other factors that *a priori* determine whether the speaker will point or not.

Note that strategy 2 can be seen as an efficient heuristic for approximating the outcomes of the Krahmer & van der Sluis (K&VdS) algorithm, particularly for those situations where the intended referent is –F. The efficiency derives from the fact that for –F referents, we expect the cheapest referring act to require a pointing act most of the time (because there will be a significant number of distractors as result of the fact that the set of distractors is not limited to those that are in domain focus), and therefore only considering referring acts that contain a pointing act – as proposed by strategy 2 – limits the search space.

In another respect, the K&VdS method does, however, differ from strategy 2: the K&VdS method does not take importance of the intended referent into account.

Finally, we would like to argue that there is a further consideration for the inclusion of pointing to –F referents (as proposed in strategy 2), that is not covered by the K&VdS method. Even if pointing is not the cheapest option in K&VdS's sense (when compared to verbal means of referring), the *attention directing* function of pointing (see, e.g., Butterworth, 2003:9) might lead to a preference for pointing. The use of pointing when the intended referent is –F, is more than simply a way of identifying the object, it also serves to *directly guide* the addressee's gaze to the referent. This is something which cannot be achieved by verbal means only (where the addressee will need to *interpret* the referring expression and then decide where to direct his or her gaze). In particular, if the addressee's gaze is in a region that is distant from the intended referent, pointing can be more adequate than verbal identification, because it allows the speaker to directly manipulate the addressee's gaze.[5]

5 Conclusions

In this paper, we have outlined our arguments against two predominant assumptions regarding modality selection in the generation of referring acts: the assumption that non-verbal means of referring are secondary to verbal ones, and the assumption that there is a single strategy that speakers follow for generating referring acts. To some extent, we have reversed these assumptions by arguing that (A) the choice regarding whether or not to point precedes the choice of verbal means of reference and (B) offering as a hypothesis for further investigation the claim that speakers might differ (irrespectively of situational factors) with regards to their inclination to refer by pointing.

Our supporting evidence was drawn from data obtained through an observational study. We specified two alternative strategies for modality selection based on correlation data from the observational study. Speakers that follow the first strategy simply abstain from pointing. Speakers that follow the other strategy make the decision whether to point dependent on whether the intended referent is in focus and/or important. Further controlled experiments are needed to verify whether the decision to point is effected by (rather than merely correlated with) whether the intended referent is in focus or important.

Let us conclude by identifying two further topics for research. These concern the precise realization of pointing acts and verbal expressions in multimodal referring acts. Firstly, on the basis of a case study of deictic gestures by Neapolitan speakers, Kendon & Versante (2003:134) suggest that 'the character of the pointing gesture itself might vary systematically in relation to semantic distinctions of various sorts'. Secondly, the choice of determiner also requires further study; most quantitative studies of the use of definite article, proximal and distal demonstrative concern their use in text, see e.g., Kirsner and Van Heuven (1988) and Maes & Noordman (1995). In our corpus, all these types of referring expression occur. Demonstratives of both the proximate and distal variety are abundant in our corpus (see Piwek & Cremers, 1996) and, interestingly, we also observed 11 instances where an indefinite noun phrase (e.g., 'Now there is, where you just were, a red square on the floor.') was used in a referring act.

[5] In other words, pointing is a form of indicating in the sense of C.S. Peirce; see Buchler (1940: chap. 7).

In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). CTIT Workshop Proceedings.
I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds), University of Aberdeen, 25-26 January 2007. pp.129-139.
ISSN 1574-0846

References

Bates, J. (1994). The role of emotion in believable agents. In: *Communications of the ACM* **37**(7), 122–125.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science* **15**(6), 415–419.

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

Buchler, J. (1940). *The Philosophy of Peirce: Selected Writings*. Routledge and Kegan Paul, London.

Butterworth, G. (2003). Pointing is the Royal Road to Language for Babies. In: Kita, S. (Ed.), *Pointing: Where Language, Culture and Cognition Meet*, Lawrence Erlbaum, New Jersey, pp.9–34.

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

Claassen, W. (1992). Generating referring expressions in a multimodal environment, in: R. Dale et al. (eds.), *Aspects of Automated Natural Language Generation*, Springer Verlag, Berlin.

Clark, H. & Bangerter, A. (2004). Changing Ideas about Reference. In: I. Noveck & D. Sperber (Eds.), *Experimental Pragmatics* (pp. 25–49). Palgrave Macmillan, New York.

Cremers, A. (1993). Transcripties dialogen blokken-experiment. IPO report no. 889, Eindhoven.

Cremers, A. (1994). Referring in a shared workspace. In: Brouwer-Janse, M.D., Harrington, T.L., (Eds.), *Human-machine communication for educational systems design*, Springer Verlag, Heidelberg, 71–78.

Cremers, A. (1996). Reference to objects: an empirically based study of task-oriented dialogues. PhD thesis, Eindhoven University of Technology.

Dale, R. and E. Reiter (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions, *Cognitive Science* **18**:233-263.

Fitts, P. (1954). The information capacity of the human motor system in controlling amplitude of movement, *Journal of Experimental Psychology* **47**, 381-391.

Gergle, D. (2006). What's There to Talk About? A Multi-Modal Model of Referring Behavior in the Presence of Shared Visual Information. In: *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2006)* Student Research Workshop.

Grice, H.P. (1975). Logic and conversation. In: Cole, P. & J. Morgan (eds), *Syntax and Semantics 3: Speech Acts*, Academic Press, New York, 41–58.

Grosz, B. (1977). The representation and use of focus in dialogue understanding. *Technical note 151*, SRI International, Menlo Park.

Huls, C. & Bos, E. (1998). Studies into Full Integration of Language and Action. In: Bunt, H., Beun, R.J. & Borghuis, T. (1998). *Multimodal Human-Computer Communication; Systems, Techniques and Experiments*, Lecture Notes in Artificial Intelligence 1374, Springer, Berlin.

Kendon, A. & Versante, L. (2003). Pointing by hand in Neapolitan. In: S. Kita, (Ed.) *Pointing: Where Language Culture and Cognition Meet*. Lawrence Erlbaum, Hillsdale, N.J.

Khan, I., Ritchie, G., van Deemter, K. (2006). The clarity-brevity trade-off in generating referring expressions. In: *Proceedings of the Fourth International Natural Language Generation Conference (INLG)*, 14-15 July 2006. Sydney, Australia, 84-86.

In: Proceedings of the Workshop on Multimodal Output Generation (MOG 2007). CTIT Workshop Proceedings.
I. Van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds), University of Aberdeen, 25-26 January 2007. pp.129-139.
ISSN 1574-0846

Kirsner, R. and Van Heuven, V. (1988). The significance of demonstrative position in modern Dutch. *Lingua* **76**, 209–248.

Krahmer, E. and Van der Sluis, I. (2003), A New Model for Generating Multimodal Referring Expressions. In: *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003)*, April 12-13, Budapest Hungary, 47–54.

Kranstedt, A., Luecking, A., Pfeiffer, T., Rieser, H. and Staudacher, M. (2006). Measuring and Reconstructing Pointing in Visual Contexts. In: *Brandial 2006: The 10th Workshop on the Semantics and Pragmatics of Dialogue*, University of Potsdam, Germany, September 2006.

Kranstedt, A. & Wachsmuth, I. (2005). Incremental Generation of Multimodal Deixis Referring to Objects In: *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG 2005)*, Aberdeen, UK, August 2005, 75–82.

Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B. and Bhogal R. (1997). The persona effect: Affective impact of animated pedagogical agents, In: *Proceedings of CHI'97 (Human Factors in Computing Systems)*, 359–366, Atlanta.

Lester, J., J. Voerman, S. Towns and C. Callaway (1999). Deictic Believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents, *Applied Artificial Intelligence* **13**(4-5):383-414.

Maes, F. and Noordman, L. (1995). Demonstrative nominal anaphors: a case of nonidentificational markedness, *Linguistics* **33**, 255-282.

Neal, J. G., Thielman, C. Y., Dobes, Z., Haller, S. M., and Shapiro, S. C. (1989). Natural language with integrated deictic and graphic gestures. In: *Proceedings of the Workshop on Speech and Natural Language* (Cape Cod, Massachusetts, October 15 - 18, 1989). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 410–423

Piwek, P. & Cremers, A. (1996). Dutch and English Demonstratives: A Comparison. *Language Sciences*, **18**(3-4), 835-851.

Prendinger, H. & Ishizuka, M. (2004)(Eds). *Life-Like Characters: Tools, Affective Functions, and Applications*, Springer, Berlin.

Scheflen, A. (1972). *Body language and the social order: communication as behavioral control*. Prentice-Hall, Englewood Cliffs, N.J.

van der Sluis, I. and Krahmer, E. (2001). Generating Referring Expressions in a Multimodal Context: An empirically motivated approach. In: W. Daelemans et al. (eds.), *Selected Papers from the 11th CLIN Meeting*, Rodopi, Amsterdam.

Wilkins, D. (1999). Manual for the 1999 Field Season. Language and Cognition Group. Max Planck Institute for Psycholinguistics, Nijmegen.