Need I Say More? On Factors Causing Referential Overspecification

Ruud Koolen (R.M.F.Koolen@uvt.nl)
Tilburg University, Department of Communication and Information Sciences
P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

Albert Gatt (A.Gatt@abdn.ac.uk)
University of Aberdeen, Department of Computing Science
Meston Building, King's College, AB24 3UE, Scotland, United Kingdom

Martijn Goudbeek (M.B.Goudbeek@uvt.nl)
Tilburg University, Department of Communication and Information Sciences
P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

Emiel Krahmer (E.J.Krahmer@uvt.nl)
Tilburg University, Department of Communication and Information Sciences
P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

Abstract

We present the results of an elicitation experiment conducted to investigate which factors cause speakers to overspecify their referential expressions, where we hypothesized properties of the target and properties of the communicative setting to play a role. The results of this experiment show that speakers tend to provide more information when referring to a target in a more complex domain and when referring to plural targets. Moreover, written and spoken referring expressions do not differ in terms of redundancy, but do differ in terms of the number of words that they contain: speakers need more words to provide the same information as people who type their expressions.

Keywords: Referential overspecification; GRE algorithms; Referring expressions.

Introduction

Referring expressions (expressions such as 'the green chair') are ubiquitous in human language production. Many studies in several fields of linguistics have investigated how people refer. Various aspects of referring expressions have been addressed by studies in both psycholinguistics and computational linguistics.

In recent computational linguistic studies on referring expressions, the focus has been on creating algorithms for the Generation of Referring Expressions (GRE), including Dale's *Full Brevity Algorithm* (1989, 1992), the *Incremental Algorithm* (Dale & Reiter, 1995; van Deemter, 2002), and the *Graph Algorithm* (Krahmer et al., 2003). Many of these algorithms implicitly base the generation of referring expressions on the Maxim of Quantity (Grice, 1975), which says that expressions should be as informative as required, but not more informative. This is in line with early work on reference by Olson (1970), who argued contrastiveness to be the primary function of referring expressions. As a result, many GRE algorithms aim for minimally specified, distinguishing references, though considerations of the computational complexity involved have motivated algorithms that do not always produce minimal descriptions. It is often assumed that human speakers refer in a similar, *distinguishing* way: that they include just enough information in their references for an addressee to single out who or what they are referring to ('the target'). Over the years, however, several psycholinguistic studies have questioned this assumption by revealing that speakers often overspecify their references and include more information than is strictly necessary for identification (e.g. Arnold, 2008; Arts, 2004; Brennan & Clark, 1996; Engelhardt et al., 2006; Pechmann, 1989). Yet, how speakers overspecify is still largely unknown. However, in order to improve the performance and human-likeness of GRE algorithms, it is important to address which factors cause referential overspecification.

We hypothesize that at least two factors cause speakers to overspecify. First, referential overspecification may be influenced by the properties of the *target referent*. In this respect, we hypothesize that targets that require more referential effort will more often result in overspecified references than targets that are easy to refer to. We expect two kinds of target properties to play a role: the *domain* in which a reference is produced and *cardinality*: whether references are singular or plural. For domain, we hypothesize that expressions produced in more complex domains are more frequently overspecified than expressions produced in a simpler domain. For cardinality, we expect that people are more likely to overspecify when they refer to plural targets (with two target objects) compared to singular targets (with one target object).

A second factor that we expect to influence referential overspecification is properties of the *communicative setting* in which references are produced. In this respect, we have two main hypotheses. First, since the production of speech is incremental, and since speaking takes arguably less effort than writing, we hypothesize that spoken references are

more frequently overspecified than written ones. Second, we hypothesize that speakers provide more information when they cannot see the addressee than when they can. When speaker and addressee can see each other, the speaker is able to receive both auditory and visual feedback from the addressee, which would make the speaker more confident about whether or not the information he provided was sufficient to single out the target referent. However, in case the addressee is not visible to the speakers, only auditory feedback is possible, which would lead to more uncertainty (e.g. Veinott et al., 1999) and, subsequently, to more referential overspecification.

Method

In order to investigate to what extent these two factors influence the information load of referring expressions, we performed a large elicitation experiment, in which participants were asked to describe target objects and distinguish them from surrounding objects. This resulted in the D-TUNA corpus, which consists of 2400 Dutch referring expressions. Data collection was inspired by the English TUNA experiment. For a detailed description of the TUNA experiment, see Gatt, van der Sluis & van Deemter (2007).

Participants

Sixty undergraduate students (14 males, 46 females) from Tilburg University participated in the experiment, either on a voluntary basis or for course credit. All participants (mean age 20.6 years old, range 18-27 years old) were native speakers of Dutch.

Materials

The materials consisted of forty trials, which all contained one or two target referents and six distractor objects. The target referents were clearly marked by red borders, so that they could easily be distinguished from the distractor objects. For each participant and each trial, the target and distractor objects were positioned randomly on the screen in a 3 (row) by 5 (column) grid. In order to manipulate the properties of the target referents, the trials varied in terms of their types of domains and in terms of cardinality.

Two types of domains A first manipulation of the target properties was that trials occurred in two different types of domains: the furniture domain and the people domain.

The twenty trials in the *furniture domain* contained pictures of four types of furniture items[1]. These items differed along four dimensions (see table 1).

[1] The pictures were taken from the Object Databank, developed by Michael Tarr at Brown University and freely distributed. URL: http://titan.cog.brown.edu:8080/TarrLab/stimuli/objects/

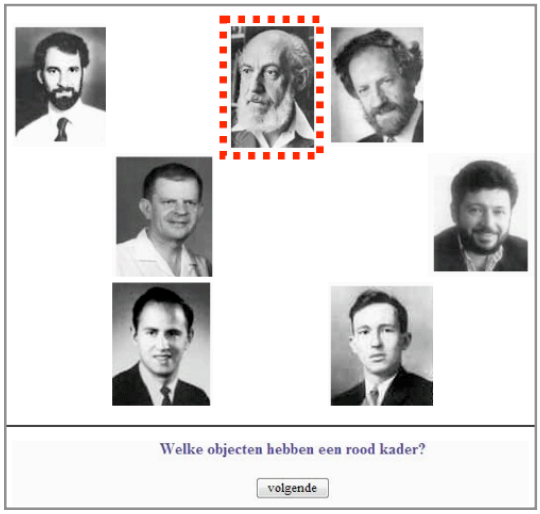

Figure 1: A trial in the people domain.

Table 1: Attributes and values of the furniture items.

Attribute	Possible values
Type	Chair, sofa, desk, fan
Colour	Blue, red, green, grey
Orientation	Front, back, left, right
Size	Large, small

The twenty trials in the *people domain* consisted of pictures of male mathematicians (for an example of a trial in this domain, see figure 1). For several reasons, this domain was the more complex of the two. First, targets in the people domain cannot be distinguished in terms of their type (since they all have 'type = person'). Second, the pictures of the persons are arguably more similar to each other than the furniture items, which makes them more difficult to distinguish from the distractor objects. Furthermore, the pictures of people were was not as controlled as the artificial pictures in the furniture domain and hence there may be more information in them that participants may use in their references. Last, the possible descriptions of people are somewhat open-ended, in that there are many unpredictable attributes that can be mentioned. However, a number of salient dimensions of variation were identified (see table 2).

Table 2: Attributes and values of the people pictures.

Attribute	Possible values
Type	Person
Orientation	Front, left, right
Age	Young, old
Hair colour	Dark, light
Has hair	0 (false), 1 (true)
Has Beard	0, 1
Has glasses	0, 1
Has shirt	0, 1
Has tie	0, 1
Has Suit	0, 1

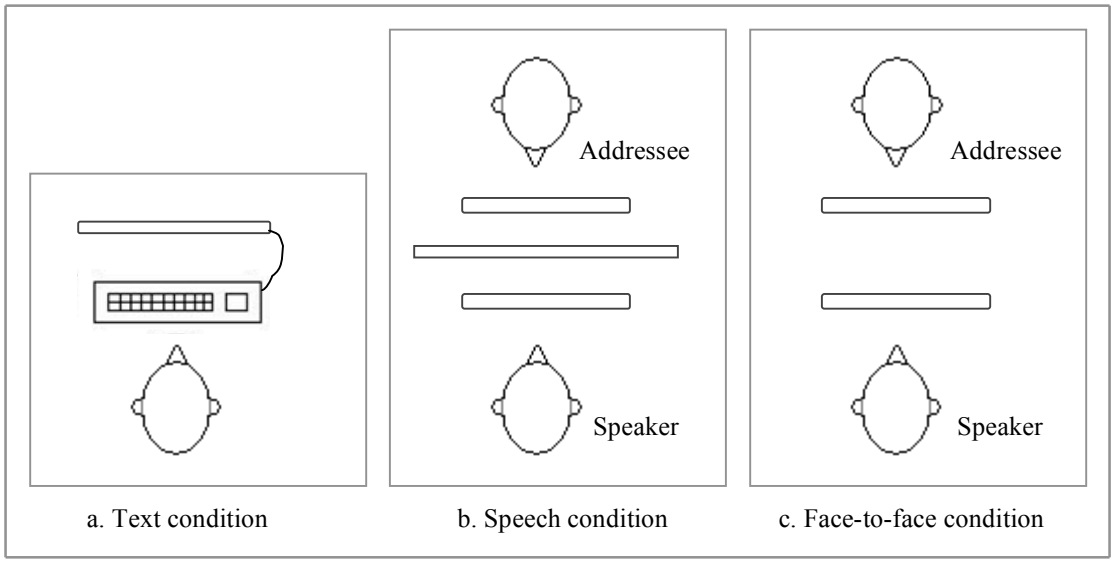

Figure 2a-c: A schematic overview of the three conditions.

Since speakers need a head noun in their references and therefore always use 'type' in their formulation (Levelt, 1989), trials were built in such a way that the attribute 'type' could never be a distinguishing attribute.

Two levels of cardinality A second manipulation of target properties was that trials differed in terms of cardinality, i.e. the number of target referents that they contained. Twenty trials were singular (SG, ten per domain) and contained one target referent. Furthermore, twenty trials (again ten per domain) were plural (PL) trials containing two target referents. An extra manipulation of the target complexity occurred by including two levels of similarity. Plural/similar (PS) trials contained two target objects with both identical distinguishing attributes, for example *'the table and the sofa that are both red'*, where the two target objects are distinguished from the distractors by means of their (shared) red colour. The plural/dissimilar (PD) trials (again five per domain) contained two target objects with different distinguishing attributes, for example *'the large fan and the red sofa'*, where the two target objects have to be distinguished by means of different attributes: size and colour.

Procedure

Each participant was presented the forty trials in a different random order. The experiments were individually performed in an experimental room, with an average running time of twenty minutes. All participants were filmed during the experiment.

The participants were asked to describe the target referents in such a way that an addressee could uniquely identify them. In order to manipulate properties of the communicative setting, the participants were randomly assigned to three conditions (text, speech and face-to-face). The *text* condition was a replication (in Dutch) of the TUNA experiment: participants produced written identifying descriptions of one or two target referents. No addressee was present, but the participants were told that their descriptions were sent to an addressee outside the experimental room. In the *speech* condition and the *face-to-face* condition, participants were asked to utter their descriptions to an addressee inside the experimental room. The addressee was a confederate of the experimenter, instructed to act as though he understood the references, but never to ask clarification questions. In the instructions, the participants were told that the location of the objects on the addressee's screen had been scrambled; hence, they could not use location. In the face-to-face condition, the addressee was visible to the participants; in the speech condition this was not the case, because a screen was placed in between speaker and addressee. A schematic overview of the three conditions is displayed in figure 2a-c.

Data annotation

The 2400 (3x20x40) identifying descriptions of the D-TUNA corpus were all semantically annotated using an XML annotation format: they were provided with information regarding attributes of both the target and distractor objects. For this annotation, we used the XML annotation scheme of the TUNA corpus (Gatt, van der Sluis & van Deemter, 2008b).

Annotating the descriptions of the D-TUNA corpus in the XML annotation scheme of the TUNA corpus is advantageous for several reasons. First, it makes the English and the Dutch corpus highly comparable. Furthermore, it makes the D-TUNA corpus a useful tool for the evaluation

Table 4: Examples of underspecified, minimally specified and overspecified references, with their number of words, their number of attributes (minus 'type'), and their number of redundant attributes.

Level of overspecification	Example	Number of words used	Total number of attributes	Number of redundant attributes
Underspecified	*'The man with the beard'*	5	1	-1
Minimally specified	*'The man with the white beard'*	6	2	0
Minimally specified	*'The white bearded man'*	4	2	0
Overspecified	*'The white bearded man without a tie'*	7	3	1

of GRE algorithms. Last, it facilitates corpus based analysis of overspecification.

Design and statistical analysis

The experiment had a 2x2x3 design (see table 3), with two within-subjects factors: *domain* (levels: furniture, people) and *cardinality* (levels: singular, plural), and one between-subjects factor representing communicative setting: *condition* (levels: text, speech, face-to-face).

Table 3: Overview of the experimental design and the number of descriptions within each cell.

	Furniture		People	
	Sing.	Plur.	Sing.	Plur.
Text	200	200	200	200
Speech	200	200	200	200
Face-to-face	200	200	200	200

We distinguish two dependent variables that indicate if and to what extent speakers overspecify their referring expressions. First, we consider the *number of words* that speakers use (filled pauses excluded) to be a (relative) indicator for referential overspecification, since the number of words is assumed to be systematically correlated to the amount of information conveyed. Second, the *number of redundant attributes* serves as a more robust indicator for referential overspecification. An attribute is considered to be redundant if removing it from the description would still result in a distinguishing reference. For example, reconsider figure 1. Several possible descriptions of the target referent (either underspecified, minimally specified or overspecified ones) are depicted in table 4, along with their corresponding number of words and number of redundant attributes. Since trials were built in such a way that 'type' could never be a distinguishing attribute, we excluded 'type' from our analysis.

The first reference given in table 4, *'The man with the beard'*, is underspecified. It contains only one attribute ('has beard = 1'), which is not enough for identification of the target. The next two references in table 4, *'The man with the white beard'* and *'The white bearded man'*, are minimally specified. They contain two attributes ('has beard = 1' and 'hair colour = light') that are both needed for identification of the target. Therefore, no redundant attributes are counted. Last, the fourth reference in table 4, *'The white-bearded man without a tie'* is overspecified, since it contains one redundant attribute that is not needed for identification of the target ('has tie = 0').

Our statistical procedure consisted of repeated measures analyses of variance (which compared means for subjects[2]) and Tukey HSD tests for multiple comparisons.

Results

The proportions of minimally specified, overspecified, and underspecified descriptions confirmed the finding in various psycholinguistic studies that human speakers tend to overspecify their references: 53.6% of the references were overspecified, which is even more than the proportion of minimally specified references: 41.4%. Only a small minority of the references (5.0%) was underspecified. More detailed analyses of the data indicated at least some factors that were responsible for referential overspecification: properties of the target referent and properties of the communicative setting.

Results for properties of the target

Results show that properties of the target serve as a first factor influencing referential overspecification.

Domain First, domain complexity can be regarded as a factor influencing referential overspecification. Figure 3 depicts the average number of words and redundant attributes as a function of domain.

[2] Repeated measures analyses of variance that compared means for items essentially gave similar results and will not be reported here.

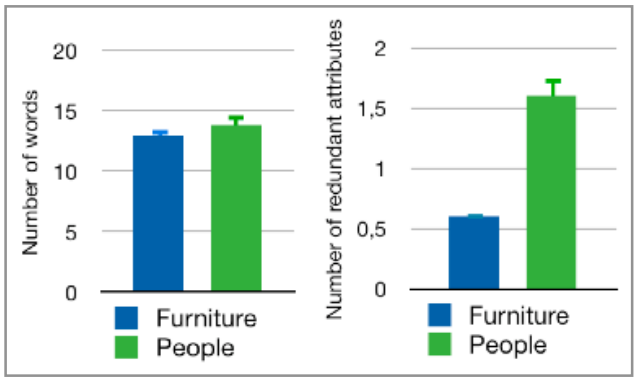

Figure 3: Average number of words and redundant attributes as a function of domain.

Although figure 3 shows a difference in number of words that were used to describe the targets, suggesting that people needed more words when referring to targets in the complex people domain ($M = 13.8$, $SD = .82$) compared to targets in the simpler furniture domain ($M = 12.9$, $SD = .56$), this difference was not statistically reliable ($F_{(1,57)} = 2.074$, *ns*). However, references to people did contain significantly more redundant attributes ($M = 1.6$, $SD = .15$) than references to furniture items ($M = .6$, $SD = .03$), $F_{(1,57)} = 9.419$, $p < .001$.

Cardinality A second factor that influences referential overspecification is cardinality. Figure 4 depicts the average number of words and redundant attributes as a function of cardinality.

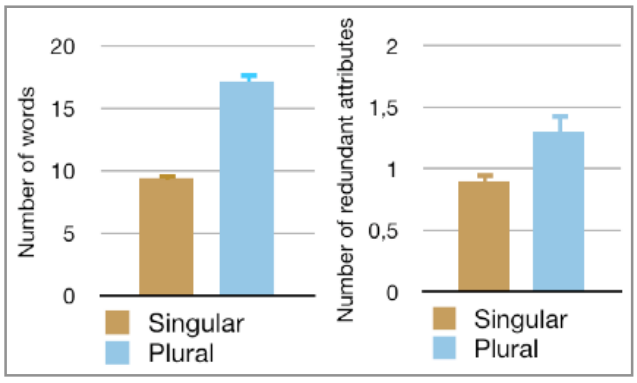

Figure 4: Average number of words and redundant attributes as a function of cardinality.

Intuitively, references to two target objects should contain more words than references to one target, which was confirmed by a significant difference between the number of words of singular references ($M = 9.4$, $SD = .47$) and plural references ($M = 17.2$, $SD = .80$), $F_{(1,57)} = 247.753$, $p < .001$. A similar pattern was observed for the difference between singular references ($M = .9$, $SD = .07$) and plural references ($M = 1.3$, $SD = .10$) with respect to the number of redundant attributes that were mentioned ($F_{(1,57)} = 42.714$, $p < .001$). These results indicate that when the target got more complex (as was the case with the two target objects in plural trials), participants tended to overspecify their references more frequently.

The effects of cardinality described above were stronger in the more complex people domain, as reflected in interactions between domain and cardinality for both of the two factors indicating referential overspecification. First, compared to the furniture domain, the effect of cardinality on the number of words that the references contained was stronger in the people domain ($F_{(1,57)} = 4.285$, $p < .05$). The same pattern was observed with the effects of cardinality on the number of redundant attributes ($F_{(1,57)} = 30.612$, $p < .001$). Thus, references to plural targets were more likely to get overspecified when they were uttered in the more complex people domain.

Similarity Results also show an effect of similarity on the number of words: the number of words used in references to dissimilar target referents ($M = 18.9$, $SD = .93$) was significantly higher than the number of words used in references to similar target referents ($M = 15.6$, $SD = .93$), $F_{(1,57)} = 35.468$, $p < .001$. However, the difference in terms of the number of redundancy between references to similar targets ($M = 1.3$, $SD = .07$) and dissimilar targets ($M = 1.4$, $SD = .15$) did not reach significance: $F_{(1,57)} = 1.809$, *ns*). Thus, plural references to two similar target referents did contain more words, but were not more redundant than plural references to two dissimilar target referents.

Results for properties of the communicative setting

Besides effects of target complexity, we also looked at the influence of communicative setting on overspecification in referring expressions. Figure 5 displays the average number of words and redundant attributes as a function of the three communicative conditions.

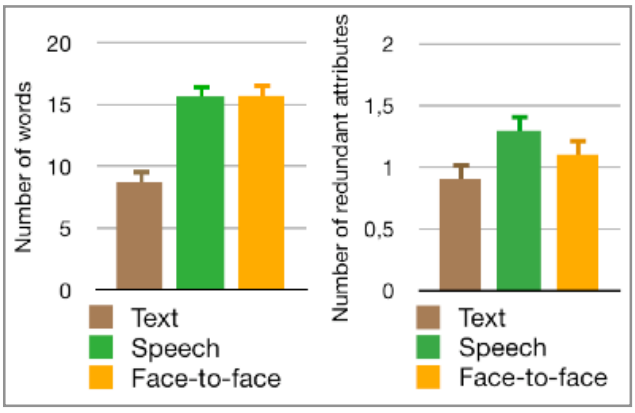

Figure 5: Average number of words and redundant attributes as a function of communicative setting.

First, there was an effect of communicative setting on the number of words that participants used when referring

($F_{(2,57)} = 13.574$; $p < .001$). More specifically, this effect was solely due to the fact that spoken references contained more words than written ones, resulting in significant differences between the text (*M = 8.7, SD = 1.1*) and speech condition (*M = 15.6, SD = 1.1*) ($p < .001$), and between the text and face-to-face condition (*M = 15.7, SD = 1.1*) ($p < .001$). However, the number of words used in spoken references in the speech and face-to-face condition did not differ. A different pattern was observed regarding the effect of communicative setting on referential overspecification in terms of the number of redundant attributes, but these effects did not reach significance ($F_{(2,57)} = 2.668$, *ns*).

The effects of communicative setting described above were stronger for plural references compared to singular references, because of an interaction between cardinality and communicative setting both of the two factors influencing referential overspecification. First, the effect of communicative setting on the number of words that speakers used were stronger for plural references, compared to singular references ($F_{(2,57)} = 15.438$; $p < .001$). However, regarding the number of redundant attributes, the interaction between communicative setting and cardinality did not reach significance ($F_{(2,57)} = 2.075$; *ns*).

Discussion

People often overspecify their referring expressions. This is a problem for GRE algorithms, since they tend to generate minimally specified references. Therefore, we have aimed to investigate which factors cause speakers to overspecify their references. In this paper, we have described an experiment in which 2400 Dutch referring expressions were gathered; over 50% of these references contained more information than strictly needed for identification of the target, which is in line with previous psycholinguistic studies on referential overspecification (e.g. Arts, 2004, Engelhardt et al., 2006; Pechmann, 1989). We hypothesized properties of the target and properties of the communicative setting to cause this overspecified references. Our analyses of the data have revealed several interesting findings.

The finding that referring expressions produced in the complex people domain are more redundant than references in the simpler furniture domain shows the influence of the *domain* in which references are produced on referential overspecification. We can conclude that when the range of possible attributes to describe a target referent gets broader (as is the case in more complex domains), language users tend to provide the addressee with more information to single out the target referent. The results also confirm our hypothesis on *cardinality* influencing overspecification in referring expressions. That is, when referring to two targets, people more frequently overspecify their expressions (in terms of both the number of words and redundancy) than when referring to a single target. The interaction between domain and similarity once again confirms the influence of domain complexity on referential overspecification.

An explanation for the differences between singular and plural references could be that, when referring, people prefer certain attributes to others. Several previous psycholinguistic studies have used attribute preference as an argument for the occurrence of referential overspecification (e.g. Pechmann, 1989; Belke & Meyer, 2002). For the case of cardinality, this would mean that when people describe two targets in plural trials, they use preferred attributes (such as *colour)* in their descriptions of both targets, even if *colour* would not have any contrastive value. This would result in using more words and more redundancy for plural references (e.g. *'The red desk facing left and the green chair facing right'*, with two times *colour* as preferred, but redundant attribute) compared to singular ones (e.g. *The red desk facing left'*, with one time *colour* as preferred, but redundant attribute). In the end, this would result in plural references containing more words and preferred (but redundant) attributes compared to singular references.

Our findings on the similarity of target referents show that plural references to two dissimilar targets are not more redundant than plural references to two similar target referents. An explanation for this could be that people have different strategies in referring to two target referents, irrespective of whether they are dissimilar of dissimilar: some always describe two targets in one reference by searching for one or more common attributes (e.g. *'The two men with…'*); others rather provide two separate descriptions of the two target referents (e.g. *'One man with… and another man with…'*), even if the targets have one or more common distinguishing attributes. These different referring strategies would explain why the results did not show redundancy effects of similarity.

Besides target complexity, we expected properties of the communicative setting to influence overspecification in referring expressions. Our findings in this respect show that written and spoken referring expressions do not differ in terms of redundancy, but do differ in terms of the number of words that they contain: speakers need more words to provide the same information as people who type their expressions. This difference is in line with the expectation that speakers are likely to use more words than writers, simply because typing requires more effort than speaking. However, a second explanation is that speakers, more than writers, tend to repeat attributes, which results in using more words, but not in more redundant attributes. This second explanation is confirmed by frequencies of repeated attributes in references uttered in the three communicative settings. It follows from table 5 that speakers are indeed, far more than writers, likely to include certain attributes more than once in their referring expressions.

Table 5: Frequencies of repeated attributes, for each communicative setting.

	Text	Speech	Face-to-face
Repeated attributes	20	205	214

The lack of difference between references uttered in the speech and face-to-face conditions shows that speakers who cannot see the addressee do not provide more information than speakers who can. Thus, the lack of auditory feedback did not lead to more referential overspecification, which contrasts with previous research (e.g. Veinott et al., 1999). An explanation for this could be that the communication between speaker and addressee in our study was rather one-sided and thus not very interactive. It was just the speaker who uttered a description; the addressee only reacted briefly when he had singled out the target. This lack of interaction could explain why we did not find significant differences between the speech and face-to-face condition in terms of referential overspecification.

In the future, we will aim to provide a more natural interaction between speaker and addressee. Furthermore, we will try to take into account that referring does not just distinguish one or two targets from others, but can also give information about the speaker's purpose to bring the object (and information associated with it) under the addressee's attention. Moreover, it would be interesting to investigate whether referring is a language dependent or a language independent process, which can be done by making a comparison between the English TUNA corpus and the Dutch D-TUNA corpus.

The results of this study reveal that generating minimally specified, distinguishing references is not sufficient for algorithms for the generation of referring expressions. These GRE algorithms will need to deal with the human tendency to overspecify references, and to acknowledge properties of the target referent and the communicative setting in which the references are uttered.

Acknowledgements

The research reported in this paper forms part of the VICI project 'Bridging the gap between psycholinguistics and computational linguistics: the case of referring expressions', funded by the Netherlands Organization for Scientific Research (NWO Grant 277-70-007).

References

Arnold, J. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, *23(4)*, 495-527.

Arts, A. (2004). *Overspecification in instructive texts.* Doctoral dissertation, Tilburg University.

Belke, E., & Meyer, A. (2002). Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during same-different decisions. *European Journal of Cognitive Psychology, 14,* 237-266.

Brennan, S., & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22(6)*, 1482-1493.

Dale, R. (1989). Cooking up referring expressions. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* (pp. 68-75). University of British Columbia, Vancouver, BC, Canada.

Dale, R. (1992). *Generating referring expressions: Building descriptions in a domain of objects and presses*. Cambridge: MIT Press.

Dale, R., & Reiter, E. (1995). Computational interpretation of the Gricean maxims of Quantity in the generation of referring expressions. *Cognitive Science, 19(8)*, 233-263.

Engelhardt, P., Bailey, K. & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language, 54*, 554-573.

Gatt, A., van der Sluis, I., & van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. *Proceedings of the eleventh European workshop on Natural Language Generation* (pp. 49-56). Saarbruecken, Germany.

Gatt, A., van der Sluis, I. & van Deemter, K. (2008b). *XML formatting guidelines for the TUNA corpus.* (Tech. rep.). Department of Computing Science, University of Aberdeen.

Grice, H.P. (1975). Logic and conversation. In Cole, P. and Morgan, J. (Eds.) *Syntax and Semantics: Speech Acts*, volume III. Academic Press.

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

Levelt, W.M.J. (1989). *Speaking: from intention to articulation*. MIT Press.

Olson, D.R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological review, 77*, 257-273.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*, 89-110.

van Deemter, K. (2002). Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, *28(1)*, 37-52.

Veinott, E., Olson, J., Olson, G. & Fu, X. (1999). Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*. (pp 302-309). Pittsburgh, Pennsylvania, United States.