Evaluating Multimodal NLG using Production Experiments

Ielka van der Sluis*, Emiel Krahmer†

*Computational Linguistics & AI
Faculty of Arts, Tilburg University
I.F.vdrSluis@uvt.nl

†Communication & Cognition
Faculty of Arts, Tilburg University
E.J.Krahmer@uvt.nl

Abstract

In this paper we report on an evaluation study for the generation of multimodal referring expressions. To test our algorithm, which allows for various gradations of preciseness in pointing, subjects performed an object identification task in a strict experimental setting. 20 subjects participated and were instructed to always use a pointing gesture (they were led to believe they were testing a new kind of 'digital pointing device'). The subjects performed their tasks on two distances: close (10 subjects) and at a distance of 2.5 meters (10 subjects). The assumption is that these conditions yield precise and imprecise pointing gestures respectively. In addition we varied the 'type' of target objects (geometrical figures versus pictures of persons). This study resulted in a corpus of 600 multimodal referring expressions. A statistical analysis (ANOVA) revealed a main effect of distance (subjects adapt their language to the kind of pointing gesture) and also a main effect of target (persons are more difficult to describe than objects). The advantages and disadvantages of this evaluation method are discussed.

1. Introduction

Because of the ongoing work on spoken dialogue systems, there is a substantial interest in Natural Language Generation (NLG) as a component of such systems. The generation of referring expressions is a central NLG task. A typical algorithm takes as input an object v ('the target') and a set of objects ('the distractors') from which the target object needs to be distinguished (borrowing terminology from Dale & Reiter 1995). The task of the algorithm is to determine which properties are needed to single out the target object from the distractors. This is known as the 'content determination' problem for referring expressions.

There are at least two motivations for the extension to multimodal referring expressions. First, in various situations a purely linguistic description may be too complex. In that case, including a deictic, pointing gesture may be the most efficient way of singling out the target referent. Second, due to the increased interest in Embodied Conversational Agents (ECAs), researchers have started exploring the possibilities of applying NLG to generate spoken language which an ECA can present. Typically, this implies the coordinated generation of language and gesture.

If we look at human communication it soon becomes apparent that referring expressions which include pointing gestures are rather common (Beun & Cremers, 1998). Various algorithms for the generation of multimodal referring expressions have been proposed (e.g., Claassen, 1992; Reithinger, 1992; Huls et al., 1995; Lester et al. 1999). Most of these are based on the assumption that a pointing gesture is precise and unambiguous. As soon as a pointing gesture is included, it directly eliminates the distractors and singles out the intended referent. As a consequence, the generated expressions tend to be relatively simple and usually contain no more than a head noun in combination with a pointing gesture. Moreover, most algorithms tend to be based on relatively simple, context-independent criteria for the decision whether a pointing gesture should be included or not. For instance, Claassen (1992) only generates a pointing gesture when referring to an object for which no distinguishing linguistic description can be produced. Lester et al. (1999), on the other hand, generate pointing gestures for all objects which cannot be referred to with a pronoun.

Recently, we developed an algorithm which differs from these earlier proposals in two ways. (We refer to Krahmer & van der Sluis 2003 for algorithmic and implementation details.) The basic assumption is that pointing should not always be precise and unambiguous. Rather we allow for various gradations of preciseness in pointing, ranging from unambiguous to vague pointing gestures. As illustrated in Figure 1, precise pointing (P) has a high precision. Its scope is restricted to the target object, and this directly rules out the distractors. But arguably, precise pointing is 'expensive'; the speaker has to make sure she points precisely to the target object in such a way that the hearer will be able to unambiguously interpret the referring expression. Imprecise pointing, on the other hand, has a lower precision –it generally includes some distractors in its scope– but is intuitively less 'expensive'. The model for pointing we propose may be likened to a 'flashlight'. If one holds a flashlight just above a surface, it will cover only a small area (the target object). Moving the flashlight away will enlarge the cone of light (shining on the target object but probably also on one or more distractors). A direct consequence of this 'flashlight model for pointing' is that we predict that the amount of linguistic properties required to generate a distinguishing multimodal referring expression is dependent on the kind of pointing gesture.

But how to evaluate our model? Evaluating content determination algorithms for natural language generation systems is known to be difficult. Corpora, for instance, which are often used for the evaluation of other natural language

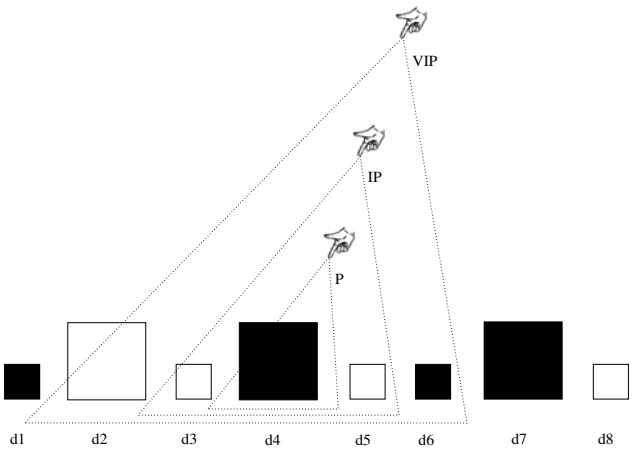

Figure 1: The distractors in the scope of a pointing act depend on the precision of pointing (Precise, ImPrecise or VeryImPrecise).

		DISTANCE	
		near	far
TARGET	object	I	II
	person	III	IV

Table 1: Overview of the experimental design with DISTANCE as between subjects and TARGET as within subjects variables.

processing applications, are not straightforwardly applicable to the evaluation of content determination algorithms, since we typically do not have access to the underlying semantic representations. Adding additional modalities, as we do here, only leads to further complications.

In this paper we propose to use production experiments for the evaluation of multimodal NLG algorithms. In such experiments, subjects are offered stimuli which they have to verbalize. It can then be checked whether the algorithm's verbalizations coincide with those of the subjects on the dimension under investigation. As a case in point, we describe a simple experiment that addresses one of the crucial ingredients of the algorithm: the claim that the linguistic part of a multimodal referring expression depends on the kind of pointing gesture. It seems likely that imprecise pointing requires more linguistic material to single out the target object, but we also would like to know in more detail *what* kind of material is used. Moreover, it might be that the *kind* of target object plays a role in this. In the experiment, we control for these factors.

2. An Experiment

2.1. General overview

A production experiment was performed to elicit multimodal referring expressions. Subjects had to perform an object identification task, in which they were first shown an isolated object which they subsequently had to single out among a set of comparable objects. Two sorts of target objects (geometrical figures and photos of famous mathematicians) were used to determine whether the kind of target influenced the results. Half of the subjects performed the tasks at a close distance (they could touch the target object directly), the other half of the subjects performed the same tasks from a small distance (and could only indicate the location in which the target appeared). The experiment has a two by two design, with TARGET as a within subjects variable and DISTANCE as between subjects variable. Table 1 summarizes the experimental design.

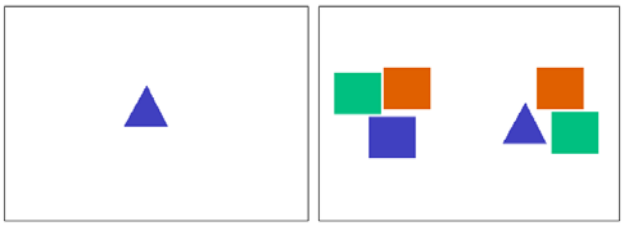

Figure 2: An example of the kind of stimuli used in the experiment. First, the target object (a geometrical object) is displayed in isolation (left). Subsequently it is presented together with a number of similar objects (right).

2.2. Method

Subjects Twenty native speakers of Dutch participated as subjects. All are students and colleagues at Tilburg University. None was familiar with the multimodal generation algorithm being tested. For each condition, the group of subjects consisted of five men and five women.

Experimental setting Subjects were led to believe they were testing a new computer system which could be operated by the combined usage of speech and gesture. They were told the system was in its testing phase; their input was required for calibration purposes. To evoke pointing gestures, the subjects were given a pen-like 'digital stick' (a pen mouse) of approximately 10 centimeters as their pointing device. They were told that the digital stick emitted a signal which the computer could detect and interpret. In addition, subjects were equipped with a headset including a microphone through which they could speak to the computer.

Their task was to identify a target object via speech and gesture. Each target object was first displayed in isolation on a 17 inch screen, after which the target object was presented among a set of distractors and the subject had to single it out. No feedback was given to subjects by the experimenter or the computer, to avoid influencing the subjects in their realizations. Half of the subjects performed the experiment in the 'near' condition; they were placed directly in front of the screen and could touch the target object with the stick (*precise pointing*). The other half of the subjects, those in the 'far' condition, were placed on approximately 2.5 meters from the screen. By definition their pointing acts were always *imprecise*.

Stimuli Two kinds of target objects were used in the experiment: (1) 15 two-dimensional geometrical objects and (2) 15 black and white photographs of persons (all famous mathematicians). The geometrical figures vary in shape (cube, circle, triangle) and color (red, blue, green). The persons display a greater variety: some are male, some female, they may wear hats, glasses, moustaches and/or

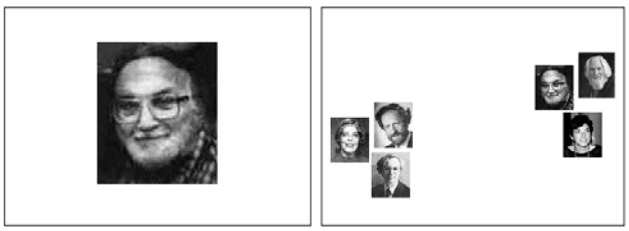

Figure 3: A second example of the kind of stimuli used in the experiment. First, the target object (a picture of a mathematician) is displayed in isolation (left). Subsequently it is presented together with a number of similar objects (right).

beards (only the men), and they may have long, short, grey or no hair.

The 30 target objects were presented to subjects in a randomized order. For the identification task, the target object was presented on a computer screen together with a number of other objects from the same domain. To facilitate pointing the objects were presented on the screen in two isolated groups of 2 or 3 objects, one containing the target (*the target group*), while the other group solely consisted of distractors (*the distractor group*). The position of the target group on the screen is systematically varied, as is the position of the target object within the target group. Figures 2 and 3 illustrate the stimuli for objects and persons respectively.

2.3. Data processing

The subjects were filmed during the experiment. The resulting data consist of 600 multimodal referring expressions (20 subjects $\times$ 30 stimuli). All utterances were transcribed and annotated. The kind of pointing gesture was classified, and the kinds of linguistic properties were determined and counted. All subjects produced a 'correct' (i.e., distinguishing) description for each target object. Below, all tests for statistical significance were done using an analysis of variance (ANOVA) with repeated measures.

2.4. Results

As intended, all subjects always used a pointing gestures. In the near condition, this pointing gesture was always a precise one, where the target object was directly touched with the pointing device. In the far condition subjects by definition employed imprecise pointing gestures, which basically denoted in which of the two groups of objects on the screen the target object is located. This indicates that the operationalization of (im)precise pointing worked as planned, and that we can test the hypothesis that the kind of pointing gesture influences linguistic realization. No gender differences were found, so we present combined results for male and female subjects.

As a first approximation, we looked at the number of words in the multimodal referring expressions as a function of the distance and the target. The results can be found in Table 2. There is a main effect of distance ($F(1,18) = 45.45, p < .001$), which indicates that in the far condition subjects use more words than in the near condition. In addition, there is a main effect of target ($F(1,18) = 53.99, p < .001$); this implies that subjects require more words to refer to the persons than to the objects. In addition, there is an interaction between distance and target ($F(1,18) = 49.09, p < .001$). This can be explained by observing that the effect of distance is stronger for persons than for objects in the far condition but not in the near condition.

		DISTANCE	
		near	far
TARGET	object	0.78 (1.21)	2.93 (0.87)
	person	0.84 (1.31)	5.45 (1.32)

Table 2: Average number of words per description as a function of distance and target. Standard deviations between brackets.

Table 3 presents a more detailed analysis of the linguistic material, making a distinction between *type* information (whether the target is a cube, a circle, person, etc., i.e., the information given in the head noun), the number of prenominal properties (*prop*, e.g., color, hair style, etc.) and the number of location markers (*loc*, e.g., left, below, etc.) Looking at the presence of type information, a main effect of distance is found ($F(1,18) = 144.6, p < .001$); no effect of target and no interaction either (in both cases $F(1,18) < 1$). That is: when subjects use a precise pointing gesture in this experiment they do not use type information, but when they use an imprecise pointing gesture, they *do* include type information (sometimes even twice, explaining the 1.01 for persons). For adjectival properties, both a main effect of distance is found ($F(1,18) = 70.01, p < .001$), and a main effect of target ($F(1,18) = 10.31, p < .01$). No interaction is found. In terms of the figures in Table 3: when subjects use a precise pointing gesture, they do not use adjectival properties, and when they use an imprecise pointing gesture they do. And in addition, when subjects describe an object they are somewhat more likely to use a prenominal adjective than when describing a person. For locations, finally, a main effect of distance is found ($F(1,18) = 2.02, p < .05$), and a main effect of target ($F(1,18) = 20.47, p < .001$). There is also an interaction between target and distance ($F(1,18) = 16.62, p < .01$). Inspection of the table reveals that these effects can be explained by the fact that location information is rare when a precise pointing act is used, but relatively common when describing a person in combination with an imprecise pointing gesture.

2.5. Discussion

The experimental results indicate that speakers indeed vary the linguistic part of a multimodal referring expression depending on distance, in that the amount of linguistic material co-varies with the kind of pointing gesture. In the near condition, eight out of ten speakers always produced multimodal referring expressions containing a demonstrative determiner ("deze", *this*) or no spoken material at all. The remaining two consistently added a head noun ("deze driehoek", *this triangle*). When, on the other hand, an imprecise pointing act is used (because of the distance to the target), the referring expressions contain much more spoken material. The kind of target object also had an influ-

			DISTANCE	
			near	far
		type	0.15 (0.32)	1.00 (0.00)
	object	prop	0.19 (0.34)	0.94 (0.13)
		loc	0.09 (0.27)	0.30 (0.43)
TARGET				
		type	0.11 (0.17)	1.01 (0.04)
	person	prop	0.03 (0.11)	0.76 (0.26)
		loc	0.12 (0.33)	0.81 (0.45)

Table 3: Average numbers of attributes given per description as a function of distance and target. The variables type, prop and loc are explained in the text. Standard deviations between brackets.

ence on this. In general, fewer words are required to single out a geometrical figure than to identify person. Closer inspection of the data reveals that both objects and persons are described in terms of their type (e.g., *triangle* and *man* respectively). In addition, objects are more often accompanied by prenominal adjectives (*blue*), while person descriptions tend to include locative expressions (*in the top left corner*). This is probably due to the fact that describing persons is inherently more difficult than describing colored geometrical objects, since the number of potentially relevant attributes is much larger for persons than for objects.

Two other things are noteworthy. First, there are some clear differences between speakers. In the 'close' condition, for instance, most speakers reduce the linguistic material in their referring expressions almost to zero. Second, in the 'far' but not in the 'close' condition, subjects tend to produce more overspecified descriptions (in line with earlier work by, for instance, Dale & Reiter (1995)). One possible explanation is that this is due to the inherent uncertainty of imprecise pointing. Speakers may not be sure whether the imprecise pointing act is sufficiently clear and to guarantee that their reference will be distinguishing they include additional properties.

Note that our algorithm is in agreement with the majority of subjects concerning the first point, but makes different prediction when it comes to the second point. This is due to the fact that the search strategy used in our algorithm is aimed at detecting minimal descriptions. It is worth stressing, though, that different search strategies are compatible with the graph-based perspective, and Krahmer et al. (2003) illustrate this by describing a different search strategy which mimics the Incremental Algorithm by Dale & Reiter (1995) and thus gives rise to a certain amount of redundancy.

3. General discussion

We have described a straightforward evaluation experiment using production data; subjects generate distinguishing descriptions for selected target objects, and the resulting descriptions are analyzed and compared with the predictions made by the algorithm. In this way, we can use spontaneous data (subjects were not told what to say), while at same time 'controlling' the input representations (the target and its properties are known). A potential disadvantage is that different aspects of an algorithm may require different experiments, and in addition that performing these experiments tends to be a time consuming process.

Naturally, the experiment described here is a first step in the direction of a full-fledged experimental evaluation. A disadvantage of the first study is that subjects were forced to point, and that the size of target object was kept constant. Therefore we conducted a second study in which subjects performed a topographical task in a more natural and interactive setting. 20 subjects (different from those of the current study) participated and were asked to locate countries on a world map. Again the subjects performed their tasks on two distances: close (10 subjects) and at a distance of 2.5 meters (10 subjects). The target objects in this study were selected in such a way that there is a distinction between the objects that are 'easy to locate' (large or isolated) and the objects that are 'difficult to locate'. The second study resulted in another corpus of 600 multimodal referring expressions. Analysis showed a confirmation of the results of the first study, plus various additional findings concerning the role of spatial relations, linguistic history and kinds of pointing gestures (dynamic, circling, static, etc.). Unfortunately, lack of space prohibits us from giving more details here, but we hope to report on these in a sequel to this paper.

Acknowledgements Thanks are due to Harry Bunt for comments on a previous version. Krahmer's work was partly done within the context of the TUNA project, funded by Engineering and Physical Sciences Research Council (EPSRC) in the UK, under grant reference GR/S13330/01 and partly within the IMOGEN project funded by the Netherlands Organization for Scientific Research (NWO).

4. References

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

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

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

Huls, C., E. Bos & W. Claassen (1995), Automatic referent resolution of deictic and anaphoric expressions, *Computational Linguistics* **21**(1):59-79.

Krahmer, E., S. van Erk & A. Verleg (2003), Graph-based Generation of Referring Expressions, *Computational Linguistics*, **29**(1):53-72.

Krahmer, E. & I. van der Sluis (2003), A New Model for Generating Multimodal Referring Expressions. Proceedings of the ENLG'03, Budapest, Hungary, pp.47-54.

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

Pechmann, T. (1989), Incremental speech production and referential overspecification, *Linguistics* **27**:98-110.

Reithinger, N. (1992), The performance of an incremental generation component for multi-modal dialog contributions, in: *Aspects of Automated Natural Language Generation*, R. Dale et al.(eds.), Springer Verlag, Berlin.