The Influence of Target Size and Distance on the Production of Speech and Gesture in Multimodal Referring Expressions

Ielka van der Sluis

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

Emiel Krahmer

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

Abstract

In this paper we report on a production experiment for multimodal referring expressions. Subjects performed an object identification task in an interactive setting. 20 subjects participated and were asked if they could identify 30 countries on a world map on the wall. 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 'size' of target objects (large or isolated objects versus small objects). 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 (smaller objects are more difficult to describe than large or isolated objects).

1. Introduction

In recent years there has been an increased interest in multimodal systems that allow combinations of speech and gesture, both on the input and the output side (see e.g., Benoit et al. 2000). Often this kind of multimodality is employed to enhance focussing on some specific target object (Clark 2003). On the input side, a number of multimodal systems allow the user to single out a target object in a visual interface using gestures ('touch pointing') accompanied with speech (as in the Smartkom system, for instance see,Whalster 2002). Examples of multimodal systems that combine gesture and speech on the output are *embodied conversational agents* (Cassel et al. 2000).

The combined usage of speech and gesture puts new constraints on the interpretation and generation modules in multimodal spoken dialogue systems. Oviatt (1999) points out, for instance, that the spoken part of multimodal language tends to be simpler than unimodal language. The downside is that the spoken information needs to be synchronized and 'fused' with gestural information. For data-driven development and testing of multimodal interpretation and generation modules it is important to collect data on how humans produce multimodal referring expressions combining speech and gesture (e.g., Piwek & Beun 2001, Kranstedt et al. 2003).

Arguably, two factors that influence object identification are target size (some targets are easier to point to than others), and target distance (an object that is close is easier to point at than an object that is further away). Interestingly, both these factors are combined and weighted in Fitts' law, an empiricial measure of the difficulty people have in reaching a target (Fitts 1954, see also Krahmer & Van der Sluis 2003).

This raises a number of questions, for example: (1) What is the influence of target size and distance on the decision to point? One would expect that people use more gestures when referring to 'easily' reachable targets (large and/or close ones), (2) What is the influence of the pointing gesture on the produced spoken language? and (3) In what way and how much do relata (Levelt 1989) occur in the referring expressions? Especially in describing 'difficult' targets, 'easily' recognizable relata could be helpful in identifying the target. To address these questions, we performed a production study in which the target size and the distance to the target were systematically varied. Two groups of subjects, one close to and one further away from a world map, had to single out countries with variable sizes. In this paper we present a statistical analysis of the resulting 600 multimodal referring expressions. In section 2 we describe the experiment we conducted, we give a general overview, and discuss subjects, experimental setting, materials and data processing. In Section 3 we present the results of our experiment: the interaction of language and speech, an analysis of the linguistic material and the gestures. We finish with a discussion in Section 4.

		DISTANCE	
		near	far
TARGET	easy	I	II
	difficult	III	IV

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

2. Method

2.1. General overview

We conducted a production experiment to elicit multimodal referring expressions. Subjects had to perform an object identification task, in which they had to identify countries on a world map. The size of the world map is 100 by 140 cm. The target countries to be identified varied in difficulty. Half of the subjects performed the tasks at a close distance (they could touch the target country directly), the other half of the subjects performed the same tasks from a distance (and could only roughly indicate the location of the target). 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.

2.2. Subjects

Twenty native speakers of Dutch participated as subjects. All are students and colleagues at Tilburg University. For each condition, the group of subjects consisted of five men and five women.

2.3. Experimental Setting

Subjects were told that their topographical knowledge was going to be tested just like in primary school. Half of the subjects performed the experiment in the ‘near’ condition; they were placed directly in front of the world map and could touch the target (*precise pointing*). The other half of the subjects, those in the ‘far’ condition, were placed on approximately 2.5 meters from the map. Subjects in the far condition could only roughly point in the direction of the target. By definition their pointing acts were always *imprecise*. In Figure 1 of each condition an example is shown. Subjects were given a stick of 40 cm they could use for pointing if they so desired. Although the subjects used different strategies to identify targets, all subjects were consistent in their behavior during the task. Subjects were asked to be more specific with unclear references.

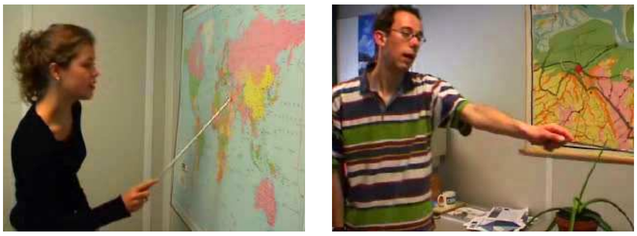

Figure 1: Example of a subjects in the near condition(left) and in the far condition (right)

2.4. Materials

We selected 30 ‘easy to find’ countries that can be divided in two kinds of target objects: 15 relatively small countries and 15 relatively large or isolated countries. Isolated countries, like islands or groups of islands, stand out because of their shape or color and we considered them to be as easy identifiable as the larger countries. The relatively small countries, like for example Italy, we call ‘difficult’ targets, because they cannot be distinguished with an imprecise pointing gesture and their description requires some effort. We call the large or isolated countries, for example Russia or New Zealand, ‘easy’ targets because they can be identified with an imprecise pointing gesture that reduces the need for linguistic information. Except for the variability in size, the countries also differ in shape and color. The 30 target objects were presented to the subjects in a random order.

2.5. Data Processing

Subjects were filmed during the experiment. The resulting data consist of 600 multimodal referring expressions (20 subjects × 30 stimuli). All utterances were transcribed and annotated. The pointing gestures were classified, and the kinds of linguistic attributes were determined and counted. All subjects produced a ‘correct’ (i.e., distinguishing) description for each target. All tests for statistical significance were done using an analysis of variance (ANOVA) with repeated measures, and with distance as between subjects variable and target as within subjects variable.

3. Results

3.1. Interaction of Language and Speech

Without being forced, all subjects always used a pointing gesture. In the near condition, this pointing gesture was always a precise one, where the target was directly touched. In the far condition subjects by definition employed imprecise pointing gestures, which basically denote in what area on the map the target is located. This indicates that the variation in distance worked as intended.

			DISTANCE	
			near	far
TARGET	easy	**words**	2.28(1.09)	15.59(3.10)
		disfl	.19(.10)	1.57(.85)
	difficult	**words**	3.23(1.63)	27.25(6.28)
		disfl	.17(.11)	2.40(.65)

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

As a first approximation, we looked at the number of words and the number of disfluencies in the multimodal referring expressions, Table 2. For the number of words there is an effect of distance ($F(1, 18) = 241.04, p < .01$), and an effect of target ($F(1, 18) = 33.12, p < .01$). These effects indicate that in the far condition subjects use more words than in the near condition and subjects require more words to refer to difficult objects than to easy ones. In addition, there is an interaction between distance and target ($F(1, 18) = 23.93, p < .01$). This can be explained by observing that the effect of target is stronger in the far condition than in the near condition. As disfluencies we count the number of repairs, repetitions, pauses and filled pauses. The number of disfluencies show an effect of distance ($F(1, 18) = 100.44, p < .01$) and an effect of target ($F(1, 18) = 6.44, p < .05$), which indicate that both in the far condition and when referring to difficult objects subjects speak less fluently. Furthermore there is an interaction between distance and target ($F(1, 18) = 7.17, p < .05$) which signals a stronger effect of target in the far condition compared to the near condition. In the near condition subjects do not use many words to refer to objects easy or difficult, consequently disfluencies are scarce.

3.2. Analysis of Linguistic Material

Thus, subjects appear to adapt their linguistic material to the kind of pointing gesture they use. Although we observed some differences among the subjects, especially in the far condition, each of the subjects displayed consistent behavior throughout the experiment. In the far condition, all subjects used imprecise pointing gestures, and hence were required to use more additional linguistic material to produce an unambiguous reference. For example a typical description of an easy object like Brasil is *dat grote groene vlak ↗ daar* (*that large green area ↗ over there*) together with an imprecise pointing gesture. As an example of a difficult object consider a description of Portugal: *Portugal ehm is het eh groene land dat ten zuid westen of dat eh in zuid europa ligt ↗ naast het roze spanje* (*Portugal uhm is the uh green country which lies on the south west or which uh lies in southern Europe ↗ next to the pinkish Spain*) together with an imprecise pointing gesture. In the near condition, a precise pointing gesture suffices to single out the target. Additionally, the name of the target is sometimes mentioned together with a *here* or a *there*.

Table 3 presents a more detailed analysis of the linguistic material, making a distinction between *name* whether the name of the target is mentioned (like *Portugal* in the example above), *type* information (whether the target is called a country, area, isle, spot, part etc., i.e., the information typically given in the head noun), the number of prenominal properties (*property*, e.g., color, size, shape, etc.) and the number of location markers (*location*). Location markers can be split into at least two types: (1) *in het zuiden* (*in the south*), as a general reference to southern part of the world, and (2) *naast het roze Spanje* (*next to the pinkish Spain*) including a relatum. In

			DISTANCE	
			near	far
TARGET	easy	**name**	.32(.26)	.84(.24)
		type	.03(.05)	.92(.29)
		property	.04(.06)	1.51(.20)
		location	.12(.13)	.18(.68)
		relata	.05(.08)	1.11(.46)
		-name	.03(.06)	.81(.60)
		-type	.00(.00)	.40(.32)
		-property	.00(.00)	.28(.46)
		-location	.03(.06)	1.35(.65)
	difficult	**name**	.33(.28)	1.07(.18)
		type	.04(.07)	.76(.16)
		property	.07(.08)	1.30(.21)
		location	.13(.18)	2.87(.98)
		relata	.11(.15)	2.21(.56)
		-name	.03(.06)	2.01(.95)
		-type	.03(.06)	.78(.45)
		-property	.00(.00)	.81(.34)
		-location	.09(.13)	2.72(.85)

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

the latter case *next to the pinkish Spain* as a whole is treated as a location marker. In addition we counted the number of *relata* used per description. In the example of Portugal the number of relata is two: *Europe* and *Spain*. The descriptions that identify relata, for example *het roze Spanje* (*the pinkish Spain*) are dealt with separately. In Table 3 we also present *name*, *type*, *property*, *location* used, for all relata used in all descriptions.

First consider the between subject effects, the near versus the far condition. The results show that for almost all features there is a significant effect of distance (*name*, $F(1,18) = 41.21, p < .01$; *type*, $F(1,18) = 132.21, p < .01$; *property*, $F(1,18) = 554.75, p < .01$; *location*, $F(1,18) = 76.57, p < .01$; *relata*, $F(1,18) = 119.787, p < .01$). Thus, in the far condition, speakers use more names, more *type*, *property* and *location* information and more *relata* to identify a target object.

Looking at the within subject effects, difficult versus easy objects. The results show that subjects tend to use more type and property information when referring to large objects (*type*, $F(1,18) = 5.96, p < .05$ and *property*, $F(1,18) = 5.94, p < .05$). Whereas in descriptions for difficult objects subjects use more *names*, *locations* and *relata* (*name*, $F(1,18) = 5.03, p < .05$; *location*, $F(1,18) = 27.72, p < .01$; *relata*, $F(1,18) = 51.157, p < .01$)

When we compare the references for easy objects with those of difficult objects, it can be noted that the differences are almost nonexistent in the near condition, while they are substantial in the far condition (*name*, $F(1,18) = 4.91, p < .05$; *type*, $F(1,18) = 6.99, p < .05$; *property*, $F(1,18) = 9.53, p < .05$; *location*, $F(1,18) = 27.24, p < .01$; *relata*, $F(1,18) = 41.149, p < .01$). Interestingly, in the far condition, easy objects are more often referred to using head nouns and properties, while descriptions of difficult objects tend to contain more locative expressions and relata.

In our separate analysis of relata, there are effects of distance for all attributes, (*name*, $F(1,18) = 33.964, p < .01$; *type*, $F(1,18) = 31.398, p < .01$; *property*, $F(1,18) = 23.139, p < .01$; *location*, $F(1,18) = 75.887, p < .01$) which can be explained by the fact that relata almost exclusively occur in the far condition. Moreover, all attributes used to describe relata display effects of target in the sense that in descriptions of easy objects all these attributes are used less compared to their occurrences in references to difficult objects (*name*, $F(1,18) = 50.562, p < .01$; *type*, ($F(1,18) = 8.491, p < .01$; *property*, $F(1,18) = 18.656, p < .01$; *location* $F(1,18) = 66.822, p < .01$). Comparing the near and the far condition the effects of target for the attributes used to describe relata are large (*name*, $F(1,18) = 49.450, p < .01$; *type*, $F(1,18) = 5.961, p < .05$; *property* $F(1,18) = 18.656, p < .01$; *location* $F(1,18) = 57.136, p < .01$). Hence, in the far condition subjects tend to use more attributes to describe relata of difficult objects, in comparison to the number of attributes used in describing relata of easy objects.

3.3. Analysis of Gestures

In Table 4 an analysis of the occurrences of gestures made during the references is presented. The total number of pointing gestures (*total pointing*) is split into pointing to the target (*to target*) and pointing *to relata*. Furthermore, we classified the kind of pointing subjects applied into *dynamic* and *static* gestures. Dynamic pointing gestures can be defined as gestures that include some kind of movement: *vertical*, *horizontal* or *circling*. Contrastively, static pointing gestures display no movement in the stroke of the pointing gesture. In general, Table 4 shows that all subjects pointed at every target at least once, no matter the distance or size. Although it is hard to distinguish difficult objects with imprecise pointing gestures, surprisingly, subjects in the far condition tend to point even more often (almost twice) to difficult objects. When we consider the number of *total pointing gestures* in more detail, it appears that subjects within the far condition, direct considerably more pointing gestures *to relata* in describing difficult objects than in describing easy objects. Apart from the distribution of pointing gestures, we looked at kinds of precise and imprecise pointing gestures. Most precise pointing gestures are of a static nature, whereas the imprecise pointing gestures display a greater variability: between static and dynamic gestures and also within the dynamic gestures.

More specifically, the total number of pointing gestures displays both an effect of distance ($F(1,18) = 24.52, p < .01$) and an effect of target ($F(1,18) = 13.45, p < .01$), which indicate that subjects in the far condition use more pointing gestures especially in references to difficult objects. Moreover the interaction between target and distance ($F(1,18) = 11.62, p < .01$) displays a difference in the effect of target. In all conditions around one pointing gesture is used to indicate the target, accordingly there are no effects of target or distance. In contrast pointing gestures that indicate a relatum display effects both of target ($F(1,18) = 14.17, p < .01$) and distance ($F(1,18) = 19.44, p < .01$). In the near condition there are no such pointing gestures because relata usually do not occur. In the far condition, except for pointing at the target, subjects also use pointing gestures to indicate relata, especially when the target is difficult to describe. The interaction between target and distance ($F(1,18) = 14.17, p < .01$) signals a difference in target effect. The type of pointing gestures used in the near condition is in almost all cases static. Whereas in the far condition the type of pointing gestures varies, dynamic pointing gestures are almost used as often as static ones. The static pointing gestures only display an effect of target ($F(1,18) = 33.82, p < .01$), which indicates that subjects tend to use more static gestures to identify difficult objects. There is no effect of distance, but there is an interaction ($F(1,18) = 17.10, p < .01$), which implies that the effect of target differs significantly between the far and the near condition. Dynamic gestures only display an effect of distance ($F(1,18) = 17.15, p < .01$), subjects use more dynamic gestures in the far condition. The effects of distance are only present for horizontal and vertical point-

			DISTANCE	
			near	far
TARGET	easy	**total**	1.00(.00)	1.32(.22)
		to target	1.00(.00)	1.13(.14)
		to relata	.00(.00)	.20(.17)
		static	.80(.32)	.73(.32)
		dynamic	.15(.31)	.90(.54)
		-vert	.00(.00)	.36(.18)
		-hor	.02(.03)	.15(.13)
		-circ	.13(.32)	.39(.34)
	difficult	**total**	1.02(.04)	1.86(.59)
		to target	1.02(.05)	1.03(.21)
		to relata	.00(.00)	.85(.64)
		static	.91(.25)	1.36(.41)
		dynamic	.12(.27)	1.03(.74)
		-vert	.02(.03)	.53(.29)
		-hor	.00(.00)	.18(.12)
		-circ	.10(.27)	.32(.48)

Table 4: Average numbers of pointing gestures given per description as a function of distance and target. The variables total, to target, to relata, static, dynamic, vert, hor and circ are explained in the text. Standard deviations between brackets.

ing gestures (vertical: $F(1, 18) = 48.710, p < .01$ and horizontal: $F(1, 18) = 29.689, p < .01$).

4. Discussion

We have described a production experiment conducted in a natural, interactive setting where subjects produce distinguishing descriptions for selected target objects. The experimental results, contrary to our expectation, indicate that speakers always include pointing gestures in their descriptions regardless of the difficulty of the target and the distance to the target. This could be a result of the fact that we equipped the subjects with a stick with which they could point, or simply because the task itself provokes pointing gestures. When the target is close, speakers tend to point only once in the direction of the target in a static fashion. When the target is located at a larger distance, the variability in the kind of pointing gestures increases. Surprisingly, speakers use more pointing gestures to refer to difficult targets in contrast to easy targets. We did not expect this, since Fitt's law (Fitts 1954) would suggest the opposite. A closer inspection of the data shows that the extra pointing gestures are directed to relata and not the target. Furthermore speakers vary the linguistic part of a multimodal referring expression depending on the distance and the kind of pointing gesture they apply. When the target is close, speakers reduce the linguistic material to almost zero, whereas subjects tend to produce overspecified descriptions if the target is located further away (in line with earlier work by, for instance, Pechmann 1989). This can be 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 information. As expected descriptions of difficult targets, often contain less fluent speech (more uhs/ums), because more speakers effort is required (see e.g., Goldman-Eisler 1968, Clark and Fox-Tree 2002). Typically the features of difficult targets are harder to recognize at a distance and there is a tendency to include descriptions of relata to indicate the location of the target. In describing difficult targets, speakers include the name of the target together with at least one property and almost three location markers. In contrast, descriptions of easy targets, generally include a head noun and one or two adjectival properties. As future work we will adjust our algorithm for the generation of multimodal referring expressions (Krahmer and Van der Sluis 2003) to the results of the empirical findings reported on in 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).

5. References

[1] Benoit, C. Martin, J.-C. Pelachaud, C. Schomaker, L. Suhm, B. (2000), Audio-Visual and Multimodal Speech Systems, in: *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, I. Mertens, R. Moore (eds.), Kluwer Academic Publishers.

[2] Cassell, J. Sullivan, J. Prevost, S. Churchill, E. (eds.): Embodied Conversational Agents. Cambridge, MA: MIT Press, 2000.

[3] Clark, H. (2003), Pointing and Placing, in *Pointing, Where Language, Culture, and Cognition Meet*, S.Kita (eds.), Lawrence Erlbaum Associates, Publishers.

[4] Clark, H. Fox Tree, J. (2002), Using uh and um in spontaneous speaking. Cognition, 84, 73-111.

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

[6] Goldman-Eisler, F. (1968): Psycholinguistics: Experiments in Spontaneous Speech, London, Academic Press.

[7] Krahmer, E. Van der Sluis, I. (2003), A New Model for Generating Multimodal Referring Expressions. Proceedings of the 9th ENLG, Budapest Hungary, pp.47- 54.

[8] Kranstedt, A. , Kuhnlein, P. & Wachsmuth, I. (2003), Deixis in Multimodal Human Computer Interaction: An Interdisciplinary Approach, *Proceedings of the 5th Gesture Workshop*, Genova, Italy, pp.112-123.

[9] Levelt, W. (1989), *Speaking, from Intention to Articulation*, MIT Press, Cambridge, MA.

[10] Oviatt, S. (1999),Ten Myths of Multimodal Interaction, Communications of the ACM, 42, 74-81, 1999.

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

[12] Piwek, P. & Beun, R. J. (2001), Multimodal Referential Acts in a Dialogue Game: From Empirical Investigations to Algorithms, *Proceedings of the IPNMD-2001*, Verona, Italy, pp.127-131.

[13] Van der Sluis, I.& Krahmer, E.(2004), Evaluating Multimodal NLG using Production Experiments, *Proceedings of the LREC-2004*, Lisbon, Portugal.

[14] Wahlster, W. (2002), SmartKom: Fusion and Fission of Speech, Gestures, and Facial Expressions *Proceedings of the 1st International Workshop on Man-Machine Symbiotic Systems*, Kyoto, Japan, pp. 213-225.