Need I say more?

On overspecification in definite reference

Ruud Koolen

Need I say more? On overspecification in definite reference

Ruud Koolen
PhD thesis
Tilburg University, 2013

TiCC Ph.D. series no. 30

Financial support was received from the Netherlands Organization for Scientific Research (NWO), for the VICI-project "Bridging the gap between psycholinguistics and computational linguistics: the case of referring expressions."

ISBN: 978-94-6203-390-0
Print: CPI Wöhrmann print service

Need I say more?

On overspecification in definite reference

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan Tilburg University
op gezag van de rector magnificus,
prof. dr. Ph. Eijlander,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen commissie
in de aula van de Universiteit
op vrijdag 20 september 2013 om 14.15 uur

door

Ruud Martinus Franciscus Koolen

geboren op 27 januari 1984 te Deurne

Promotores:

Prof. Dr. E. J. Krahmer
Prof. Dr. M. G. J. Swerts

Promotiecommissie:

Prof. Dr. K. van Deemter
Dr. N. Katsos
Prof. Dr. A. Koller
Prof. Dr. A. A. Maes
Dr. P. Piwek

Contents

1

General introduction

The Doors are one of the bands from the late sixties that are still popular today: even forty-two years after the mysterious death of The Doors' lead singer Jim Morrison (who died in 1971 in Paris), people are still talking about this legendary band, and about the four musicians that created all those brilliant pieces of psychedelic music. So what kind of language do speakers produce when talking about The Doors?

For one thing, talking about the band's musicians involves *referring* to them: for example, if you want to describe The Doors' front man (who in that case is the *target referent*), you can use his common name ("Jim Morrison"), his first name ("Jim"), his full name ("James Douglas Morrison"), or one of his honorific nicknames (e.g., "the Lizard King", which he made up himself in a song called *Not to touch the earth*). But what if your addressee is in the unfortunate position of not knowing The Doors and their famous lead singer? In that case, you will have to come up with an alternative description of your target referent. For example, you could use the band shot depicted in Fig. 1.

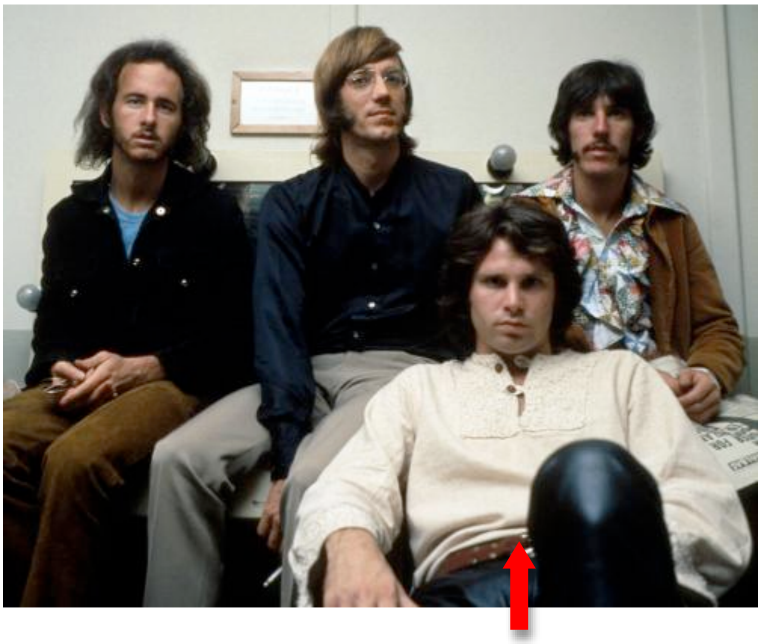


Fig. 1: The Doors. The red arrow marks the intended target referent (Jim Morrison).

If you were to use this picture to explain who Jim Morrison was, the first thing you would have to do is to make sure that your addressee knows whom of the four persons in Fig. 1 you are talking about. For that purpose, "the guy sitting in the front", for example, would do: it would uniquely distinguish your target person from the three other persons that are visible (i.e., the *distractors*). Or, if you do not want to use location information to identify your target, "the man wearing leather pants", "the guy with the white shirt", or "the guy wearing a white shirt and leather pants" would also suffice as uniquely identifying descriptions of the target referent. In general, these are all *definite target descriptions* consisting of a definite article, a head noun, and one or more modifiers.

The focus in this dissertation is on target descriptions such as these, whose content relies on the information available in the direct visual context of the target. As we will see, such descriptions are often *overspecified*, meaning that they contain one or more *redundant* modifiers that are – strictly speaking – not necessary for unique target identification. For example, "the guy wearing a white shirt *and* leather pants" is overspecified in the context of Fig. 1, since mentioning the white shirt *or* the leather pants would suffice to identify the target. Previous research has shown that human speakers often produce overspecified descriptions (e.g., Arts, 2004; Pechmann, 1989), and that redundant attributes have an effect on listeners during the target identification process (e.g., Arts, Maes, Noordman, & Jansen, 2011; Engelhardt, Bailey, & Ferreira, 2006).

Throughout this dissertation, various issues related to the production (chapter 2, 3, 4, 5) and processing (chapter 6) of overspecified descriptions are addressed from two perspectives: *psycholinguistics* and *computational linguistics*. In general, we report on results of psycholinguistic experiments, aiming to formulate implications for computational models that are built to automatically generate distinguishing descriptions of target referents (such as people or objects). In the next section, we therefore discuss existing models of human reference production, and the basic principles behind computational interpretations of such models. Needless to say, these models will be explained in more detail later on in the dissertation.

Existing models of reference production

In psycholinguistics, various models for the speech production process have been proposed. These models generally assume that human speech production consists of several modules, where a common distinction is made between how speakers "decide-what-to-say", and how they "decide-how-to-say-it". In Levelt's (1989) *Blueprint for the speaker*, the Conceptualizer is responsible for the former, regulating several mental activities for the speaker such as "conceiving the intention to produce an utterance, selecting the relevant information to be expressed for the realization of this purpose, keeping track of what was said before, and so on" (Levelt, 1989, p. 9). The output of this conceptualization process (the *preverbal message*) serves as the input of the next processing component, which Levelt calls the Formulator. In this formulation stage, the preverbal message is turned into a linguistic structure via grammatical and phonological encoding. The end product is an internal representation of how the planned utterance should be articulated (Levelt, 1989).

The conceptualization and formulation stages can also be applied to the production of definite descriptions. If one regards target identification as the core purpose of reference (which is often done in psycholinguistic research on reference production and overspecification; for example Olson, 1970; Pechmann, 1989; Engelhardt et al., 2006), the conceptualization stage mainly involves speakers' decision on which information they want to include in order to make the target identifiable. For example, in Fig. 1, a speaker might decide to mention that the intended target referent is wearing a white shirt, because this is an attribute that can be ascribed to the target person, but not to the other three persons in the visual scene. In this way, also other attributes might be selected, for example that the target referent is a male person, that he is wearing leather pants, or that he has curly black hair. After this selection process (which thus appears to be a matter of choice), it is the Formulator's task to encode all selected attributes into a grammatical noun phrase (e.g., "the guy wearing leather pants and a white shirt"). Given that properties can be realized in different ways (e.g., "the leather pants" versus "the pants made of leather"), problems of choice need to be solved here as well.

The production of referring expressions has also been studied extensively from a computational perspective, namely in the field of Natural Language Generation (NLG). NLG is a subfield of Artificial Intelligence (AI), and researchers in this field generally aim to build computational algorithms that are able to automatically convert non-linguistic information (e.g. from a database) into natural language text or speech. These NLG algorithms typically require a component that allows them to compute descriptions of objects, regardless of their purpose (Mellish et al., 2006). Therefore, several Referring Expression Generation (REG) algorithms that are able to do this have been developed. Many of these algorithms are primarily concerned with the conceptualization stage in Levelt's (1989) *Blueprint for the speaker*: they aim to select attributes to distinguish a target from one or more distractors. How do the current algorithms do this?

Fig. 2 shows the common schema (taken from Viethen, Dale, & Guhe, submitted) that classic REG algorithms and their descendants generally follow when generating a distinguishing target description (see also Dale and Reiter (1995), and Krahmer and Van Deemter (2012) for more detailed descriptions of the traditional REG problem).

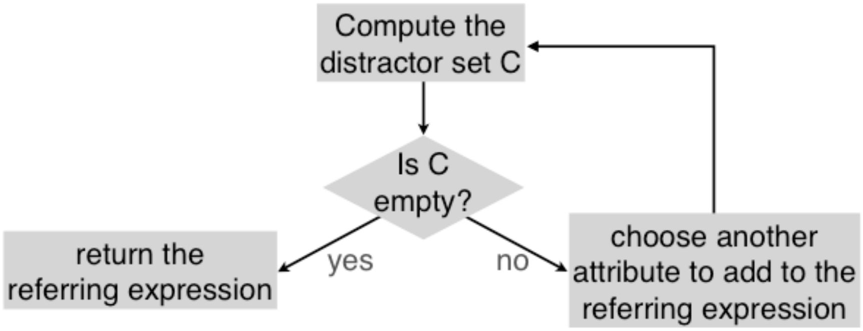

Compute the distractor set C
Is C empty?
yes
no
return the referring expression
choose another attribute to add to the referring expression

Fig. 2: The traditional REG model.

As can be seen in Fig. 2, REG algorithms generally repeat two steps as long as the expression under construction is not fully distinguishing: (1) add an attribute to the description; (2) recompute if any distractor remains. The distractor set C is usually taken to be the set of objects that are present in the direct visual context of the target. In Fig. 1, for example, the target is Jim Morrison, while the distractor set C consists of the three people in the back (i.e., the other musicians, who are in this case logical candidates to become members of the distractor set C). In order to generate a distinguishing description of the target in this specific case, an algorithm could for example consider the attribute *hair color* (which has the value *black* for this target). As can be seen in Fig. 1, this attribute-value pair <*hair color = black*> would exclude the one blonde guy (i.e., the middle person in the back) from the distractor set. Thus, least one extra attribute should be added to rule out the two other persons as well, since the algorithms' stop criterion is an empty distractor set C.

What remains is a problem of choice: which attribute to consider first, if there are many attributes to choose from that all rule out at least one distractor object? And, if more than one attribute is needed: which attribute to consider next? In order to solve this issue, some REG algorithms are built around the notion of *discriminatory power*, which can be defined by the number of distractor objects that can be ruled out by an attribute or a set of attributes. The various algorithms interpret this notion in different ways: from strict to more relaxed. One strict interpretation is used in Dale's (1989) *Greedy Algorithm*, which selects the attribute with the highest distinguishing value at every stage of the selection process; one more relaxed interpretation comes from Dale and Reiter's (1995) *Incremental Algorithm* (IA), considering preferred attributes for inclusion before less preferred ones. The IA uses a Preference Order (PO) to do this (i.e., a domain-dependent ranking of attributes that typically follows from empirical data), only selecting an attribute if it rules out at least one of the remaining distractors. Other REG algorithms do not necessarily use the notion of discriminatory power. For example, the *Full Brevity Algorithm* (Dale, 1989) always seeks to find the shortest possible description (thus relying on a different strategy than the one outlined in Fig. 2), while the *Graph Algorithm* (Krahmer, van Erk, & Verleg, 2003) uses preferences of attribute-value pairs that are modelled as cost functions (where cheaper is more preferred). In this way, the Graph Algorithm searches for the cheapest overall description, and might opt for one single, relatively dispreferred attribute (e.g., eye color) that rules out all distractor objects at once, instead of combining several more preferred properties.

In recent years, researchers have started asking to what extent the descriptions that these various models generate are comparable to those produced by human speakers (e.g., Viethen & Dale, 2006; Van Deemter, Gatt, Van der Sluis, & Power, 2012a), which raises the question how *humanlike* these descriptions are. In this respect, it is interesting to note that, like human speakers, algorithms such as the IA do sometimes generate target descriptions that turn out to be overspecified in the end (they may for example select an attribute that renders all attributes selected earlier obsolete). Does this imply that the algorithms' output is comparable to human-produced descriptions? And, related to this, to what extent are the current REG algorithms actually intended to mimic human referential behavior? Van Deemter, Gatt, Van Gompel and Krahmer (2012b) point out that this is not necessarily true. In this respect, Van Deemter et al. (2012b) signal that Dale and Reiter (1995) on the one hand argue that "one way to create a computational model is to determine how speakers generate texts and build algorithms based on these observations" (p. 252), but also that "psychological realism is not the most important consideration for developing an algorithm" (p. 253). In any case, Van Deemter et al. (2012b) conclude that algorithms such as the IA are at least often interpreted as aiming to mimic human-produced referring expressions, because they often build on the Maxims of conversational implicature (Grice, 1975), as well as on psycholinguistic theories of incremental speech production (Pechmann, 1989).

This ambiguity regarding the goal of the REG current algorithms is also relevant from an evaluation point of view. If the aim is to generate target descriptions that lead to short identification times (Garoufi & Koller, 2011), or descriptions that are easy to identify (Paraboni, Van Deemter, & Masthoff, 2007), humanlikeness is perhaps not the most suitable evaluation criterion.

However, a substantial amount of work in REG has taken the perspective of humanlikeness and uses human-produced descriptions to evaluate the algorithms' output. One recent example comes from Van Deemter et al., (2012a), who used a big corpus of human-produced descriptions (the TUNA corpus) to evaluate the algorithms discussed in Dale and Reiter (1995).

Irrespective of whether the main goal of the current REG algorithms is to generate humanlike and realistic output, they can make interesting predictions that are relevant to psycholinguistics. For example, as explained above, some of the current algorithms predict that target descriptions can actually be overspecified. In this dissertation, we build on this observation, addressing the following research question:

"Which factors cause human speakers to overspecify their target descriptions, and to what extent can these be modeled by existing REG algorithms?"

In general, we report on the results of psycholinguistic experiments, seeking to find patterns in the human production and processing of overspecified object descriptions, and aiming to formulate implications for the performance of current REG algorithms. A detailed outline of the dissertation is provided in the next section.

Focus and outline

This dissertation reports on five studies related to the production and perception of overspecified definite target descriptions. Given that we aim to formulate implications for the automatic generation of target descriptions, we adopt several basic principles that are generally used by the classic REG algorithms. Firstly, the descriptions under study are initial references to concrete entities in the world (such as chairs or persons) rather than to abstract entities (such as "reality" or "democracy"). Secondly, as said earlier, our focus is on so-called *exophoric* descriptions, whose content follows from the knowledge that is available in the visual context of the target referent (that is, the target itself and the surrounding distractors). This means that descriptions containing information that cannot be derived from this context (such as names, or descriptions like "the guy with the beautiful baritone voice") are beyond the scope of our research. The same goes for target descriptions containing negative attributes (such as "the guy without glasses"). Lastly, our focus is on so-called "one shot" descriptions, implying that discourse-related factors that may influence speakers' reference production (such as recency) are not addressed either.

The first study presents the D-TUNA corpus, which is a corpus of semantically annotated Dutch target descriptions; that is, all its expressions (that were produced by participants in a large-scale production experiment) were annotated with information regarding attributes of both the target and distractor objects. The data collection and annotation was inspired by the English TUNA corpus (collected by Van Deemter, et al., 2012a), which has been used before for the evaluation of REG algorithms (e.g., Gatt & Belz, 2010). In this first study, we use the D-TUNA corpus to explore various factors that might cause speakers to overspecify their target descriptions: the *domain* in which a description is uttered (i.e., artificial pictures of furniture items vs. realistic pictures of people), the *number of targets* that a scene contains (one or two), and the *communicative setting* in which a description is produced, focusing on the effects of modality (speech vs. writing) and interactivity (monologue vs. dialogue settings).

The second study investigates *learning curves* for REG algorithms: how many human-produced references are needed to make a good estimate of which attributes are preferred in a given domain? For previously unstudied domains, Van Deemter et al. (2012a) argue that systematically considering all possible preference orders quickly becomes impractical, because there are often too many possibilities to test. Given that this problem exists for Dale and Reiter's (1995) Incremental Algorithm *and* Krahmer et al.'s (2003) Graph Algorithm (which both – albeit in different ways – use attribute preferences), this second study explores how difficult it actually is to determine which attributes are preferred in a new domain. Are hundreds of human-produced instances needed, or can a few of them do? This question is answered for two *algorithms* (IA and Graph), in two *domains* (furniture and people), and for two *languages* (English and Dutch; we use the TUNA and D-TUNA corpora).

The third study explores the link between visual scene perception and referential overspecification, and in particular how the amount of *variation* in a scene relates to speakers' tendency to overspecify. The motivation for this study comes from one of the findings presented in the first study, being that speakers are more likely to include one or more redundant attributes when they are to refer to photo-realistic pictures of people rather than to artificial pictures of furniture items. What causes this difference is difficult to explain: it could conceivably be due to the difference in the extent to which the pictures in the two domains were realistic, but another explanation could be that the pictures in the people domain left speakers with more possible attributes to distinguish the target. In other words, the amount of visual variation might play a role here. In order to test this directly, the third study reports on three experiments that in a different (but related) way manipulate the amount of variation within a *single* domain. Thus, more specifically, we systematically test whether speakers are more likely to overspecify when they are presented with *high* variation scenes as compared to *low* variation scenes.

The fourth study goes further into the assumed link between scene perception and referential overspecification by exploring various factors that may determine whether or not an object in a scene is regarded as a relevant distractor of the target. Regarding REG algorithms, Dale and Reiter (1995) define the distractor set as the set of entities that the addressee is currently assumed to be attending to (also referred to as C in Fig. 2), but they do not explain explicitly how this set of distractors should be determined for a scene, which implies that this set is often taken to exist of all elements in the context set (except for the target referent). In line with Krahmer and Theune (2002), who present an extension of the IA that is able to restrict the distractor set by relying on linguistic salience, this fourth study explores to what extent *visual salience* might play a role in this as well. Following Itti and Koch's (2000) model of selective visual attention, we test how both *bottom-up*, perceptual saliency cues (such as the color and type of a distractor, the presence of visual clutter, and the distance between the target and a potential distractor), and *top-down*, conceptual saliency cues (in this case the specificity of the referential task) guide speakers in restricting the distractor set, and in their redundant use of color.

The fifth and final study investigates how listeners *perceive* redundant attributes. While previous work on reference production and Referring Expression Generation has generally taken target identification as the core purpose of reference (e.g., Dale & Reiter, 1995; Olson, 1970; Pechmann, 1989), this fifth study explores to what extent objective redundant modifiers (providing color and shape information) guide children in two age groups (six- and nine-years-old) in their choices for sweets. Especially young children have a high chance to derive false implicatures in the sense of Grice (1975), since their pragmatic capabilities are taken to be under development (e.g., Davies & Katsos, 2010), meaning that they must learn to understand the implications of information that is provided to them when they are in a conversation (Siegal & Surian, 2004). Therefore, in this last study, we investigate to what extent there are developmental differences between the children in the two age groups in how they are guided by redundant information in their choices for sweets.

The first three studies are based on papers that have been published in scientific journals, and the other two have been submitted as full papers. Being self-contained, the chapters all have their own abstract, introduction, discussion and reference list. Due to this self-contained nature of the chapters, a small amount of redundancy in the respective introductions was unavoidable. Chapter 7 contains a general discussion and conclusion.

References

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the 27th annual meeting of the association for Computational Linguistics*, 68-75. University of British Columbia, Vancouver, BC, Canada.

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

Davies, C. and Katsos, N. (2010). Over-informative children: Production / comprehension asymmetry or tolerance to pragmatic violations? *Lingua*, 120, 1956-1972.

Engelhardt, P. E., Bailey, K. G. D., and Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language,* 54, 554-573.

Gatt, A., and Belz, A. (2010). Introducing shared task evaluation to NLG: the TUNA shared task evaluation challenges. In E. Krahmer & M. Theune (Eds.), *Empirical methods in Natural Language Generation.* Berlin and Heidelberg: Springer (LNCS 5790).

Garoufi, K., & Koller, A. (2011). Combining symbolic and corpus-based approaches for the generation of successful referring expressions. In *Proceedings of the 13th European workshop on Natural Language Generation (ENLG)*, 121-131. Nancy, France.

Grice, H. P. (1975). Logic and conversation. In: P. Cole, & J. L. Morgan (Eds.), *Speech Acts*. Academic Press, New York.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40, 1489-1506.

Krahmer, E., van Erk, S., Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, 29 (1), 53-72.

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In: K. van Deemter & R. Kibble (Eds.). *Information sharing: Givenness and newness in language processing* (pp. 223-264). CSLI publications, Stanford.

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

Mellish, C., Scott, D., Cahill, L., Evans, R., Paiva, D., Reape, M. (2006). Reference architecture for Natural Language Generation systems. *Natural Language Engineering*, 12 (1), 1-34.

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

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

Siegal M. and Surian, L. (2004). Conceptual development and conversational understanding. *Trends in Cognitive Sciences*, 8, 534-538.

Van Deemter, K., Gatt, A., Van der Sluis, I., & Power, R. (2012a). Generation of referring expressions: assessing the Incremental Algorithm. *Cognitive Science*, 36, 799-836.

Van Deemter, K., Gatt, A., Van Gompel, R., & Krahmer, E. (2012b). Toward a computational psycholinguistics of reference production. *Topics in Cognitive*

Science, 4 (2), 166-183.

Viethen, H., & Dale, R. (2006). Algorithms for generating referring expressions: do they do what people do? In *Proceedings of the 4th International conference on Natural Language Generation (INLG)*, 63-70. Sydney, Australia.

Viethen, H., Dale, R., & Guhe, M. Referring in dialogue: alignment or construction? *Submitted to Language and Cognitive Processes*.

Need I say more?

2

Factors causing overspecification

Abstract

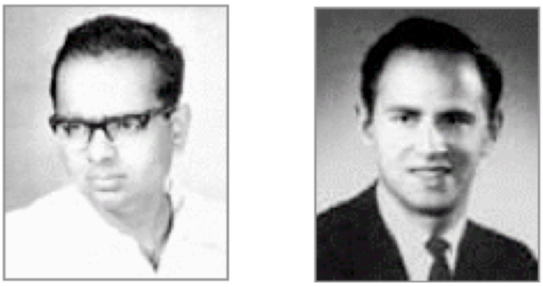
Speakers often overspecify their target descriptions and include more information than necessary for unique identification of the target referent. In the current chapter, we study the production of definite target descriptions, and explore several factors that might influence the amount of information that is included in these descriptions. First, we present the results of a production experiment investigating referential overspecification as a function of the properties of a target referent and the communicative setting. The results show that speakers (both in written and oral conditions) tend to provide more information when a target is plural rather than singular, and in domains where the speaker has more referential possibilities to describe the target. However, written and spoken referring expressions do not differ in terms of semantic redundancy. We conclude our chapter by discussing the implications of our empirical findings for pragmatic theory and for language production models.

This chapter is based on:

- Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43 (13), 3231-3250.

Introduction

Speaking often involves referring to things in order to identify them. Speakers already learn how to refer from an early age (Ford & Olson, 1975; Yip & Matthews, 2007), and referring expressions can be found across all languages. When referring, language users constantly need to decide about which and how much information to provide to make communication successful. For example, consider the two persons in Fig. 1.



Person A Person B

Fig. 1: Two possible target referents

Suppose you want to point out person A (the *target*) to an addressee. To do this adequately requires describing person A in such a way that the addressee can distinguish him from person B (the *distractor*). It stands to reason that there is a choice between different distinguishing attributes in order to make person A identifiable; definite descriptions like “The man wearing glasses”, “The man with the white shirt”, and “The man without a tie” all contain sufficient information for the addressee to single out the target.

Although the aforementioned target descriptions contain different attributes to distinguish person A, they have in common that they are all minimal identifying descriptions of the target referent. That is, apart from the type, or basic category to which the target entity belongs (namely, ‘person’), they contain just enough additional information (*attributes*) for the addressee to single out the target. However, several psycholinguistic studies have revealed that speakers often overspecify their references and include more information than is strictly necessary for identification (e.g., Arts, 2004; Brennan & Clark, 1996; Engelhardt et al., 2006; Pechmann, 1989). Thus, when referring to person A in Fig. 1, a speaker could produce an overspecified expression like "The black-haired man wearing glasses" instead of a minimally specified expression.

While the phenomenon of referential overspecification is by now fairly well established, it remains largely unexplored how factors such as properties of the target referent and properties of the communicative setting influence it. In this chapter we therefore present the results of a large-scale language production experiment in order to investigate whether and to what extent this is the case.

The production of definite reference

When speakers refer, they usually aim to identify a particular target referent to a listener (e.g., Searle, 1969). How do speakers decide which attributes to include in a target description? A straightforward way to do this is by selecting those attributes that rule out other entities in the context of the target. Olson (1970) emphasized this primarily contrastive function of referring expressions by revealing that decisions regarding the semantic content of a referring expression (the attributes of a referent to include) are determined by the speaker's knowledge of the intended target referent and the surrounding distractor objects. This implies that referring expressions are contrastive by means of their distinguishing attributes: speakers identify a target referent by ascribing a combination of attributes to it that cannot be ascribed to the distractor objects (Bach, 1994). In everyday discourse, the attributes that speakers include in their referring expressions are determined by endophoric or exophoric considerations (or, often, by a combination of the two).

In case referring expressions are endophorically used, this means that the attributes they contain are determined on the basis of previously described objects in the unfolding discourse. Sometimes, however, the previous expressions in the discourse are not taken into account; the referring expressions are then exophorically used. When producing exophoric expressions, the speaker only uses the knowledge available in the direct physical context of the target (that is, the target itself and its immediate distractors) to determine which and how much information to mention. Since exophoric referring expressions do not take previous expressions into account, and especially serve to identify objects in an environment, they are considered more basic than endophoric ones.

Although referring expressions can occur in numerous forms, the focus in this chapter is on definite noun phrases, rather than on pronouns, nominalizations or other linguistic mechanisms used for reference. The definite noun phrases under study here consist of a definite article, a head noun and possibly one or more modifiers (e.g., "The man wearing glasses" or "The red button"). These modifiers may be realized in different ways, for example adjectives ("red"), restrictive relative clauses ("which is red"), and so on. Furthermore, the noun phrases we study refer to physical objects in the world (such as persons or chairs), and not to abstract entities (such as "reality" or "democracy"). It needs to be emphasized that we take a purely exophoric perspective in studying reference: the expressions we study are first mention definite descriptions, and are solely intended to identify a target referent to an addressee without serving any other communicative goal. These kinds of definite descriptions are, among others, commonly used in written text (Poesio & Vieira, 1998), and in written and spoken instructive discourse (see for example, Maes et al., 2004).

The occurrence of referential overspecification

As the example used in the introduction of this chapter shows, speakers can describe a target in many different ways, using different types and amounts of information. In this respect, several studies have shown that speakers often include more information in their descriptions than strictly necessary for unique target identification. Thus, instead of being minimally specified, referring expressions are often overspecified. For example, Deutsch and Pechmann (1982) designed a language production game in which speakers were presented with eight objects (in four domains) and were asked to refer to one object that they would like as a birthday present.

Once the addressee was able to identify the speaker's target, he placed it in a box. Deutsch and Pechmann found 28 percent of the adults' referring expressions to contain one or more redundant attributes of the target. In a language production experiment by Pechmann (1989) speakers were asked to uniquely identify one target object from a set of distractor objects (such as cars, chairs, etc.). The objects differed in type and/or color and/or size. Pechmann found 21 percent of the speakers' expressions to contain at least one redundant attribute and thus to be overspecified. Nadig and Sedivy (2002) also found evidence for referential overspecification in an experiment in which speakers were asked to describe a target object (for example a big glass) in a display of four objects to an addressee. One contrasting distractor object of the same class as the target (for example a small glass) was either absent, visible only to the speaker, or visible to both speaker and addressee. Results revealed that when the contrasting target object was not visible to the addressee, over 50 percent of the descriptions contained redundant modifiers. Similar results were found by Maes et al. (2004), who performed an experiment in which participants produced written instructive texts on how to use a radio alarm. Their results showed both initial references (52 percent) and anaphoric referring expressions (75 percent) to be frequently overspecified. In Engelhardt et al. (2006), speakers produced referring expressions by asking them to instruct an addressee (who was a confederate) to move real target objects on or in some other object (e.g., an apple on a towel). In the *matching* location condition, the target object had to be moved from one location to another of the same type (e.g., from one towel to another towel), while in the *different* location condition the target object was moved to a different type of location (e.g., from a towel to a box). Overall, referring expressions contained redundant information approximately one third of the time.

Overspecification vs. the Gricean Maxim of Quantity

The occurrence of overspecification seems not to be in agreement with the Cooperative Principle formulated by Grice (1975), and with the Maxim of Quantity in particular. This conversational maxim consists of the following two components (Grice, 1975: 45):

1. Make your contribution as informative as is required (for the current purposes of the exchange);
2. Do not make your contribution more informative than is required.

With respect to the amount of information that a speaker should provide to identify a target to an addressee, a strict interpretation of the Maxim of Quantity seems to suggest that the speaker should provide enough information to identify the target (Quantity 1), but not more information (Quantity 2). Hence, this would imply that any attribute that is not needed for target identification should be considered redundant, and that the inclusion of redundant attributes would lead to overspecified descriptions.

Given that people are known to cooperate when they are in a conversation (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986), addressees will expect all information included in an expression to be relevant. Therefore, interpreting the Maxim of Quantity in a strict manner (such as described above) seems not to be in line with what Grice intended (Bach, 2006): one can think of other reasons why a speaker would – consciously or not – include information that does not directly serve the target identification goal. As Grice states himself, the information level of a contribution should depend on the purpose that a speaker has in uttering it. Mooney (2004) further elaborates on this by arguing that the Maxims that are part of Grice's Cooperative Principle are generally contextually sensitive, which – with regard to the Maxim of Quantity – implies that it depends on the context in which an utterance takes place whether all information provided is relevant. According to Mooney, the notion of activity types by Levinson (1979) can be used to contextualize the Cooperative Principle, since these activity types include (non-linguistic) goals of discourse that involve more than the communication of information alone.

In any case, it is hard to give an a priori definition of what determines whether a given conversational contribution is overspecified in a given context. However, for reasons of making the Maxim of Quantity (and the phenomenon of overspecification in particular) testable in a controlled setting, there is a variety of papers in pragmatics (e.g., Arts et al., 2011; Engelhardt et al, 2006), psycholinguistics (e.g., Olson, 1970; Pechmann, 1989), and computational linguistics (e.g., Dale, 1989; Dale & Reiter, 1995; Jordan, 2000; 2002) that investigate overspecification in the light of target identification. As stated in the beginning of this chapter, this is also what we aim to do here, and we therefore regard descriptions as overspecified in case more information is included than strictly necessary for unambiguous identification of the target. In this way, we investigate several factors that may influence the specification level of definite target descriptions, aiming to uncover the cognitive processes that underlie the production of such descriptions with a view to explaining why speakers, even in apparently 'simple' settings, overspecify.

Why do speakers overspecify?

Why do speakers often provide more information than strictly necessary when describing a target referent? The traditional cognitive view on human referring behaviour emphasizes that speakers design their referring expressions in such a way that the addressee can efficiently identify the target (Arnold, 2008). This suggests that referring is an addressee-oriented process, and that one reason why speakers often overspecify their referring expressions could be that it helps the addressee to identify the target more quickly. This view is also compatible with some pragmatic accounts of reference, notably Searle's (1969). Several empirical studies have provided support for this addressee-oriented view by revealing that listeners find it easier to identify a target referent when the speaker's description is overspecified, rather than minimally specified (Deutsch, 1976; Nadig & Sedivy, 2002; Paraboni et al. 2006; Sonnenschein, 1984; Sonnenschein & Whitehurst, 1982). Further support came from Engelhardt et al. (2006), who demonstrated that addressees judging the quality of instructions do not rate overspecified referring expressions to be any worse than minimally specified ones.

One possible explanation for overspecified references being beneficial for addressees could be that addressees create a *conceptual gestalt* of the target object for which they have to search (Levelt, 1989). Perceptually salient target attributes such as the basic-level category (type) and color are central to a gestalt, because they help the addressee to construct a mental representation of the target. This is in line with studies by Eikmeyer and Ahlsèn (1996), Pechmann (1989), and Schriefers and Pechmann (1988), who found type and color to be omnipresent in referring expressions, irrespective of their contrastive value. This would suggest that speakers redundantly mention such perceptually salient target attributes because they help listeners to complete their conceptual gestalt and speed up the identification process. This was confirmed by Arts et al. (2011), who revealed that in individual expressions, overspecified referring expressions that contained perceptual information (e.g., about shape, color, or size) led to significantly shorter identification times than minimally specified expressions. Davies and Katsos (2009) have also emphasized this pragmatic contribution of perceptually salient attributes to successful communication.

However, there is also some evidence that overspecification is detrimental for listeners, at least beyond a certain point. In a separate experiment using eye tracking, Engelhardt et al. found that listeners tended to be slowed down by overspecified expressions during identification. This result has since been replicated using a different methodology by Engelhardt et al. (2011). In a related vein, eye tracking on reference resolution in the Visual World Paradigm (Tanenhaus et al., 1995) has consistently observed what is known as a *point of disambiguation effect*, whereby listeners process incoming information in a description incrementally as a description is heard, circumscribing the domain of possible referents to exclude distractors that do not have the attributes heard so far in the description until a point is reached where the referent has been identified (Eberhard et al., 1995; Sedivy et al., 1999). However, this seems to suggest that no further domain circumscription would occur beyond the point where a target has been identified, even if further information were available in the incoming description.

Previous research on reference production has shown that facilitating the identification process is not the only reason why speakers overspecify their referring expressions. Speaker-oriented processes may also play a role. For example, speakers often contrast a particular target referent with a previously mentioned one, despite the fact that this previous target referent is no longer visually present in the visual domain (Levelt, 1989). This kind of displaced communication (Spivey & Richardson, 2008) makes the speaker include the distinguishing attribute of the previous target object in the reference to the new target object, irrespective of its contrastive value. Further evidence for speaker-oriented processes comes from Belke and Meyer (2002), who conducted an eye tracking study to explore speakers' preferences for certain attributes. In their experiment, Belke and Meyer had participants judge whether certain target and distractor objects were similar or different. Results showed a high proportion of overspecified referring expressions, due to the speakers' tendencies to redundantly mention color in their descriptions. Belke and Meyer claim that this result was due to the fact that speakers preferred perceptually salient attributes (such as color). Furthermore, because these attributes are absolute, they do not require the speaker to compare the target with the distractor object to determine the right value. In contrast, Belke and Meyer also found the opposite to be the case for attributes such as size, which do require a target referent to be compared to its distractors in order to determine the right value (e.g., whether the referent is large or small).

Computational implications of referential overspecification

The production of referring expressions has received considerable attention in Natural Language Generation (NLG), a subfield of Natural Language Processing and Artificial Intelligence (AI) that aims to build systems that automatically generate natural language text or speech from non-linguistic information (e.g., from a database; Reiter & Dale, 2000). Practical applications of NLG include, among others, the automatic generation of weather forecasts (Goldberg, 1994; Reiter et al., 2005), summarization of medical information (Gatt et al., 2009), and generation of instructions in contexts such as museum tours (Stock et al., 2007).

Given the ubiquity of referring expressions in natural language, it is no surprise that NLG systems typically also require algorithms that compute distinguishing descriptions for objects (Mellish et al. 2006). Various Referring Expression Generation (REG) algorithms have been proposed. As in the case of psycholinguistic research in this area, REG algorithms have typically taken the Gricean Maxim of Quantity as a starting point. For example, Dale's Full Brevity algorithm (Dale, 1989, 1992) was based on a strict interpretation of the Maxim, seeking to find the shortest possible description (in terms of the number of attributes included) for a target referent. This turned out to be computationally expensive, because it involved exhaustive search through the space of possible combinations of the target attributes in order of their length. A more relaxed interpretation of the Maxim is afforded by the Incremental Algorithm (Dale & Reiter, 1995), which was partly motivated by some of the psycholinguistic phenomena identified above, particularly the fact that some attributes are preferred over others. To do this, the Incremental Algorithm performs a "hillclimbing" search along a predetermined *preference order*, which lists attributes in order of preference (for example, placing color before size). Given an intended referent, the algorithm searches along the preference order, adding an attribute to a description if it has some contrastive value and terminating once the description is fully identified. Since the algorithm does not backtrack to remove attributes that turn out to be redundant, a description can be overspecified.

Since the work of Dale and Reiter (1995), there have been many further developments of the original Incremental Algorithm, for example to handle plural referring expressions, in particular, those which involve logical disjunction of attributes, and are typically realised as coordinate NPs like "the blue and brown books" (Van Deemter, 2002; Gatt, 2007). In addition, there has also been some work on developing frameworks that can accommodate different algorithms, such as graph-theoretic approaches (Krahmer et al., 2003).

The performance of REG algorithms is usually evaluated by comparing them to a corpus of human-authored *reference* descriptions, the idea being that the quality of an algorithm is reflected by the extent to which its output matches that of humans on the same input domains (see, among others, Viethen & Dale, 2006; Gupta & Stent, 2005; Gatt et al., 2007; Gatt & Belz, 2010). From a comprehension perspective, *effectiveness* is also considered a conceivable criterion of adequacy of an algorithm's output (van Deemter, Gatt, van Gompel & Krahmer, 2012). Paraboni, van Deemter and Masthoff (2007) argue that referring expressions can be regarded as effective when they are easy for an addressee to comprehend and/or resolve.

As we have observed, considerations of computational efficiency, as well as psycholinguistic considerations have led to a relaxation of the Gricean Maxim of Quantity in determining how REG algorithms perform their task. Still, there remain several open questions concerning when an algorithm should indeed overspecify, and how. Part of the reason for this is that psycholinguistic results have to date provided only a partial picture of the processes underlying referential communication. For example, while a preference for certain attributes that results in increased likelihood of overspecification has been established, it is not clear whether speakers are also influenced by the type of domain they are considering (e.g., whether they are talking about items such as chairs and tables, or completely different entities such as human beings). Nor is it clear to what extent overspecification is influenced by the nature of the target referent, in particular, whether both singular and plural references are equally likely to be overspecified. For example, van Deemter's (2002) extension of the Incremental Algorithm to handle plurals was found to return highly complex expressions with a great degree of redundancy, leading some researchers to propose a return to strict Gricean principles for plurals (Gardent, 2002). Whether this is justified remains something of an open question (but see Gatt, 2007, for evidence that overspecification occurs in plurals as well).

Another issue that has not received much attention in the computational literature is the possible influence of different communicative settings on reference. Many of the algorithms discussed here are agnostic as to whether they are used in the context of an NLG system for text or speech, or whether the system is generating language in a situated setting or in a dialogue. There are some notable exceptions, such as Heeman and Hirst's (1995) computational interpretation of Clark & Wilkes-Gibbs's (1986) collaborative model for reference and, more recently, work in situated dialogue settings (Stoia et al., 2006; Kelleher & Kruijff, 2006; Byron et al., 2009). However, these computational approaches tend to have little to say about overspecification per se. An interesting exception is the work of Jordan (2000; 2002), whose investigation of redundancy in dialogues involving a joint task showed that interlocutors frequently produced descriptions that were overspecified in the context of dialogue, for instance, by repeating information that had already been used previously. Jordan argued that one reason for this could be that identification in the context of dialogues is often not the only goal of a referring expression. For example, a speaker may repeat information to signal agreement with an interlocutor. A computational implementation of this model showed that it matched human behaviour (as reflected in a dialogue corpus) better than some existing algorithms (Jordan & Walker, 2005). This effect of repetition is related to Pickering and Garrod's (2004) notion of *alignment*, which holds that the expressions that are produced earlier in the interaction influence the intonation patterns and syntactic structures of the ones that are produced later in the interaction. Goudbeek and Krahmer (2012) showed that alignment in referring expressions also occurs at the level of content selection (i.e., the target attributes that are included).

In summary, there are a number of open questions in pragmatics whose answers may help the development of better REG algorithms. One set of questions concerns the aspects of the domain that result in overspecification, and also how descriptions to singular and plural targets differ in terms of overspecification. Another concerns the difference between modality (speech vs. writing) and interactivity (monologue vs. dialogue settings). It is one of the motivations of the present study to contribute to answering these sets of questions.

Factors causing referential overspecification

In order to make REG algorithms described above perform better in terms of the amount of information that they include in generated output, some of the factors causing human speakers to produce overspecified references need to be further addressed and discussed. We therefore investigate two factors that we expect to cause referential overspecification: properties of the target and properties of the communicative setting.

Properties of the target referent. We expect referential overspecification to be influenced by the properties of the target referent. We hypothesize that targets that require more effort to refer to – in a sense to be made more precise below – will more often result in overspecified references than targets that are easy to refer to. This is in line with an observation by van der Sluis and Krahmer (2007). In a study looking at the production of deictic (pointing) gestures, they found the difficulty of the referential task to influence overspecification: in difficult tasks (i.e. pointing to far away targets), speakers used more words and included more locative relations in their descriptions compared to simple tasks. We investigate two kinds of properties of the target that may influence referential overspecification: the type of domain in which a target occurs, and the cardinality of the target (i.e. whether it consists of one referent or more than one).

For the *domain* in which a reference is produced, we hypothesize that expressions produced in domains where the objects vary more in terms of the number of potential attributes (and therefore afford speakers with more referential possibilities) will be more likely to contain more information than expressions produced in a domain where the speaker has less attributes to choose from. In order to investigate this, we made speakers refer to target objects in two different domains: one domain consisting of pictures of furniture items that vary in terms of a small number of attributes, and one domain in which entities are photographs of real people that differ on a higher number of attributes. We expect references to targets in the latter domain to be more frequently overspecified. By studying this, we aim to find out whether the specification level of target descriptions is affected by the referential possibilities that speakers have in a given domain, and thus, how different domains may lead to different specification levels. Clearly, this is not the only factor that determines the 'complexity' of a domain. For example, world knowledge will have an impact on what speakers choose to say about an entity. However, our hypothesis targets the specific issue of the amount of choice available to a speaker and the impact that this has on the ultimate decision of how to identify an object.

For *cardinality* (i.e. whether references are singular or plural), we hypothesize that plural references are more likely to be overspecified than singular references. More specifically, we expect that speakers are more likely to include redundant attributes in their references to two target objects than in their references to one target object. The reasoning behind this is twofold. First, there may be competition between the two target objects in plural references, which might force speakers to divide attention. As a result, the referring task could become more difficult, since dividing attention may lower the activation level that each of the two targets has in the speaker's mind. This explanation is in part based on findings by Arnold and Griffin (2007), who found evidence for an effect of cardinality on referential overspecification, albeit in a different context from ours, looking at when speakers use a pronoun. They report the results of two experiments demonstrating that speakers are less likely to use a pronoun when they need to describe two characters in a cartoon than when they need to describe a single character. In the present case, we expect that the effect of competition and divided attention may result in less control over the amount of information given about an entity in a definite description. A second reason why we expect plural references to be more frequently overspecified than singular references is related to attribute preference. As described earlier, studies by (among others) Pechmann (1989) and Belke and Meyer (2002) show that speakers tend to include preferred attributes (such as color) in references to one singular target, irrespective of whether these attributes have distinguishing value or not. This suggests that having more than one target may compound these effects.

Properties of the communicative setting. We also hypothesize that the properties of the communicative setting in which references are produced may cause overspecification. In this respect, we have two main hypotheses.

First, we hypothesize that spoken references are more frequently overspecified than written ones, thus containing more redundant attributes. This expectation is in line with the findings of Cohen (1984), who found that speakers provide more identification information than writers in instructive discourse. We expect similar results for the production of referring expressions, and the reason behind this is twofold. First, writing generally goes in the direction of an addressee that is not physically present, while speaking often involves a real, physically present addressee. Spoken expressions may therefore be more frequently overspecified, which would be in line with a study by Van der Wege (2009), who found instructions to an imaginary addressee to be more attenuated than instructions to a real addressee. Second, speaking, unlike writing, is an incremental process. This means that a speaker may start formulating a reference before he has fully planned what to say. We expect the same to hold in the case of spoken referring expressions, and that speakers start formulating a referring expression before having completed their scanning of the whole domain. Indeed, the incremental nature of speech production has been used to explain why language users often produce overspecified referring expressions (Pechmann, 1989). On the other hand, writers arguably are less likely to produce their expressions incrementally: they are more likely to plan the content of their references before starting to produce them, in part because a writer is typically under less time pressure than a speaker in a real-time conversation. Moreover, writers have the ability to edit the expressions they have produced, which allows them to correct their expressions for redundant attributes.

Our second hypothesis with respect to communicative setting is that speakers provide more redundant 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, meaning that he can rely on the addressee's flagging the need for further information should this arise. This would fit well with the principle of mutual responsibility proposed by Clark and Wilkes-Gibbs (1986), which says that language partners collaborate in order to establish the mutual belief that the listener has understood the speaker's reference. Clark and Wilkes-Gibbs based this principle on experimental findings showing that speakers' later references to an addressee tend to contain more definite descriptions (such as "the ice skater") than early references, and also that later references contain fewer words and required fewer turns per target than early references. Clark and Wilkes-Gibbs explain these findings by claiming that speaker and addressee collaborate by establishing conceptual pacts. Arguably, this collaboration becomes more difficult when language partners cannot see each other, simply because the speaker cannot receive visual feedback from the addressee. In the absence of such feedback, a speaker may be unsure whether his or her referring expression is sufficiently precise for the addressee to single out the target, and hence may be more inclined to add extra information. In this context, Eriksson's (2009) discussion of referential communication as an interplay between verbal and bodily practices is highly relevant. This implies that communication is less successful when the speaker and addressee cannot see each other. Further support comes from a study by Mol et al. (2009), who showed positive effects of mutual visibility on a speaker's gesture production, and, as a consequence, on the amount of information provided.

Experiment: which factors cause speakers to overspecify?

Below, we investigate to what extent properties of the target referent and properties of the communicative setting influence the amount of information included in referring expressions.

Method

The data for this study comes from a large-scale elicitation experiment with Dutch speakers, which yielded a corpus of referring expressions. The design and methodology was based on the TUNA corpus of referring expressions (Gatt, van der Sluis & van Deemter, 2007), which was collected through an elicitation study with English speakers, who were asked to refer to targets in visual domains consisting of objects arrayed on a screen. However, while the English TUNA corpus consists entirely of written descriptions, the corpus described here, D(utch)-TUNA, contains both spoken and written expressions and, in addition, manipulates a number of other factors related to communicative setting, which we describe below.

Materials

The materials consisted of forty trials, all of which contained one or two target referents and six distractor objects in a visual domain. The target referents were clearly marked by red borders, so that they could be easily distinguished from the distractor objects. For each participant and each trial, the target and distractor objects were positioned randomly on a screen in a partially filled 3 (row) by 5 (column) grid. In constructing the trials, two principal factors related to the properties of the target referent were manipulated, namely the type of domain and 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. For examples of trials in these domains, see Fig. 2.

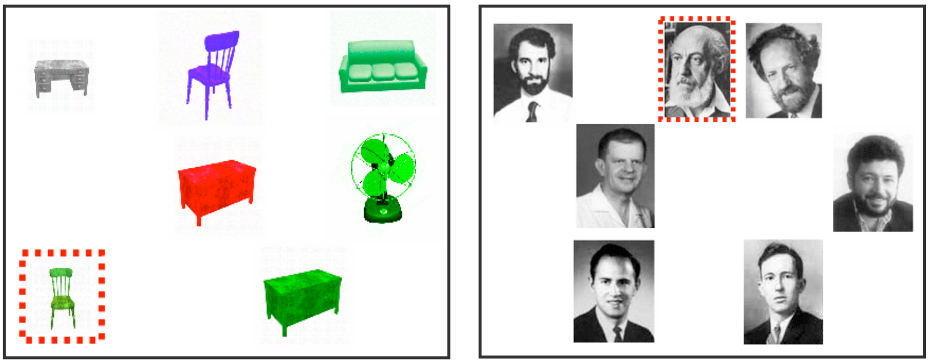


Fig. 2: Examples of (singular) trials in the furniture domain and the people domain.

In each case, the attributes or dimensions along which the objects were included in the trials were defined in advance, based on previous work on the construction of the English TUNA corpus (see also Gatt, van der Sluis & van Deemter, 2007). The twenty trials in the *furniture* domain contained artificially constructed pictures of four types of furniture items[1]. These items differed along four dimensions (see table 1).

Table 1: Attributes and values of the furniture items.

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

The twenty trials in the *people* domain consisted of pictures of males, which had been used before in an earlier study by van der Sluis and Krahmer (2007). There were several clear differences between the two domains[2]. First, since the pictures of people were real photographs, they were not as controlled as the artificial pictures in the furniture domain. Hence, there may be more information in them that participants could use in their references, making the set of possible descriptions of a target in this domain somewhat open-ended (in that many unpredictable attributes could be mentioned). However, the earlier study by Van der Sluis and Krahmer showed that speakers tend to make use of a subset of the available attributes much more frequently than others. Our estimate of the number of available attributes is based on this subset. Second, the target objects in the people domain could not be distinguished in terms of their type (since they are all male persons). Last, the pictures of the persons are arguably more

[1] The pictures were taken from the Object Databank, developed by Michael Tarr at Carnegie Mellon University and freely distributed. URL: http://www.tarrlab.org/

[2] We are aware of the fact that these motivations are specific to our stimulus material, and that they cannot be applied to all kinds of pictures of furniture items and people. We will elaborate on this in the discussion of this chapter.

perceptually similar to each other than the furniture items, which might make them more difficult to distinguish from the distractor objects.

As in the furniture domain, a number of salient dimensions of variation were identified in the people domain, based on earlier work by van der Sluis and Krahmer (2007) and van der Sluis et al. (2007). These dimensions are shown in table 2.

Table 2: Attributes and values of the people pictures.

Attribute	**Possible values**
Type	Person
Orientation	Front, left, right
Age	Young, old
Hair color	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

As in the original TUNA experiment, the construction of trials was such that, for each possible combination of the attributes in tables 1 and 2, there was one trial in the relevant domain where that combination was minimally required in order to distinguish the referent. For example, there was a furniture trial in which the target could be distinguished using color only, another one in which the target could be distinguished using a combination of color and orientation, and so on. Since speakers always need to include a head noun in their references and therefore tend to 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 (for example, there was no furniture trial in which a target could be uniquely distinguished by the description "The chair"). However, an attempt was made to balance the proportion of trials in the furniture domain in which objects of different types (e.g., chair and sofa) were used as target referents.

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 (ten per domain), containing one target referent; the remaining twenty (again ten per domain) were plural trials containing two targets. Fig. 2 shows examples of singular trials in the two domains, while Fig. 3 shows examples of plural trials.

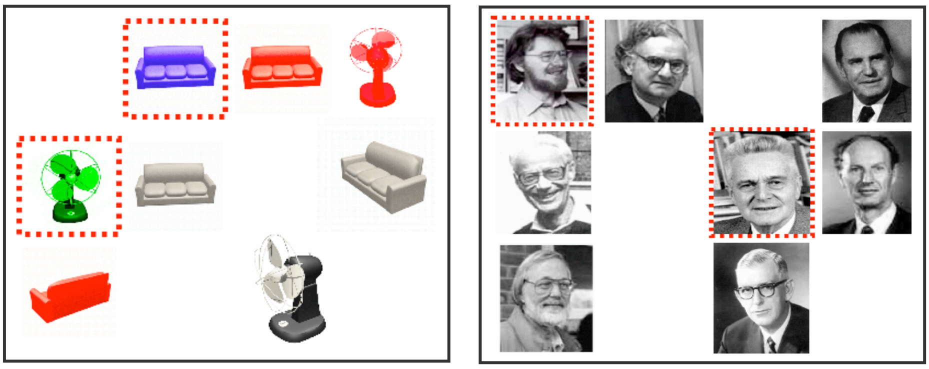


Fig. 3: Examples of plural trials in the furniture domain and the people domain.

These kinds of plural trials will generally result in participants producing semantic plurals (Schwarzschild, 1996): an example of a description could be "The brown chair and the grey desk"). Since plural references can be produced in different ways, we included two kinds of plural trials. Half of the trials (five per domain) contained two target objects with identical values for their distinguishing attributes. For example, both targets might be red, and color was the distinguishing attribute, so that a description like "The table and the sofa that are both red" could distinguish them. The other half (again five per domain) contained two target objects with different values on their distinguishing attributes. For example, they could consist of a fan and a sofa, both of which needed to be distinguished via their size attribute, but were of different sizes. An example of a distinguishing description of such a trial would be "The large fan and the small sofa".

Participants

Sixty undergraduate students (14 males, 46 females) from Tilburg University participated in the experiment. All participants (mean age 20.6 years old, range 18-27 years old) were native speakers of Dutch.

Procedure

Each participant was presented with the same forty trials in a different randomized order. The experiments were individually performed in an experimental room, with an average running time of twenty minutes. The participants could take as much time as they needed to describe the pictures. All participants were filmed during the experiment, mainly in order to capture their speech.

The participants were asked to describe the target referents in such a way that their addressee could uniquely identify them. In order to manipulate properties of the communicative setting, the participants were randomly assigned to one of 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 the target referents. No addressee was present, but the participants were told that their descriptions would be 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. This was done to enable a focus on content planning of initial descriptions ('first mentions'), and to make descriptions comparable across conditions. 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 information in their descriptions. 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 Fig. 4.

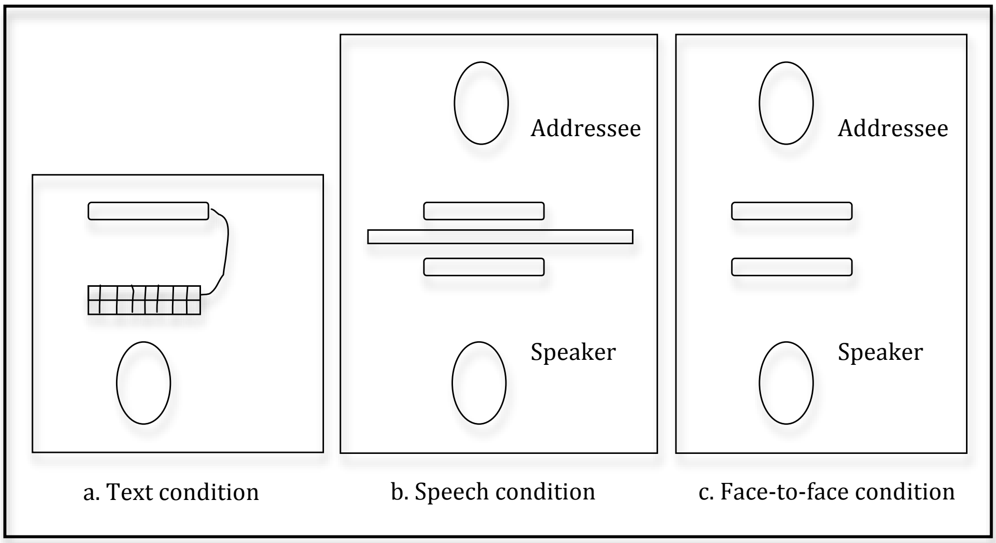


Fig. 4: Schematic overview of the three experimental conditions.

Our procedure differed from that used in the original TUNA experiment in three ways. First, the present study used a laboratory-based setup, whereas the original TUNA study was conducted online in a relatively uncontrolled setting. Second, as noted above, the original study only had the text condition, and participants were told that they would be writing descriptions that would be resolved in real time by an automatic language understanding system, rather than by an addressee who was in another room. Third, one of the between-groups factors manipulated in the English-language version was whether or not participants could refer to objects using their location on the screen, whereas locative descriptions were excluded from D-TUNA.

Data annotation

The 2400 (60 participants x 40 trials) identifying descriptions of the D-TUNA corpus were all semantically annotated in XML format, and paired with a full representation of the objects in a trial, consisting of the attributes of both the target and distractor objects. We used the annotation scheme of the TUNA corpus (Gatt, van der Sluis & van Deemter, 2008b), and carried out the annotation with the annotation tool Callisto[3]. An example of an XML representation of the people trial shown in Fig. 2 is depicted in Fig. 5 on the next page, together with one of the references to the target in the corpus. In this expression, the target is referred to (in Dutch) as *"De man met de witte baard en bril"* (meaning "The man with the white beard and glasses").

All 2400 files in the corpus consist of a TRIAL node, containing a trial ID and detailing the specific conditions under which the expression was produced (such as domain, modality and cardinality). Furthermore, each TRIAL node subsumes four nodes: a DOMAIN node, a STRING-DESCRIPTION node, a DESCRIPTION node and an ATTRIBUTE-SET node.

The DOMAIN node contains a representation of the domain of the particular trial and consists of seven or eight ENTITY nodes: one or two target entities (depending on cardinality) and six distractor entities. Each entity node lists the properties of the particular entity. The STRING-DESCRIPTION node contains the full target description, as produced by the participant. The DESCRIPTION node contains the annotated version of the target description. All determiners and content words that are part of the string description were marked up with the attributes that they represent. For example, the adjective *"witte"* (meaning "white") corresponds to the attribute <hair color = light>. In case a participant mentioned an attribute that was not predefined for the domain at all (e.g., "The laughing man"), that attribute was annotated as <other = other>. The ATTRIBUTE-SET contains an overview of all the properties (attributes and their values) that are mentioned in the string description and thus represents the "flat" semantic structure of the referring expression.

[3] See http://callisto.mitre.org/

```
<TRIAL ID="A03t21" CARDINALITY="1" CONDITION="text" DOMAIN="people"
   MODALITY="written">
  <DOMAIN>
       <ENTITY ID="54" IMAGE="Eilenberg.jpg" TYPE="target">
            <ATTRIBUTE NAME="hasBeard" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="hasTie" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="type" TYPE="literal" VALUE="person"/>
            <ATTRIBUTE NAME="hasHair" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="hasGlasses" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="hasSuit" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="age" TYPE="literal" VALUE="old"/>
            <ATTRIBUTE NAME="hairColor" TYPE="literal" VALUE="light"/>
            <ATTRIBUTE NAME="orientation" TYPE="literal" VALUE="left"/>
            <ATTRIBUTE NAME="hasShirt" TYPE="boolean" VALUE="1"/>
       </ENTITY>
            <ENTITY ID="4" IMAGE="Fefferman.jpg" TYPE="distractor">
            <ATTRIBUTE NAME="hasBeard" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="hasTie" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="type" TYPE="literal" VALUE="person"/>
            <ATTRIBUTE NAME="hasHair" TYPE="boolean" VALUE="1"/>
            <ATTRIBUTE NAME="hasSuit" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="hasGlasses" TYPE="boolean" VALUE="0"/>
            <ATTRIBUTE NAME="age" TYPE="literal" VALUE="young"/>
            <ATTRIBUTE NAME="hairColor" TYPE="literal" VALUE="dark"/>
            <ATTRIBUTE NAME="orientation" TYPE="literal" VALUE="front"/>
            <ATTRIBUTE NAME="hasShirt" TYPE="boolean" VALUE="1"/>
         . . . . .
   </DOMAIN>
   <STRING-DESCRIPTION>
         De man met een witte baard en zonder bril.
   </STRING-DESCRIPTION>
   <DESCRIPTION NUM="singular">
       <DET VALUE="definite">De</DET>
       <ATTRIBUTE ID="a1" NAME="type" VALUE="person">man</ATTRIBUTE>
                   met
       <ATTRIBUTE ID="a3" NAME="hasBeard" VALUE="1">een<ATTRIBUTE ID="a2"
         NAME=”hairColor” VALUE=”light”>witte</ATTRIBUTE>baard</ATTRIBUTE>
                   en
       <ATTRIBUTE ID="a4" NAME="hasGlasses" VALUE="0">bril</ATTRIBUTE>
   </DESCRIPTION>
   <ATTRIBUTE-SET>
       <ATTRIBUTE ID="a1" NAME="type" VALUE="person"></ATTRIBUTE>
       <ATTRIBUTE ID="a2" NAME="hairColor" VALUE="light"></ATTRIBUTE>
       <ATTRIBUTE ID="a3" NAME="hasBeard" VALUE="1"></ATTRIBUTE>
       <ATTRIBUTE ID="a4" NAME="hasGlasses" VALUE="0"></ATTRIBUTE>
   </ATTRIBUTE-SET>
</TRIAL>
```

Fig. 5: Example of an XML file of a reference in the people domain.

Annotating the descriptions of the D-TUNA corpus in the XML annotation scheme of the English TUNA corpus is advantageous for several reasons. Firstly, it makes the English and Dutch TUNA corpora highly comparable. Secondly, it makes the D-TUNA corpus a useful tool for the evaluation of algorithms that aim to automatically generate referring expressions (REG algorithms), because the annotation is machine-readable and provides semantic information. Lastly, our annotation facilitates a corpus-based analysis of referential overspecification, since it explicitly marks up the semantic/conceptual content of object descriptions, abstracting away from individual variation in the way these are realised syntactically.

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 regard the *number of redundant attributes* as a dependent variable indicating the amount of overspecification in referring expressions. An attribute is considered to be redundant if removing it from the description would still result in a distinguishing reference. For example, consider Fig. 2 once again. 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 redundant attributes. Since trials were built in such a way that type could never be a distinguishing attribute, we excluded type from our analysis of attributes.

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	Total number of attributes	Number of redundant attributes
Underspecified	"The man with the beard."	1	-1
Minimally specified	"The man with the white beard."	2	0
Minimally specified	"The white bearded man."	2	0
Overspecified	"The white bearded man without a tie."	3	1

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 express two attributes (<has beard =1> and <hair color = light>) that are both needed for identification of the target object. Therefore, no redundant attributes are counted in that case. Finally, 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>). Note that, as this last example illustrates, attributes can sometimes specify what is absent in the picture.

Our statistical procedure consisted of mixed-model repeated measures ANOVAs, and Tukey HSD tests for post-hoc multiple comparisons. We report on main effects and interactions where these are significant.

Results

The overall proportions of minimal, overspecified, and underspecified descriptions confirmed the finding of various psycholinguistic studies that speakers tend to overspecify their referring expressions: 53.6% of the references were overspecified, as compared to 41.4% minimally specified references. A small minority of the references (5.0%) was underspecified. Given the fact that referential underspecification was rare in the sample (especially compared to the rates of overspecification) and did not result in significant differences on any of the relevant factors, we decided not to further address this topic here.

More detailed analyses of the data indicated at least some factors that were responsible for the occurrence of referential overspecification: properties of the target referent and properties of the communicative setting. Table 5 depicts the descriptive statistics (means and standard deviations) of all the conditions analyzed.

Table 5: The number of redundant attributes (means and standard deviations) as a function of all the conditions analyzed.

	Domain		Cardinality		Comm. setting		
	Furn.	People	Sing.	Plur.	Text	Speech	F-to-F
Redundant attributes	.6 (.03)	1.6 (.15)	0.9 (.07)	1.3 (.10)	0.9 (.14)	1.3 (.14)	1.1 (.14)

Results for properties of the target

Domain. First, the results show that domain complexity can be regarded as a factor causing referential overspecification. Fig. 6 depicts the average number of redundant attributes as a function of domain.

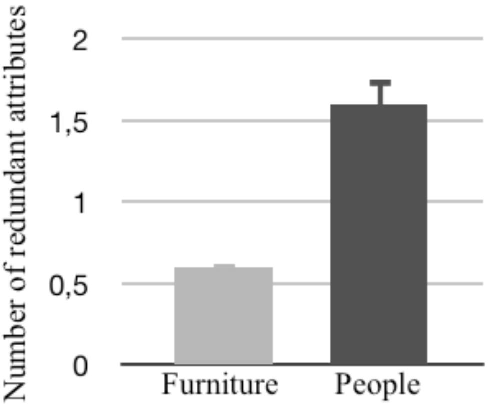


Fig. 6: Average number of redundant attributes as a function of domain.

Fig. 6 shows that references to people did contain significantly more redundant attributes (M = 1.6, SD = .15) than references to furniture items (M = .6, SD = .03), $F_{(1,57)}$ = 54.12, $p < .001$, $\eta^2 = .49$. These results suggest that speakers are more likely to overspecify their references when they refer to targets that occur in a domain where the objects differ along a relatively high number of dimensions.

Cardinality. A second factor that was hypothesized to influence referential overspecification is cardinality. We first checked whether the two different kinds of plural trials differed in terms of referential overspecification, but no significant differences were found.

Fig. 7 shows the average number of redundant attributes as a function of cardinality.

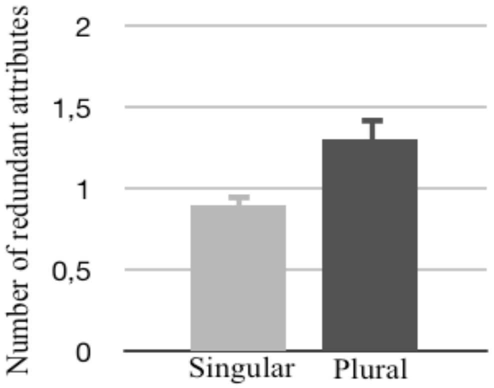


Fig. 7: Average number of redundant attributes as a function of cardinality.

Our hypothesis that references to two target objects should contain more redundant attributes than references to one target was confirmed by the significant difference between the number of redundant attributes that were mentioned in singular references (M = .9, SD = .07) and plural references (M = 1.3, SD = .10), $F_{(1,57)}$ = 44.27, $p < .001$, $\eta^2 = .44$. These results indicate that speakers overspecify their references more frequently when they need to refer to plural targets.

The effect of cardinality described above was stronger in the more complex people domain, as reflected in interactions between domain and cardinality for the number of redundant attributes. That is, compared to the furniture domain, the effect of cardinality on the number of redundant attributes that the references contained was stronger in the people domain ($F_{(1,57)}$ = 30.61, p < .001, η^2 = .35). In order to study this interaction in more detail, we conducted two separate statistical analyses to study whether the effect of cardinality was significant in both domains. In the furniture domain, we found an effect of cardinality ($\mathrm{F}_{(1,57)}$ = 4.27, p < .05, η^2 = .07), meaning that descriptions of two furniture items (M = .64*,* *SD* = .05) contained significantly more redundant target attributes than descriptions of one furniture item (M = .55, *SD* = .03). We found a similar (but somewhat stronger) effect of cardinality in the people domain ($\mathrm{F}_{(1,57)}$ = 42.98, p < .001, η^2 = .43), showing that descriptions of two persons (M = 2.0, SD = .19) contained significantly more redundant target attributes than target descriptions of one person (M = 1.2, *SD* = .12).

Results for properties of the communicative setting

Besides effects of properties of the target, we also looked at the influence of communicative setting. Fig. 8 displays the average number of redundant attributes as a function of the three communicative conditions.

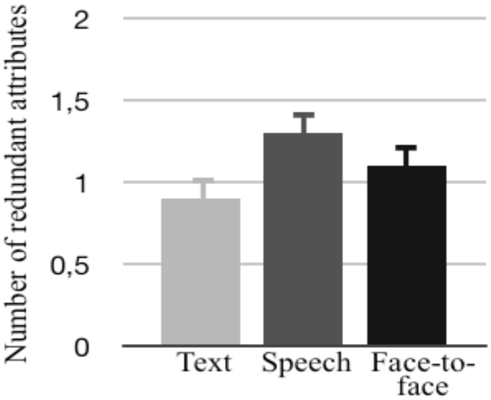


Fig. 8: Average number of redundant attributes as a function of communicative setting.

Although Fig. 8 shows differences in the average number of redundant attributes that were mentioned in the three respective conditions, suggesting that speakers in the speech condition (M = 1.3, SD = .14) were more likely to include redundant attributes in their descriptions compared to speakers in the face-to-face condition (M = 1.1, *SD* = .14) and to writers in the text condition (M = .9, *SD* = .14), these differences did not reach significance ($F_{(2,57)} = 1.61$, *ns*, $\eta^2 = .05$).

Summary

In the above experiment, we have investigated whether properties of the target and properties of the communicative setting influence speakers in terms of the amount of redundant target attributes that they include in their referring expressions.

Properties of the target serve as a first factor causing referential overspecification. First, we found an effect of domain, which suggests that references to targets that occur in domains in which objects afford more referential possibilities (due to there being more attributes available) contain significantly more redundant attributes than references to targets that occur in domains where speakers have fewer referential possibilities. Second, we found effects of cardinality, which indicate that references to plural targets contain significantly more redundant attributes than references to singular targets. These effects of cardinality were present in both the furniture and the people domain.

The results also show that written and spoken referring expressions do not differ in terms of overspecification: speakers do not include significantly more redundant attributes in their descriptions than people who type their expressions. The results also show that people who cannot see the addressee do not use lengthier descriptions (in terms of the number of redundant attributes) than people who can, which means that a lack of visual feedback does not directly cause speakers to overspecify their referring expressions in this type of communicative context.

General discussion

What makes speakers include redundant information in their referring expressions? In the current chapter, we have presented a quantitative study on factors that might influence referential overspecification. The results have revealed several findings. In general, we found around 50% of the speakers' referring expressions to contain more information than needed for identification of the target, which is in line with previous studies on referential overspecification (e.g., Engelhardt et al., 2006; Maes et al., 2004; Pechmann, 1989).

Factors causing referential overspecification

We first described a language production experiment in which we collected the D-TUNA corpus, which consists of 2400 Dutch referring expressions. These were all definite noun phrases, describing furniture items and people. We used this corpus to investigate to what extent properties of the target and properties of the communicative setting cause such definite references to be overspecificied.

Properties of the target referent

In line with our hypotheses, we found that properties of the target influence referential overspecification.

Domain. We have found that references to furniture items were less likely to be overspecified than references to people. These results confirm our hypothesis that the number of choices available to a speaker in the domain in which a target referent occurs affects the amount of overspecification, and that different domains can thus lead to different specification levels. This finding has at least two implications.

First, the effect of domain on referential overspecification suggests that when the range of attributes to describe a target referent gets broader (as was the case with the target objects in the people domain), speakers include more information in their target descriptions and provide the addressee with more information to single out the target. The precise effect of the number of choices available to a speaker could be studied more accurately by manipulating the number of possible attributes that is available for object descriptions within a single domain. The results of such an experiment will be presented in Chapter 4 of this dissertation.

Earlier in this chapter, we have described several differences between the furniture and the people domain. For one thing, contrary to the furniture domain, the type in the people domain was constant, which may have caused the participants to encode each people referent with an attribute anyway. One related difference was that the pictures in the people domain are more perceptually similar, which leads us to a second implication of the effect of domain on overspecification: it implies that perceptual similarity is a potentially strong indicator for the complexity of a given domain. This would mean that the complexity of a domain and the ease with which certain variables can be accessed are closely related. More specifically, it suggests that as the difficulty of selecting the discriminatory attributes increases, the speaker tends to mention more (possibly redundant) attributes in order to be sure that the addressee can identify the target referent.

It needs to be emphasized that the motivations for why we consider the furniture domain and the people domain to be different are specific to our stimulus material: they cannot be generally applied to all possible kinds of pictures of furniture items and people. For example, the situation can be different when a speaker wants to refer to people that he or she is familiar with (such as family members of friends, or Hollywood stars), or to furniture items that are all similar-looking. Furthermore, it should be noted that the furniture items could have colors that are atypical of the noun that was modified (e.g., one would normally not expect to see a bright green desk or fan). Since color typicality has been shown to influence reference production (Sedivy, 2003; Westerbeek, Koolen, & Maes, 2013), this may have influenced the speakers in our study as well.

Cardinality. Regarding cardinality, we would like to emphasize that the focus in this chapter is on semantic plurality (Schwarzschild, 1996): rather than morphological plurals (e.g., "The brown chairs"), participants generally produced semantic plurals (e.g., "The brown chair and the grey desk"). In line with our hypothesis, we have found that plural target descriptions are more frequently overspecified than singular target descriptions. In the introduction of this chapter, we formulated two arguments that may explain the effect of cardinality on overspecification.

First, there may be competition between the two target objects, which forces speakers to divide attention. This may make the referring task more difficult, since dividing attention lowers the activation level that each of the two targets has in the speaker's mind. Such lower activation levels may cause speakers to have difficulties in avoiding redundancy, because redundancy avoidance requires comparison between a target and its distractors, and the target has to be in the focus of attention in order for this to take place successfully. While our results show that speakers often do not carry out this process of comparison exhaustively for singular targets (as shown by the high proportion of overspecified singular descriptions), these observations would also explain why this tendency appears to be even greater for plurals. This is in line with the work of Arnold and Griffin (2007), who demonstrated that plural references are more likely to contain specific forms (such as descriptions) instead of pronouns, and therefore are less attenuated than singular references. Our findings extend this observation further by highlighting another sense in which plural references are more specific than singular ones, namely in terms of the number of (redundant) attributes they contain.

Second, when referring, speakers prefer certain attributes to others. As we have observed in the introduction, several psycholinguistic studies use attribute preference as an argument for the occurrence of referential overspecification (e.g., Pechmann, 1989; Belke & Meyer, 2002). These studies showed that speakers tend to include preferred attributes (such as color) when referring to one singular target, irrespective of whether these attributes have contrastive value or not. Our findings suggest that this tendency to include preferred attributes is even greater when speakers refer to two target referents. This suggestion is in line with Gatt and van

Deemter's work (2007), which shows that speakers tend to conceptualize two target objects in a parallel fashion, that is, if they describe one of two target referent using color, they are bound to do the same for the other one. In the end, this leads to overspecification. For example, the plural reference "The red desk facing left and the green chair facing right" contains color twice as a preferred but redundant attribute, and is therefore more redundant than the singular reference "The red desk facing left" (which contains color as a preferred but redundant attribute only once).

Properties of the communicative setting

The results fail to confirm our hypotheses on the influence of the communicative setting on the redundancy level of referring expressions.

Although we have signalled a numeric trend that spoken referring expressions contain more redundant attributes than written ones, there are no significant differences between the written and the spoken conditions in terms of redundancy. This finding contrasts with previous research. Based on Cohen (1984), who found speakers to provide more identification information than writers in instruction-giving discourse, we expected this to also count for referring expressions, in the sense that written and spoken references would differ in terms of their number of redundant attributes. We further based our expectations on Van der Wege (2009), who found instructions to an imaginary addressee to be more attenuated than instructions to a real addressee. However, our results fail to confirm these expectations. One possible explanation may be that in the two spoken conditions of our experiment, the communication between speaker and addressee was rather one-sided and thus not very interactive: It was the task of the addressee to only react briefly when he had singled out the target. This lack of interaction may have reduced the role of the addressee in the two spoken conditions, in the end making it comparable to the (marginal) role of the imaginary addressee in the written condition. Some support for this interpretation comes from studies by Jordan (2000, 2002), who found that interlocutors in a collaborative dialogue task were likely to produce redundant references (including attributes which are repeated across turns, even though the referent has been established in the joint focus of attention), in order to achieve other goals in addition to identification. For example, a speaker may redundantly use the attribute color in “The red sofa” in response to her interlocutor’s suggestion to “buy the red sofa”, in order to signal agreement with her interlocutor about the proposal.

Even though speakers do not use more redundant attributes than writers (contrary to our expectations), they do use more words. More precisely, speakers in the speech condition (M = 15.6) and face-to-face condition (M = 15.7) use almost twice as many words than people in the text condition (M = 8.7). We believe that the physical presence of an addressee, as well as the incremental nature of speech production indeed causes this difference between speaking and writing. However, this did not result in mentioning more redundant attributes, which suggests that the number of words and the numbers of redundant attributes are not directly linked. An explanation for this is that speakers, more than writers, tend to repeat attributes, which may arise from disfluencies. For example, repetition of attributes in expressions such as “The red… red… chair” obviously increases the number of words used. This explanation is confirmed by the frequencies of repeated attributes in references produced in the three communicative settings. These show that spoken references in the speech and face-to-face condition contained 205 and 214 repeated attributes respectively, while written references in the text condition contained only 20 repeated attributes.

Although we have also signalled a numeric trend that speakers who cannot see an addressee provide more information than speakers who can, the referring expressions uttered in the two respective conditions did not differ in terms of the number of redundant attributes that were mentioned. This result implies that a lack of visual feedback does not lead to more referential overspecification, which contrasts with our hypothesis and with previous research (e.g., with the principle of mutual visibility by Clark and Wilkes-Gibbs, 1986). We believe that the lack of direct interaction between speaker and addressee may also explain this apparent contradiction, since it may have reduced the role of the addressee in both conditions. In future research, we aim to investigate whether this is indeed the case.

Given the lack of significant differences that we have found regarding the three communicative conditions, some concern with respect to the naturalness of our experimental setting needs to be expressed. As we have explained above, it seems that the participants in our experiment seemed not so much focused on their addressee, but more on their primary task: they had to distinguish the target referent from the contrast set of distractors. However, in naturalistic dialogue, as Jordan and Walker (2005) argue, speakers are not solely guided by this contrast set when selecting the content of their target descriptions; also other factors might play a role. First, Jordan and Walker stress the importance of intentional factors, where they explain that the attributes included in a description might have other communicative purposes than target identification alone. Furthermore, Jordan and Walker emphasize the role of the addressee in conversation, where they argue that also conceptual pact factors are important in a speaker's attribute selection process (they base themselves on papers by Brennan and Clark (1996) and Clark and Wilkes-Gibbs (1986) here). Since also major pragmatic theories take the role of the addressee in conversation as a starting point, we will discuss the implications of our results for these theories below.

Implications of referential overspecification for pragmatic theory

One strict interpretation of the Maxim of Quantity (Grice, 1975) is that a speaker should provide just enough information for the addressee to identify a target referent (Quantity 1), but not more (Quantity 2), and that speakers should ideally produce minimally distinguishing target descriptions. However, in the beginning of this chapter, we have suggested that considering overspecification as a violation of the Maxim of Quantity would be an over-simplification: the information level of expressions generally depends on the purpose that a speaker has in uttering it. Still, several pragmatic theories on reference (which we will discuss below) argue that overspecification can have a negative effect on how an addressee comprehends an expression.

As we have seen in the introduction of this chapter, one can draw a distinction between speaker-driven and addressee-oriented factors that may cause speakers to overspecify (Arnold, 2008). Although the Maxim of Quantity intuitively seems to focus on what is said and meant by the speaker in a conversation, Grice himself makes explicit that the maxims apply to both the speaker and the addressee (Grice, 1989, p. 31). This perspective is also taken by neo-Griceans such as Horn and Levinson. Horn (1984; 2005) mirrors the two components of the Maxim of Quantity by introducing the Q-Principle (that is addressee-oriented) and the R-Principle (that is speaker-oriented). Levinson (2000) also stresses the role of the addressee by stating that violations of the Maxim of Quantity (i.e. in case of overspecification) should be viewed as inferences drawn by the addressee. This suggests that providing redundant information could negatively affect the addressee’s comprehension process: they may lead to false implicatures, or require unnecessary effort for the addressee to process the utterance.

Also Relevance theory (Carston, 1991; Sperber & Wilson, 1986; Wilson & Sperber, 1981) suggests that referential overspecification might have a negative effect on comprehension. According to this theory, the speaker aims to tailor an utterance so that the addressee’s search for relevance will reach the intended interpretation. Hence, utterance comprehension should be interpreted in a psychologically plausible way, implying that an addressee always aims to arrive at an interpretation of an expression without spending much (needless) effort. In this view, expressions that meet this aim are considered ‘optimally relevant’. Arguably, referring expressions that contain redundant information may not be labeled as such, since they may require unnecessary effort for the addressee to comprehend them.

In this chapter, we have looked at various factors that may influence referential overspecification. We have seen that both domain properties (furniture vs. people) and target properties (singular vs. plural) may cause speakers to include more redundant properties in their referring expressions. It seems likely that this is primarily a speaker-driven process, as we have argued. In this study, we did not look explicitly at addressee-oriented aspects of referring, but it would be interesting to see how addressees process these overspecified references. As we have noted earlier, how sensitive addressees are to overspecified information remains somewhat unclear. According to Grice and his followers, overspecified information may trigger false implicatures (see also Dale and Reiter, 1995). When hearing a reference to a table as ``the brown wooden table", in a situation where there is only one table, an addressee might infer that the fact that the table is brown and wooden is somehow significant; why else would the speaker mention these attributes? At the same time, it seems perhaps unlikely that the additional overspecification in the case of references to people or to plural targets, creates more false implicatures, but we are not aware of any studies that directly test this.

Interestingly, we found no effects of communicative setting on the amount of overspecification. Whether references where typed or spoken did not result in significant differences in terms of the number of redundant attributes, nor did it matter whether the participant could see the addressee or nor. Although we have suggested that this result may be due to a lack of direct interaction between speakers and addressees, it suggests that a speaker-oriented factor is involved here as well. If addressee-oriented factors would be dominant (which pragmatic theories such as Relevance theory seem to suggest), one would expect to find differences between the numbers of redundant attributes that the expressions contain in the three conditions.

Overspecification in Referring Expression Generation

As discussed at the beginning of the chapter, the present chapter was partly motivated by some open questions related to computational models of Referring Expression Generation (REG). We turn once again to this topic below, and discuss some points of convergence and divergence between the terms of the present study and the computational literature, as well as the implications that our results may have for REG algorithms.

Definitions of redundancy. As observed earlier, concerns about referential overspecification and redundancy, and the tension between these and computational efficiency, have been at the heart of the development of REG algorithms. The definition of "redundant attribute" used in this chapter – that is, an attribute that is redundant in relation to the rest of the description – has a precedent in the computational REG literature in Reiter's (1990) definition of *Local Brevity*. Reiter notes that a definition of redundancy as incorporated in Dale's (1989) *Full Brevity* algorithm results in a computational procedure that is intractable. Since this definition focuses on finding the smallest set of attributes that distinguish a target, it leads in the worst case to an exhaustive search through the space of all possible descriptions, starting with the shortest. Reiter's Local Brevity alternative avoids this by generating a distinguishing description and then checking whether it contains any attributes that could be removed without sacrificing identification. If this is not possible, the description is regarded as minimally specified.

The difference between the two can be illustrated through an example. Consider a domain in which "The bearded man with glasses" and "The man with the tie" both contain just enough information to make the target identifiable. That is, we assume that removing the attribute "has beard" from the first description would result in a non-identifying description, as would removing "has glasses". It is only in conjunction that these attributes identify the object. Similarly, we assume that "has tie" is required in the second description. In such a case, the Full Brevity algorithm would prefer the second description, since it contains only one attribute (besides type), and therefore is the shortest description of the two. Therefore, under this definition, "The bearded man with glasses" is not strictly minimal, since there still exist shorter descriptions (such as "The man with the tie"). By contrast, under a Local Brevity assumption, both of these would qualify as "non-redundant" descriptions, since neither contains attributes that can be removed while satisfying the identification goal.

Clearly, there will be cases where the two definitions converge, but it is worth noting that from a computational perspective, only the Local Brevity assumption is tractable in the worst case. The definition of redundant attribute used in this chapter is in line with the Local Brevity assumption, although separate analyses of the same data show that defining referential overspecification in line with Full Brevity also evinces similar trends.

The above discussion should not of course be taken to imply that either of the two algorithmic interpretations of redundancy discussed here is psycholinguistically plausible, but this does not hamper the impact of this study, which does rely not on any such algorithmic interpretation. For one thing, neither of these two algorithms is incremental, and incrementality is one of the core aspects of psycholinguistic models that seek to explain referential overspecification (e.g., Pechmann, 1989). As discussed earlier, incremental REG procedures have since been proposed – notably by Dale and Reiter (1995) – and have been claimed to provide a better approximation to what people do, in part because their output may be overspecified. It is therefore worth exploring the possible implications that our findings have for such algorithms.

Computational implications of overspecification. Some psycholinguistic research shows that referential overspecification can be beneficial to readers or listeners, since it facilitates the target identification process (e.g., Arts et al., 2011). This would suggest that automatically generated expressions containing redundant attributes would be easier to resolve. For example, recent work by Paraboni, Masthoff and van Deemter (2006) shows that an optimal level of overspecification can help readers to resolve references to parts of documents. On the other hand, current REG algorithms, including the Incremental Algorithm of Dale & Reiter (1995), do not systematically generate overspecified expressions. For example, the Incremental Algorithm discussed earlier relies on an a priori specification of a preference ordering of attributes, but will only add an attribute (even one which is highly preferred) if it has discriminatory value, that is, if it excludes some of the remaining distractors of the target. Therefore, this algorithm will not use color for a target in a domain in which all other objects have the same color. Moreover, also other algorithms that overspecify deliberately have been designed, such as the ones by Paraboni et al. (2007) and Zender, Kruijff and Kruijff-Korbayová (2009), but these do not systematically address factors related to domain, cardinality and communicative setting as manipulated here. So how does the present chapter contribute to improving the performance of REG algorithms in the amount of information they should include in their referring expressions?

On the basis of our empirical findings, we can formulate several important implications for automatic Referring Expression Generation. First, REG algorithms should include more redundant information when referring to more complex target referents, based on effects of domain and cardinality on referential overspecification in our human corpus data. More specifically, this means that automatically generated referring expressions to two target referents that are uttered in a complex domain should contain more redundant information than those to one target referent that are uttered in a simpler domain. Second, our findings suggest that modality (text or speech) and whether or not a speaker has the possibility to give visual feedback should not influence the redundancy level of automatically generated referring expressions (but see the caveats raised earlier).

How can REG algorithms determine which attributes should be redundantly added to their referring expressions? Based on Maes et al. (2004), we recommend that the kinds of properties that are used to describe a target should be of a perceptual nature. In particular, speakers tend to include perceptually salient target attributes (such as color) in their references, because both speaker and addressee easily perceive such attributes (Belke & Meyer, 2002). Since Arts (2004) and Arts et al. (2011) found that locative expressions also facilitate object identification, presumably because they help to orient attention to the right location in space, it would seem to make sense to redundantly add location information to automatically generated referring expressions in such spatial domains. Stipulating that algorithms should include perceptually salient or spatially useful attributes even when they are redundant has some implications for algorithm design. In particular, it would seem that a strategy such as that incorporated in the Incremental Algorithm, where salient attributes are only included if they have some discriminatory value, might not always return results which are compatible with what humans would do in the same situation. On the other hand, it is plausible to assume that humans exhibit a degree of non-determinism in their referential behaviour, so that the inclusion of, say color in a domain where all objects are of the same color, may be probabilistic.

As described earlier, evaluation of REG algorithms is not only a case of aiming to emulate human referring behaviour, but also to produce *effective* referring expressions (van Deemter et al., 2010), that is, expressions which are easy for an addressee to comprehend and/or resolve (Paraboni et al., 2007). Previous research has shown that including redundant information in a referring expression facilitates object identification (e.g., Arts et al., 2011; Deutsch, 1976; Nadig & Sedivy, 2002; Engelhardt et al., 2006; Paraboni et al. 2006). This implies that REG algorithms might include redundant information in their referring expressions, because this makes them easier to comprehend for addressees. But what level of redundancy is optimal in any given situation? Arguably, pragmatic factors related to the communicative setting are likely to affect the amount of redundant information (van Deemter et al., 2007). On the one hand, there may be fault-critical settings, where accurate understanding is crucial; On the other hand, there may be situations in which the reader of listener's ease of comprehension is not at the forefront of the speaker or writer's attention, for instance when the speaker or writer is under time pressure. It stands to reason that redundancy is more important in the former case.

Conclusion

The current study addresses several factors that cause speakers to overspecify their referring expressions, or, more specifically, the definite target descriptions that they produce. The results of our language production experiment show that the domain in which a target occurs, and whether this target is singular or plural, determines the amount of redundant information that language users include in their definite descriptions. However, the results do not show an effect of communicative setting (concerning the modality in which an expression is produced, and the speaker's possibility to receive visual feedback from the addressee) on overspecification. This was possibly due to a lack of natural interaction between speaker and addressee. Since current Referring Expression Generation (REG) algorithms are not able to deal with referential overspecification in a systematic way, the above pragmatic findings are helpful in improving these algorithms in terms of the amount of information that they should include in their referring expressions.

Acknowledgments

We thank three anonymous reviewers, Anja Arts, Fons Maes and Marc Swerts for their helpful comments concerning earlier drafts of this chapter, and Carel van Wijk for his help in the statistical analyses. The D-TUNA corpus is publicly available via the TST-Centrale (Dutch HLT Agency; http://www.inl.nl/nl/corpora/d-tuna-corpus/) of the Nederlandse Taalunie (Dutch Language Union; http://taalunieversum.org/).

References

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

Arnold, J., & Griffin, Z. (2007). The effect of additional characters on choice of referring expressions: Everyone competes. *Journal of Memory and Language*, 56, 521-536.

Arts, A. (2004). *Overspecification in instructive texts*. Dissertation, Tilburg University. Wolf Publishers, Nijmegen.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, 43 (1), 361-374.

Bach, K. (1994). *Thought and reference*. Oxford University Press, Oxford.

Bach, K. (2006). The top ten misconceptions about implicature. In: *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor*. Laurence R. Horn, B. Birner & G. Ward (eds). Amsterdam: John Benjamins, 21-30.

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.

Byron, D., Koller, A., Striegnitz, K., Cassell, J., Dale, R., Moore, J., & Oberlander, J. (2009). Report on the first NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proceedings of the 12th European Workshop on Natural Language Generation*, Athens, Greece.

Carston, R. (1991). Implicature, explicature and truth-theoretical semantics. In: Davis, S. (Ed.), *Pragmatics*. Oxford University Press, New York.

Clark, H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1-39.

Cohen, P. (1984). The pragmatics of referring and the modality of communication. *Computational Linguistics*, 10, 97-146.

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the 27th annual meeting of the association for Computational Linguistics*. University of British Columbia, Vancouver, BC, Canada. 68-75

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

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

Davies, C., & Katsos, N. (2009). Are interlocutors as sensitive to over-informativeness as they are to under-informativeness? In *Proceedings of the CogSci workshop on the Production of Referring Expressions (PRE-CogSci 2009)*. Amsterdam, The Netherlands.

Deutsch, W. (1976). *Sprachliche Redundanz und Objekt Identifikation*. University of Marburg, Dissertation; Marburg.

Deutsch, W. & Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition*, 11, 159-184.

Eberhard, K., Spivey-Knowlton, M., Sedivy, J., & Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24, 409-436.

Eikmeyer, H., & Ahlsén, E. (1996). The cognitive process of referring to an object: A comparative study of German and Swedish. In *Proceedings of the 16th Scandinavian Conference on Linguistics*. Turku, Finland.

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.

Engelhardt, P., Demiral, Ş., & Ferreira, F. (2011). Over-specified referential expressions impair comprehension: an ERP study. *Brain and Cognition*, 77, 304-314.

Eriksson, M. (2009). Referring as interaction: On the interplay between linguistic and bodily practices. *Journal of Pragmatics*, 41 (2), 240-262.

Ford, W., & Olson, D. (1975). The elaboration of the noun phrase in children's description of objects. *Journal of Experimental Child Psychology*, 19 (3), 371-382.

Gardent, C. (2002). Generating minimal definite descriptions. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)*, 96-103. Philadelphia, USA.

Gatt, A. (2007). Generating coherent references to multiple entities. Unpublished PhD thesis, University of Aberdeen.

Gatt, A., & Belz, A. (2010). Introducing shared task evaluation to NLG: The TUNA shared task evaluation challenges. In: Krahmer, E. & Theune, M. (Eds.), *Empirical methods in Natural Language Generation*. Berlin and Heidelberg: Springer.

Gatt, A., & Van Deemter, K. (2007). Incremental generation of plural descriptions: Similarity and partitioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2007)*. Prague, Czech Republic.

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

Gatt, A., Van der Sluis, I., Van Deemter, K. (2008b). XML formatting guidelines for the TUNA corpus. Technical report. Department of Computing Science, University of Aberdeen.

Goldberg, E., Driedger, N., Kittredge, R. (1994). Using Natural-Language Processing to produce weather forecasts. *IEEE Expert: Intelligent Systems and their applications*, 9 (2), 45-53.

Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: content planning, modifier ordering, and referential overspecification. *Topics in Cognitive Science* 4 (2), 269-289.

Grice, H. (1975). Logic and conversation. In: Cole, P. & Morgan, J. L. (Eds.), *Speech Acts*. Academic Press, New York, pp. 41-58.

Grice, H. (1989). Studies in the way of words. Harvard University Press, Cambridge, MA.

Gupta, S., & Stent, A. (2005). Automatic evaluation of referring expression generation using corpora. In *Proceedings of the 1st workshop on Using Corpora in Natural Language Generation*, 1-6. Brighton, UK.

Heeman, P., & Hirst, G. (1995). Collaborating on referring Expressions. *Computational Linguistics*, 21(3), 351-282.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In: Schriffrin, D. (Ed.), *Meaning, form and use in context*: Linguistic applications. Georgetown University Press, Washington.

Horn, L. (2005). Current issues in neo-Gricean pragmatics. *Intercultural Pragmatics*, 2, 191-204.

Jordan, P. (2000). Influences on attribute selection in redescriptions: A corpus study. In: *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 250-255, Pennsylvania, USA.

Jordan, P. (2002). Contextual Influences on attribute selection for repeated descriptions. In: K. van Deemter and R. Kibble (Eds.), *Information Sharing: Reference and Presupposition in Natural Language Generation and Understanding*. Stanford, Ca.: CSLI.

Jordan, P., & Walker, M. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24, 157-194.

Kelleher, J., & Kruijff, G.J. (2006). Incremental generation of spatial referring expressions in situated dialogue. In: *Proceedings of the joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*.

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

Levelt, W. (1989). *Speaking: From Intention to Articulation*. MIT Press, Cambridge/London.

Levinson, S. (1979). Activity types and language. *Linguistics*, 17, 365-399.

Levinson, S. (1997). From outer to inner space: Linguistic categories and nonlinguistic thinking. In: Nuyts, J. & Pedersen, E. (Eds.), *Language and Conceptualization*. Cambridge: Cambridge University Press, pp. 13-45.

Levinson, S. (2000). *Presumptive meaning. The theory of generalized conversational implicature*. MIT Press, Cambridge MA.

Maes, A., Arts, A., & Noordman, L. (2004). Reference management in instructive discourse. *Discourse Processes*, 37 (2), 117-144.

Mellish, C., Scott, D., Cahill, L., Evans, R., Paiva, D., & Reape, M. (2006). A reference architecture for Natural Language Generation systems. *Natural Language Engineering*, 12 (1), 1-34.

Mol, L., Krahmer, E., Maes, A., Swerts, M. (2009). The communicative import of gestures: Evidence from a comparative analysis of human-human and human-machine interactions. *Gesture*, 9 (1), 98-127.

Mooney, A. (2004). Co-operation, violations and making sense. *Journal of Pragmatics*, 36 (5), 899-920.

Nadig, A., & Sedivy, J. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, 13, 329-336.

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

Paraboni, I., Masthoff, J., & Van Deemter, K. (2006). Overspecified reference in hierarchical domains: Measuring the benefits for readers. In *Proceedings of the 4th International Conference on Natural Language Generation (INLG-06)*, 55-62. Sydney, Australia.

Paraboni, I., Van Deemter, K., & Masthoff, J. (2007). Making referents easy to identify. *Computational Linguistics*, 33 (2), 229-254.

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

Pickering, M., & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27, 169-226.

Poesio, M., & Vieira, R. (1998). A corpus-based investigation of definite reference use. *Computational Linguistics*, 24, 183-216.

Portet, F., & Gatt, A. (2009). Towards a possibility-theoretic approach to uncertainty in medical data interpretation for text generation. In *Proceedings of the workshop on Knowledge Representation for HealthCare (KR4HC-09)*. Verona, Italy.

Reiter, E. (1990). The computational complexity of avoiding conversational implicatures. In *Proceedings of the 28th annual meeting of the Association for Computational Linguistics (ACL-1990)*, 97-104. Pittsburgh, Pennsylvania, USA.

Reiter, E., & Dale, R. (2000). *Building Natural Language Generation systems*. Cambridge University Press, Cambridge.

Reiter, E., Sripada, S., Hunter, J., Yu, J., & Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167, 137-169.

Schriefers, H., & Pechmann, T. (1988). Incremental production of referential noun phrases by human speakers. In: Zock, M. and Sabah, G. (Eds.), *Advances in Natural Language Generation*, volume 1. Pinter, London, 172-179.

Schwarzschild, R. (1996). *Plurality*. Kluwer, Dordrecht, The Netherlands.

Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Oxford University Press, Oxford.

Sedivy, J., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71, 109-147.

Sedivy, J. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32 (1), 3-23.

Sonnenschein, S. (1984). The effect of redundant communication on listeners: Why different types may have different effects. *Journal of Psycholinguistic Research*, 13, 147-166.

Sonnenschein, S., & Whitehurst, G. (1982). The effects of redundant communications on the behavior of listeners: Does a picture need a thousand words? *Journal of Psycholinguistic Research*, 11 (2), 115-125.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Harvard University Press, Cambridge MA.

Spivey, M., & Richardson, D. (2008). Language embedded in the environment. In: Robbins, P. & Aydede, M. (Eds.), *The Cambridge handbook of situated cognition*. Cambridge University Press, Cambridge, UK, 383-400.

Stock, O., Zancanaro, M., Busetta, P., Callaway, C., Krüger, A., Kruppa, M., Kuflik, T., Not, E., & Rocchi, C. (2007). Adaptive, intelligent presentation of information for the museum visitor in PEACH. *User Modeling and User-Adapted interaction*, 17 (3), 257-304.

Stoia, L., Shockley, D., Byron, D. & Fosler-Lussier, E. (2006). Noun phrase generation for situated dialogs. In *Proceedings of the 4th International Conference on Natural Language Generation*, 81-88. Morristown, NJ, USA.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.

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

Van Deemter, K., Gatt, A., Van Gompel, R., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4, 166-183.

Van der Sluis, I., Gatt, A., & Van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions: Going beyond toy domains. In *Proceedings of the international conference on Recent Advances in Natural Language Processing (RANLP-07)*, Borovets, Bulgaria.

Van der Sluis, I., & Krahmer, E. (2007). Generating multimodal references. *Discourse Processes*, 44, 145-174.

Van der Wege, M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, 60, 448-463.

Viethen, J., & Dale, R. (2006). *Algorithms for generating referring expressions: do they do what people do? In Proceedings of the 4th international conference on Natural Language Generation (INLG-06)*, 63-70. Sydney, Australia.

Westerbeek, H., Koolen, R., & Maes, A. (2013). Color typicality and content planning in definite reference. In *Proceedings of the CogSci workshop on the Production of Referring Expressions: Bridging the gap between empirical and computational approaches to reference (PRE-CogSci 2013)*. Berlin, Germany.

Wilson, D., & Sperber, D. (1981). On Grice's theory of conversation, in: Werth, P. (Ed.), *Conversation and discourse*. Croom Helm, London.

Yip, V., & Matthews, S. (2007). *The bilingual child: Early development and language contact*. Cambridge University Press. Cambridge.

Zender, H., Kruijff, G.J., & Kruijff-Korbayová (2009). Situated resolution and generation of spatial referring expressions for robotic assistants. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, 1604-1609. Pasadena, California, USA.

Need I say more?

3

Learning preferences for Referring Expression Generation: effects of domain, language and algorithm

Abstract

One important subtask of Referring Expression Generation (REG) algorithms is to select the attributes in a definite description for a given object. In this chapter, we study how much training data is required for algorithms to do this properly. We compare two REG algorithms in terms of their performance: the classic Incremental Algorithm and the more recent Graph algorithm. Both rely on a notion of preferred attributes that can be learned from human descriptions. In our experiments, preferences are learned from training sets that vary in size, in two domains and languages. The results show that depending on the algorithm and the complexity of the domain, training on a handful of descriptions can already lead to a performance that is not significantly different from training on a much larger data set.

This chapter is based on:

- Koolen, R., Krahmer, E., & Theune, M. (2012). Learning preferences for Referring Expression Generation: Effects of domain, language and algorithm. In *Proceedings of the 8th International conference on Natural Language Generation (INLG)*. Chicago, USA.
- Krahmer, E., Koolen, R., & Theune, M. (2012). Is it that difficult to find a good preference order for the Incremental Algorithm? *Cognitive Science*, 36 (5), 837-841.
- Theune, M., Koolen, R., Krahmer, E. & Wubben, S. (2011). Does size matter – How much data is required to train a REG algorithm. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics (ACL).* Portland, Oregon, USA.

Introduction

When speakers use a description to refer to an object ("the green sofa") or a person ("the tall man"), they have to determine which properties to include so that an addressee knows which object or person is intended. Clearly, there are many ways in which a sofa can be distinguished from other furniture items ("the large sofa", "the sofa left of the modern-looking chair", etc.), and there are even more possibilities when we consider references to persons. The question, then, is how speakers decide which properties to include in an object description, and - since most practical Natural Language Generation systems include a dedicated module for Referring Expression Generation (REG) in one form or another (Mellish et al., 2006) - how this decision process can be modeled in a REG algorithm.

Various REG algorithms have been proposed that compute which set of properties distinguish a target, where properties themselves are often represented as attribute-value pairs, such as <color, green>, indicating that the target has the value *green* for the attribute 'color'. In a seminal study, Dale and Reiter (1995) present and compare various algorithms that accomplish this task. One algorithm, which they call Greedy Heuristic (GR), relies on discriminatory power: it chooses attributes one by one, at each iteration selecting that attribute of the intended referent that excludes most of the distractors not previously ruled out. An alternative that Dale and Reiter discuss is the Incremental Algorithm (IA), which relies on the assumption that some attributes are more preferred than others. For example, when trying to identify a chair, its color is probably more helpful than its size. This intuition of preferred attributes is formalized using a preference order (PO), which is a list of attributes through which the IA iterates, selecting an attribute-value pair if it helps distinguishing the target from one or more of the distractors.

Even though the IA is exceptional in that it relies on a complete ordering of attributes, also most other current REG algorithms make use of preferences in some way (Di Fabbrizio et al., 2008; Gervás et al., 2008; Kelleher, 2007; Spanger et al., 2008; Viethen and Dale, 2010). The Graph-based algorithm (Krahmer et al., 2003), for example, models preferences in terms of costs, where cheaper is more preferred. Contrary to the IA, the Graph-based algorithm assumes that preferences operate at the level of attribute-value pairs (or properties) rather than at the level of attributes; in this way it becomes possible to prefer a straightforward size (e.g., large) over a subtle color (e.g., mauve or taupe). Moreover, the graph-based algorithm looks for the cheapest overall description, and might opt for a description with a single, relatively dispreferred property ("the man with the blue eyes") when the alternative would be to combine many, relatively preferred properties ("the large, balding man with the bow tie and the striped tuxedo"). This flexibility is arguably one of the reasons why the graph-based REG approach works well: it was the best performing system in the most recent REG Challenge (Gatt et al., 2009).

But where do the preferences used in REG algorithms come from? Dale and Reiter (1995) point out that preferences are domain dependent, and that determining them for a given domain is essentially an empirical question. Unfortunately, they do not specify how this particular empirical question should be answered. The general preference for color over size is experimentally well-established (e.g., Pechmann, 1989), but for most other cases experimental data are not readily available. An alternative would be to look at human data, preferably in a "semantically transparent" corpus (Van Deemter et al., 2006), that is: a corpus that contains the attributes and values of all domain objects, together with the attribute-value pairs actually included in a target reference. Such corpora are typically collected using human participants, who are asked to produce referring expressions for targets in controlled visual scenes. One example is the TUNA corpus, which is a publicly available data set containing 2280 human-produced descriptions in total, and which formed the basis of various REG Challenges. Clearly, building a corpus such as TUNA is a time consuming and labour intensive exercise, so it will not be surprising that only a handful of such corpora exists (and often only for English).

This raises at least one important question: how many human-produced references are needed to make a good estimate of which attributes and properties are preferred? Or, in the words of Van Deemter, Gatt, Van der

Sluis and Power (2012a, p. 804): “how might ‘good’ POs be selected?” In their paper, Van Deemter and colleagues present an extensive evaluation of the various algorithms discussed in Dale and Reiter (1995), and, for the case of the IA, they emphasize the crucial role of the PO in order for the algorithm to perform well. More specifically, they show that, both for descriptions of furniture and people, the IA outperforms the GR algorithm with a ‘good’ PO, while the opposite holds when the IA relies on a ‘bad’ PO. Van Deemter et al. point out that systematically testing all POs for a given domain – searching for the best performing one - quickly becomes impractical, since for a domain where objects can be described using *n* different attributes, there are *n*! POs to consider. Based on considerations such as these, they conclude that “someone who is looking for a GRE algorithm for a previously unstudied application domain might do better choosing GR (...), instead of an unproven version of the IA” (Van Deemter et al., 2012a, p. 829). Arguably, the above problems also exist for the graph-based algorithm (Krahmer et al., 2003), since this algorithm makes use of attribute preferences for the generation of its descriptions as well (as we have explained earlier).

In any case, it seems to us that it should first be determined how difficult it really is to find a ‘good’ PO. Do we really need hundreds of instances, or is it conceivable that a few of them (collected in a semantically transparent way) will do? This is not an easy matter, since various factors might play a role: from which data set are example references sampled, what are the domains of interest, and, perhaps most importantly, which REG algorithm is considered? In this chapter, we address these questions by running learning curve experiments, where we systematically train two leading REG algorithms (the Incremental Algorithm and the graph-based REG algorithm) on sets of human-produced descriptions of increasing size. Subsequently, we evaluate them on a held-out test set. We do this for two different domains (people and furniture descriptions) and two data sets in two different languages (TUNA and D-TUNA, the Dutch version of TUNA). In this chapter, we use the exact same evaluation metrics as Van Deemter et al., thereby extending the findings of their experiments.

Below we first explain in more detail which algorithms (Section 2) and corpora (Section 3) we used for our experiments. Then we describe how we derived costs and preference orders from subsets of these corpora (Section 4), and report the results of our experiments focusing on effects of domain, language and size of the training set (Section 5). We end with a discussion and conclusion (Section 6), where we also compare the performance of the IA trained on small set sizes with that of the classical Full Brevity and Greedy algorithms (Dale & Reiter, 1995).

The Algorithms

In this section we briefly describe the two algorithms, and their settings, used in our experiments. For details about the algorithms we refer to the original publications.

The Incremental Algorithm

The basic assumption underlying the Incremental Algorithm (Dale and Reiter, 1995) is that speakers "prefer" certain attributes over others when referring to objects. This intuition is formalized in the notion of a list of attributes, ranked in order of preference. When generating a description for a target, the algorithm iterates through this list, adding an attribute to the description under construction if its value helps rule out any of the distractors not previously ruled out. There is no backtracking in the IA, which means that a selected attribute is always realized in the final description, even if the inclusion of later attributes renders it redundant. In this way, the IA is capable of generating overspecified descriptions, in accordance with the human tendency to mention redundant information (Pechmann, 1989; Engelhardt et al., 2006; Arts et al., 2011). The type attribute (typically realized as the head noun) has a special status in the IA. After running the algorithm it is checked whether type is in the description; if not, it is added, so that type is always included even if it does not rule out any distractors.

To derive preference orders from human-produced descriptions we proceeded as follows: given a set of *n* descriptions sampled from a larger corpus (where *n* is the set size, a variable we systematically control in our experiments), we counted the numbers of times a certain attribute occurred in the *n* descriptions. The most frequently occurring attribute was placed at the first position of the preferred attributes list, followed by the second most frequent attribute, etc. In the case of a tie (i.e., when two attributes occurred equally often, which typically is more likely to happen in small training sets), the attributes were ordered alphabetically. In this way, we made sure that all ties were treated in the same, comparable manner, which resulted in a complete ranking of attributes, as required by the IA.

The graph-based algorithm

In the graph-based algorithm (Krahmer et al., 2003), which we refer to as Graph, information about domain objects is represented as a labelled directed graph, and REG is modeled as a graph-search problem. The output of the algorithm is the cheapest distinguishing subgraph, given a particular cost function assigning costs to properties (i.e., attribute-value pairs). By assigning zero costs to some properties Graph is also capable of generating overspecified descriptions, including redundant properties. To ensure that the graph search does not terminate before the free properties are added, the search order must be explicitly controlled (Viethen et al., 2008). To ensure a fair comparison with the IA, we make sure that if the target's type property was not originally selected by the algorithm, it is added afterwards.

In this study, both the costs and orders required by Graph are derived from corpus data. We base the property order on the frequency with which each attribute-value pair is mentioned in a training corpus, relative to the number of target objects with this property. The properties are then listed in order of decreasing frequency. To derive costs, we used the same corpus frequencies: we created different cost functions (mapping properties to costs) by means of *k*-means clustering, using the Weka toolkit. The *k*-means clustering algorithm assigns *n* points in a vector space to *k* clusters (S_1 to S_k) by assigning each point to the cluster with the nearest centroid. The total intra-cluster variance *V* is minimized by the following function:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

where μ_i is the centroid of all the points $x_j \in S_i$. In our case, the points *n* are properties, the vector space is one-dimensional (frequency being the only dimension) and μ_i is the average frequency of the properties in S_i. The cluster-based costs are defined as follows:

$$\forall x_j \in S_i, cost(x_j) = i - 1$$

where S_1 is the cluster with the most frequent properties, S_2 is the cluster with the next most frequent properties, and so on.

Given that Theune et al. (2011) achieved the best performance with $k = 2$ (meaning that the properties are divided in two groups based on their frequency), we use this two cluster option in the current study as well[1]. The properties in the group with the highest frequency get cost 0. These 'free' properties are always included in the description if they help distinguish the target. The properties in the less frequent group get cost 1; of these properties, the algorithm only adds the minimum number necessary to achieve a distinguishing description. Ties due to properties occurring with the same frequency need not be resolved when determining the cost function, since Graph does not assume the existence of a complete ordering (Krahmer et al., 2003). Properties that did not occur in a training corpus

[1] For comparison, we also evaluated the Free Naïve and Stochastic cost functions on our test data. We found that the Free-Naïve cost function performed better than the stochastic cost function (Theune et al., 2010), but not significantly different from *k*-means clustering with $k = 2$ (Theune et al., 2011). These findings show that *k*-means clustering can be regarded as a systematic alternative for the manual assignment of frequency-based costs.

were automatically assigned cost 1. Like we did for the IA, we listed attribute-value pairs with the same frequency in alphabetical order.

The corpora

Training and test data for our experiments were taken from two corpora of referring expressions, one English (TUNA) and one Dutch (D-TUNA).

English data: the TUNA corpus

The TUNA corpus (Gatt et al., 2007) is a semantically transparent corpus consisting of object descriptions in two domains (furniture and people). The corpus was collected in an on-line production experiment, in which participants were presented with visual scenes containing one target object and six distractor objects. These objects were ordered in a 5 × 3 grid, and the participants were asked to describe the target in such a way that it could be uniquely distinguished from its distractors. Table 1 shows the attributes and values that were annotated for the descriptions in the two domains.

Table 1: Attributes and values in the furniture and people domains. X- and Y-DIMENSION refer to an object's horizontal and vertical position in a scene grid and only occur in the English TUNA corpus.

Furniture	
Attribute	**Possible values**
Type	Chair, desk, sofa, fan
Color	Green, red, blue, gray
Orientation	Front, back, left, right
Size	Large, small
X-dimension	1, 2, 3, 4, 5
Y-dimension	1, 2, 3

People	
Attribute	**Possible values**
Type	Person
Age	Old, young
Hair color	Light, dark
Orientation	Front, left, right
Has Beard	0 (false), 1 (true)
Has Glasses	0, 1
Has Shirt	0, 1
Has Suit	0, 1
Has Tie	0, 1
X-dimension	1, 2, 3, 4, 5
Y-dimension	1, 2, 3

There were two experimental conditions: in the +LOC condition, the participants were free to describe the target using any of its properties, including its location on the screen (represented in Table 1 as the X- and Y-DIMENSION), whereas in the -LOC condition they were discouraged (but not prevented) from mentioning object locations. However, some descriptions in the -LOC condition contained location information anyway.

Dutch data: the D-TUNA corpus

For Dutch, we used the D-TUNA corpus (Koolen and Krahmer, 2010; see also Chapter 2 of this dissertation). This corpus uses the same visual scenes and annotation scheme as the TUNA corpus, but consists of Dutch instead of English target descriptions. Since the D-TUNA experiment was performed in laboratory conditions, its data is relatively 'cleaner' than the TUNA data, which means that it contains fewer descriptions that are not fully distinguishing and that its descriptions do not contain X- and Y-DIMENSION attributes. Although the descriptions in D-TUNA were collected in three different experimental conditions (written, spoken, and face-to-face), we only use the written descriptions in this chapter, as this condition is most similar to the data collection in TUNA.

Method

To find out how much training data is required to achieve an acceptable attribute selection performance for the IA and Graph, we derived orders and costs from different sized training sets. We then evaluated the algorithms on a held-out test set, using the derived orders and costs. Training and test sets were taken from TUNA and D-TUNA.

As Dutch training data, we used 160 furniture and 160 people items, randomly selected from the textual descriptions in the D-TUNA corpus. The remaining furniture and people descriptions (40 items each) were used for testing. As English training data, we took all -LOC data from the training set of the REG Challenge 2009 (Gatt et al., 2009): 165 furniture and 136 people descriptions. As English test data we used all -LOC data from the REG 2009 development set: 38 furniture and 38 people descriptions. We only used -LOC data to increase comparability to the Dutch data.

From the Dutch and English furniture and people training data, we selected random subsets of 1, 5, 10, 20, 30, 40 and 50 descriptions. Five different sets of each size were created, since the accidental composition of a training set could strongly influence the results. All training sets were built up in a cumulative fashion, starting with five randomly selected sets of size 1, then adding 4 items to each of them to create five sets of size 5, and so on, for each combination of language and domain. We used these different training sets to derive preference orders of attributes for the IA, and costs and property orders for Graph, as outlined above.

We evaluated the performance of the derived preference orders and cost functions on the test data for the corresponding domain and language, using the standard Dice and Accuracy metrics for evaluation. Dice measures the overlap between attribute sets, producing a value between 1 and 0, where 1 stands for a perfect match and 0 for no overlap at all. Dice scores are calculated by scaling the number of attributes that two descriptions have in common, by the overall size of the two sets:

$$\text{dice}\ (D_{\mathrm{H}}, D_{\mathrm{A}}) = \frac{2 \times |D_{\mathrm{H}} \cap D_{\mathrm{A}}|}{|D_{\mathrm{H}}| + |D_{\mathrm{A}}|}$$

where D_{H} is the (set of attributes in) the description produced by a human author and D_{A} the description generated by an algorithm. Accuracy is the percentage of perfect matches between the generated attribute sets and the human descriptions in the test set. Both metrics were used in the REG Generation Challenges (Gatt & Belz, 2010).

Results

Effects of domain, language and algorithm

To determine the effect of domain and language on the performance of REG algorithms, we applied repeated measures analyses of variance

(ANOVA) to the Dice and Accuracy scores, using *domain* (levels: furniture, people) as a within variable, and *algorithm* (levels: IA, Graph) and *language* (levels: English, Dutch) as between variables.

The results show main effects of *domain* (Dice: $F_{(1,152)}$ = 56.10, p < .001; Acc.: $F_{(1,152)}$ = 76.36, p < .001) and *language* (Dice: $F_{(1,152)}$ = 30.30, p < .001; Acc.: $F_{(1,152)}$ = 3.380, p = .07). Regarding the two domains, these results indicate that both the IA and the Graph algorithm generally performed better in the furniture domain (Dice: M = .86, SD = .01; Acc.: M = .56, SD = .03) than in the people domain (Dice: M = .72, SD = .01;Acc.: M = .20,SD = .02). Regarding the two languages, the results show that both algorithms generally performed better on the Dutch data (Dice: M = .84, SD = .01; Acc.: M = .41, SD = .03) than on the English data (Dice: M = .74, SD = .01; Acc.: M = .34, SD = .03). There is no main effect of *algorithm*, meaning that overall, the two algorithms had an equal performance. However, this is different when we look separately at each domain and language, as we do below.

Learning curves per domain and language

Given the main effects of domain and language described above, we ran separate ANOVAs for the different domains and languages. For these four analyses, we used *set size* (levels: 1, 5, 10, 20, 30, 40, 50, all) as a within variable, and *algorithm* as a between variable. To determine the effects of set size, we calculated the means of the scores of the five training sets for each set size, so that we could compare them with the scores of the entire set. The results are shown in Tables 2 and 3 on the next page.

Table 2: Performance for each set size in the furniture domain. For sizes 1 to 50, means over five sets are given. The full sets are 165 English and 160 Dutch descriptions.

	English Furniture			
	IA		Graph	
Set size	Dice	Acc. (%)	Dice	Acc. (%)
1	0.764	36.8	0.693	24.7
5	0.829	55.3	0.756	33.7
10	0.829	55.3	0.777	39.5
20	0.829	55.3	0.788	40.5
30	0.829	55.3	0.782	40.5
40	0.829	55.3	0.793	45.3
50	0.829	55.3	0.797	45.8
All	0.829	55.3	0.810	50.0

	Dutch Furniture			
	IA		Graph	
Set size	Dice	Acc. (%)	Dice	Acc. (%)
1	0.925	63.0	0.876	44.5
5	0.935	67.5	0.917	62.0
10	0.929	68.5	0.923	66.0
20	0.930	65.5	0.923	64.0
30	0.931	67.0	0.924	65.5
40	0.931	67.0	0.931	67.5
50	0.929	66.0	0.929	67.0
All	0.926	65.0	0.929	67.5

Table 3: Performance for each set size in the people domain. For sizes 1 to 50, means over five sets are given. The full sets are 136 English and 160 Dutch descriptions.

	English People			
	IA		Graph	
Set size	Dice	Acc. (%)	Dice	Acc. (%)
1	0.519	7.4	0.558	12.6
5	0.605	15.8	0.617	14.5
10	0.682	21.1	0.683	20.0
20	0.710	22.1	0.716	24.7
30	0.682	15.3	0.716	26.8
40	0.716	26.3	0.723	26.3
50	0.718	27.9	0.727	26.3
All	0.724	31.6	0.730	28.9

	Dutch People			
	IA		Graph	
Set size	Dice	Acc. (%)	Dice	Acc. (%)
1	0.626	4.5	0.682	17.5
5	0.737	16.0	0.738	21.0
10	0.738	12.5	0.741	19.5
20	0.765	12.5	0.778	25.5
30	0.762	14.5	0.789	25.0
40	0.763	11.5	0.792	25.0
50	0.764	10.5	0.798	26.0
All	0.812	12.5	0.812	32.5

We made planned post hoc comparisons to test which is the smallest set size that does not perform significantly different from the entire training set in terms of Dice or Accuracy scores (we call this the "ceiling"). We report results both for the standard Bonferroni method, which corrects for multiple comparisons, and for the less strict HSD method from Fisher, which does not. Note that with the Bonferroni method we are inherently less likely to find statistically significant differences between the set sizes, which implies that we can expect to reach a ceiling earlier than with the HSD method. Table 4 shows the ceilings we found for the algorithms, per domain and language.

Table 4: Ceiling set sizes computed using HSD, with Bonferroni between brackets.

	English Furniture		**Dutch Furniture**	
	Dice	Accuracy	Dice	Accuracy
IA	5 (5)	5 (5)	1 (1)	1 (1)
Graph	10 (5)	40 (5)	5 (1)	5 (1)
	English People		**Dutch People**	
	Dice	Accuracy	Dice	Accuracy
IA	10 (10)	40 (1)	20 (5)	1 (1)
Graph	20 (10)	20 (1)	30 (20)	5 (1)

The furniture domain. Table 2 shows the Dice and Accuracy scores in the furniture domain. We found significant effects of *set size* for both the English data (Dice: $F_{(7,518)} = 15.59$, $p < .001$; Acc.: $F_{(7,518)} = 17.42$, $p < .001$) and the Dutch data (Dice: $F_{(7,546)} = 5.322$, $p < .001$; Acc.: $F_{(7,546)} = 5.872$, $p < .001$), indicating that for both languages, the number of descriptions used for training influenced the performance of both algorithms in terms of both Dice and Accuracy. Although we did not find a main effect of algorithm, suggesting that the two algorithms performed equally well, we did find several interactions between *set size* and *algorithm* for both the English data (Dice: $F_{(7,518)} = 1.604$, ns; Acc.: $F_{(7,518)} = 2.282$, $p < .05$) and the Dutch data

(Dice: $F_{(7,546)} = 3.970$, $p < .001$; Acc.: $F_{(7,546)} = 3.225$, $p < .01$). For the English furniture data, this interaction implies that small set sizes have a bigger impact for the IA than for Graph. For example, moving from set size 1 to 5 yielded a Dice improvement of .18 for the IA, while this was only .09 for Graph. For the Dutch furniture data, however, a reverse pattern was observed; moving from set size 1 to 5 yielded an improvement of .01 (Dice) and .05 (Acc.) for the IA, while this was .11 (Dice) and .18 (Acc.) for Graph.

As depicted in table 4, post hoc tests showed that small set sizes were generally sufficient to reach ceiling performance: the general pattern for both algorithms and both languages was that the scores increased with the size of the training set, but that the increase got smaller as the set sizes became larger. For the English furniture data, Graph reached the ceiling at set size 10 for Dice (5 with the Bonferroni test), and at set size 40 for Accuracy (again 5 with Bonferroni), while this was the case for the IA at set size 5 for both Dice and Accuracy (also 5 with Bonferroni). For the Dutch furniture data, Graph reached the ceiling at set size 5 for both Dice and Accuracy (and even at 1 with the Bonferroni test), while this was at set size 1 for the IA (again 1 with Bonferroni).

The people domain. Table 3 shows the Dice and Accuracy scores in the people domain. Again, we found significant effects of *set size* for both the English data (Dice: $F_{(7,518)} = 39.46$, $p < .001$; Acc.: $F_{(7,518)} = 11.77$, $p < .001$) and the Dutch data (Dice: $F_{(7,546)} = 33.90$, $p < .001$; Acc.: $F_{(7,546)} = 3.235$, $p < .01$). Again, this implies that for both languages, the number of descriptions used for training influenced the performance of both algorithms in terms of both Dice and Accuracy. Unlike we did in the furniture domain, we found no interactions between *set size* and *algorithm*, but we did find a main effect of algorithm for the Dutch people data (Dice: $F_{(1,78)} = .751$, ns; Acc.: $F_{(1,78)} = 5.099$, $p < .05$), showing that Graph generated Dutch descriptions that were more accurate than those generated by the IA.

As in the furniture domain, post hoc tests showed that small set sizes were generally sufficient to reach ceiling performance (see table 4). For the English data, Graph reached the ceiling at set size 20 for both Dice and

Accuracy (with Bonferroni: 10 for Dice, 1 for Accuracy), while this was the case for the IA at set size 10 for Dice (also 10 with Bonferroni), and at set size 40 for Accuracy (and even at 1 with Bonferroni). For the Dutch data, Graph reached the ceiling at set size 30 for Dice (20 with Bonferroni), and at set size 5 for Accuracy (1 with Bonferroni). For the IA, ceiling was reached at set size 20 for Dice (Bonferroni: 5), and already at 1 for Accuracy (Bonferroni: 1).

Discussion

Our main goal was to investigate how many human-produced references are required by two REG algorithms (the Incremental Algorithm and the Graph-based algorithm) to determine preferences and costs for a new domain, and to generate "human-like" descriptions for new objects in these domains. Our results show that with relatively small sets of descriptions (much smaller than the entire TUNA or D-TUNA corpus) we obtained results that are not significantly different from the best performing variants of the two algorithms. In the simple furniture domain even one training item can already be sufficient, at least for the IA. In other words, even based on a small set of human-produced descriptions we can make an informed guess about what a 'good' preference order or cost is for a given domain. Arguably, these results allow two important, related conclusions: (a) Even though, in theory, there are indeed *n*! different possible preference rankings of *n* attributes (or attribute-value pairs), it is both more important and less difficult to get a 'good' ordering for the head of the ranking than for its tail. And (b) relatively many human-produced descriptions will include properties that are preferred, so that it does not really matter how dispreferred properties are ordered.

As shown in Table 4, on the whole the IA needed fewer training data than Graph (except in the English people domain, where Graph only needed a set size of 10 to hit the ceiling for Dice, while the IA needed a set size of 20). Given that the IA ranks attributes, while the graph-based REG algorithm ranks attribute-value pairs, the difference in required training data is not surprising. In any domain, there will probably be more attribute-value pairs than attributes, so determining an attribute ranking is an easier task than determining a ranking of attribute-value pairs. Another advantage of ranking attributes rather than attribute-value pairs is that it is less vulnerable to the problem of "missing data". More specifically, the chance that a specific attribute does not occur in a small training set is much smaller than the chance that a specific attribute-value pair does not occur. As a consequence, the IA needs fewer data to obtain complete attribute orderings than Graph needs to obtain costs for all attribute-value pairs.

Interestingly, we only found interactions between training set size and algorithm in the furniture domain. In the people domain, there was no significant difference between the size of the training sets required by the algorithms. This could be explained by the fact that the people domain has about twice as many attributes as the furniture domain, and fewer values per attribute (see Table 1). This means that for people the difference between the number of attributes (IA) and the number of attribute-value pairs (Graph) is not as big as for furniture, so the two algorithms are on more equal grounds.

Both algorithms performed better on furniture than on people. Arguably, the people pictures in the TUNA experiment can be described in many more different ways than the furniture pictures can, so it stands to reason that ranking potential attributes and values is more difficult in the people than in the furniture domain. In a similar vein, we might expect Graph's flexible generation strategy to be more useful in the people domain, where more can be gained by the use of costs, than in the furniture domain, where there are relatively few options anyway, and a simple linear ordering may be quite sufficient.

This expectation was at least partially confirmed by the results: although in most cases the differences are not significant, Graph tends to perform numerically better than the IA in the people domain. Here we may see the pay-off of Graph's more fine-grained preference ranking, which allows it to distinguish between more and less salient attribute values. In the furniture domain, most attribute values appear to be more or less equally salient (e.g., none of the colors gets notably mentioned more often), but in the people domain certain values are clearly more salient than others. In particular, the attributes HasBeard and HasGlasses are among the most frequent attributes in the people domain when their value is true (i.e., the target object can be distinguished by his beard or glasses), but they hardly get mentioned when their value is false. Graph quickly learns this distinction, assigning low costs and a high ranking to <*HasBeard, true*> and <*HasGlasses, true*> while assigning high costs and a low ranking to <*HasBeard, false*> and <*HasGlasses, false*>. The IA, on the other hand, does not distinguish between the values of these attributes.

Moreover, the graph-based algorithm is arguably more generic than the Incremental Algorithm, as it can straightforwardly deal with relational properties and lends itself to various extensions (Krahmer et al., 2003). In short, the larger training investment required for Graph in simple domains may be compensated by its versatility and better performance on more complex domains. To test this assumption, our experiments should be repeated using data from a more realistic and complex domain, e.g., geographic descriptions (Turner et al., 2008). Unfortunately, currently no such data sets are available.

Finally, we found that the results of both algorithms were better for the Dutch data than for the English ones. We think that this is not so much an effect of the language (as English and Dutch are highly comparable) but rather of the way the TUNA and D- TUNA corpora were constructed. The D-TUNA corpus was collected in more controlled conditions than TUNA and as a result, arguably, it contains training data of a higher quality. Also, because the D-TUNA corpus does not contain any location properties (X- and Y-dimension) its furniture and people domains are slightly less complex than their TUNA counter- parts, making the attribute selection task a bit easier.

One caveat of our study is that so far we have only used the standard automatic metrics on REG evaluation (albeit in accordance with many other studies in this area). However, it has been found that these do not always correspond to the results of human-based evaluations, so it would be interesting to see whether the same learning curve effects are obtained for extrinsic, task based evaluations involving human subjects. Following Belz and Gatt (2010), this can be done by measuring reading times, identification times or error rates as a function of training set size.

Comparing IA with FB and GR. We have shown that small set sizes are sufficient to reach ceiling for the IA. But which preference orders (POs) do we find with these small set sizes? And how does the IA's performance with these orders compare to the results obtained by alternative algorithms such as Dale and Reiter's (1995) classic Full Brevity (FB) and Greedy Algorithm (GR)? – a question explicitly asked by Van Deemter et al. (2012a) and by Van Deemter, Gatt, Van der Sluis and Power (2012b). In the furniture domain, all five English training sets of size 5 yield a PO for which Van Deemter et al. (2012a) showed that it causes the IA to significantly outperform FB and GR (i.e., either C(olor)O(rientation)S(ize) or CSO; note that here we abstract over type which Van Deemter and colleagues do not consider). When we look at the English people domain and consider set size 10 (ceiling for Dice), we find that four out of five sets have a preference order where HairColor, HasBeard and HasGlasses are in the top three (again disregarding type); one of these (hasGlasses, HasBeard, HairColor) is the best performing preference order found by Van Deemter and colleagues (2012a), another performs slightly worse but still significantly better than FB and GR (HasBeard, HasGlasses, HairColor); the other two score statistically comparable to the classical algorithms. The fifth people PO includes X- and Y-dimension in the top three, which Van Deemter et al. ignore. In sum: in almost all cases, small set sizes (5 and 10 respectively) yield POs with which the IA performs at least as well as the FB and GR algorithms, and in most cases significantly better.

Conclusion

We have shown that with few training instances, acceptable attribute selection results can be achieved; that is, results that do not significantly differ from those obtained using a much larger training set. Given the scarcity of resources in this field, we feel that this is an important result for researchers working on REG and Natural Language Generation in general.

We found that less training data is needed in simple domains with few attributes, such as the furniture domain, and more in relatively more complex domains such as the people domain. The data set being used is also of influence: better results were achieved with D-TUNA than with the TUNA corpus, which probably not so much reflects a language difference, but a difference in the way the corpora were collected.

We found some interesting differences between the IA and Graph algorithms, which can be largely explained by the fact that the former ranks attributes, and the latter attribute-value pairs. The advantage of the former (coarser) approach is that overall, fewer training items are required, while the latter (more fine-grained) approach is better equipped to deal with more complex domains. In the furniture domain both algorithms had a similar performance, while in the people domain Graph did slightly better than the IA.

It is worth pointing out that our approach still requires a semantically balanced corpus. However, the evaluation results suggest that a surprisingly small one may be sufficient. Of course, it should be kept in mind that these conclusions, while intuitive, are based on data from the relatively simple domains studied by Van Deemter et al. (2012), and evaluated by using a limited (though standard) set of evaluation metrics. Moreover, here we have only considered references to single targets, while Van Deemter et al. (2012) also consider references to sets, for which determining a preference ranking might well be more complicated.

It will certainly be interesting to see what happens when the field of Referring Expression Generation moves toward more complex references and more open-ended domains. For now, we conclude that someone who is looking for a REG algorithm for a new application domain might consider collecting a handful of human-produced descriptions, before discarding the IA or the graph-based algorithm in favor of another algorithm.

Acknowledgments

We thank Albert Gatt for allowing us to use his implementation of the IA, and Sander Wubben for help with *k*-means clustering.

Need I say more?

References

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, 43 (1), 361–374.

Belz, A., & Gatt, A. (2008). Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, 197–200.

Dale, R., & Reiter, E. (1995). Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2), 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.

Di Fabbrizio, G., Stent, A., & Bangalore, S. (2008). Trainable speaker-based referring expression generation. In *Twelfth Conference on Computational Natural Language Learning (CoNLL- 2008)*, 151–158.

Gatt, A., Van der Sluis, I., & Van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG 2007)*, 49–56.

Gatt, A., Belz, A., & Kow, E. (2009). The TUNA-REG Challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, 174–182.

Gatt, A., & Belz, A. (2010). Introducing shared task evaluation to NLG: the TUNA shared task evaluation challenges. In E. Krahmer & M. Theune (Eds.), Empirical methods in Natural Language Generation. Berlin and Heidelberg: Springer (LNCS 5790).

Gervàs, P., Hervàs, R., & Léon, C. (2008). NIL-UCM: Most-frequent-value-first attribute selection and best-scoring-choice realization. In *Proceedings of the 5th International Natural Language Generation Conference (INLG 2008)*, 215–218.

Jordan, P., & Walker, M. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24, 157–194.

Kelleher. J. (2007). DIT - frequency based incremental attribute selection for GRE. In *Proceedings of the MT Summit XI Workshop Using Corpora for Natural Language Generation: Language Generation and Machine Translation (UCNLG+MT)*, 90–92.

Koolen, R., & Krahmer, E. (2010). The D-TUNA corpus: A Dutch dataset for the evaluation of referring expression generation algorithms. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010)*.

Krahmer, E., Van Erk, S., & Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, 29(1), 53–72.

Mellish, C., Scott, D., Cahill, L., Paiva, D., Evans, R., & Reape, M. (2006). A reference architecture for natural language generation systems. *Natural Language Engineering*, 12, 1–34.

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

Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating NLG systems. *Computational Linguistics*, 35(4), 529–558.

Spanger, P., Kurosawa, T., & Tokunaga, T. (2008). On "redundancy" in selecting attributes for generating referring expressions. In *COLING 2008: Companion volume: Posters*, 115– 118.

Theune, M., Koolen, R., Krahmer, E., & Wubben, S. (2011). Does size matter – How much data is required to train a REG algorithm? In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, 660-664.

Turner, R., Sripada, S., Reiter, E., & Davy, I. (2008). Using spatial reference frames to generate grounded textual summaries of georeferenced data. In *Proceedings of the 5th International Natural Language Generation Conference (INLG)*, 16–24.

Van Deemter, K., Van der Sluis, I., & Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Natural Language Generation Conference (INLG 2006)*, 130–132.

Van Deemter, K., Gatt, A., Van der Sluis, I., & Power, R. (2012a). Generation of referring expressions: Assessing the Incremental Algorithm. *Cognitive Science*, 36 (5), 799-836.

Van Deemter, K., Gatt. A., Van der Sluis, I., & Power, R. (2012b). Assessing the Incremental Algorithm: a response to Krahmer et al. *Cognitive Science*, 36 (5), 842-845.

Viethen, J., & Dale, R. (2010). Speaker-dependent variation in content selection for referring expression generation. In *Proceedings of the 8th Australasian Language Technology Workshop*, 81–89.

Viethen, J., Dale, R., Krahmer, E., Theune, M., & Touset, P. (2008). Controlling redundancy in referring expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 239–246.

Need I say more?

4

The effect of scene variation on the redundant use of color

Abstract

This chapter investigates to what extent the amount of variation in a visual scene causes speakers to mention the attribute color in their definite target descriptions, focusing on scenes in which this attribute is not needed for identification of the target. The results of our three experiments show that speakers are more likely to redundantly include a color attribute when the scene variation is high as compared to when this variation is low (even if this leads to overspecified descriptions). We argue that these findings are problematic for existing algorithms that generally aim to automatically generate psychologically realistic target descriptions, such as the Incremental Algorithm, since these algorithms make use of a fixed preference order per domain and do not take visual scene variation into account.

This chapter is based on:

- Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, 37 (2), 395-411.

Introduction

In everyday language use, speakers often produce definite descriptions of *target* objects (such as “the brown chair”), and they aim to do this in such a way that their addressee is able to uniquely distinguish the target from its surrounding *distractor* objects (e.g., other furniture items). It is well-known that speakers tend to *overspecify* their descriptions by providing *redundant* attributes that are not needed for target identification (e.g., Arts, 2004; 2011; Engelhardt, Bailey, & Ferreira, 2006), which is, as some have argued, in conflict with the Maxim of Quantity as proposed by Grice (1975). In any case, overspecification is hard to capture for Referring Expression Generation algorithms, which are computational models that aim to generate definite object descriptions (Reiter & Dale, 2000). These algorithms generally focus on *Content Determination*: what attributes should be included to distinguish the target? Most algorithms (including Dale and Reiter’s Incremental Algorithm, introduced in 1995) rest on the assumption that some attributes are preferred over others, and that these preferences are fixed for a given domain. As we explain later on, this assumption gives rise to some amount of overspecification. However, given that there is growing awareness that speech production and visual scene perception are closely intertwined (e.g., Griffin & Bock, 2000; Hanna & Brennan, 2007; Meyer, Sleiderink, & Levelt, 1998), we argue that whether speakers redundantly mention a certain attribute (in particular: color) does not merely depend on how preferred this attribute is in a given domain, but also on the visual variation in a scene.

Psycholinguistic evidence for overspecification

It is often assumed that speakers tend to obey the Maxim of Quantity (Grice, 1975), stating that speakers should make their contribution as informative as is required (for the current purpose of the exchange), but not more informative than that. If one regards identification as a core purpose of uttering a target description (as we do here, following many previous papers on referential overspecification, for example: Engelhardt et al., 2006; Olson, 1970; Pechmann, 1989), the Maxim of Quantity would result in descriptions that contain just enough attributes for the addressee to identify the target. This prediction is at odds with the observation that speakers often overspecify and provide their addressees with redundant attributes (e.g., Arts, 2004; 2011; Engelhardt et al., 2006). One plausible reason for this is that speakers tend to include attributes that they prefer, even if mentioning these attributes leads to overspecified target descriptions. Attributes that are known to be generally preferred tend to be perceptually salient, for example color (e.g., Belke & Meyer, 2002; Pechmann, 1989). The focus in this chapter lies therefore on the redundant use of the attribute color in definite object descriptions.

Overspecification in Referring Expression Generation

These findings concerning the occurrence of referential overspecification and speakers' preferences for certain attributes have important implications for researchers in the field of Natural Language Generation (NLG). NLG is a subfield of Artificial Intelligence that aims to build models for automatically generating natural language text or speech, usually from non-linguistic information (e.g., from a database; Reiter & Dale, 2000). NLG systems typically use Referring Expression Generation (REG) algorithms that generate distinguishing descriptions of objects (Mellish, Cahill, Evans, Paiva, & Reape, 2006). Many of these REG algorithms can be seen as (implicit or explicit) computational interpretations of the Gricean maxims of conversational implicature. For example, the algorithms discussed by Dale and Reiter (1995) explicitly take the Maxim of Quantity as a starting point, aiming to approximate the referential behavior of human speakers.

So how should the Maxim of Quantity be interpreted in the context of a referring expression generation task? It is worth mentioning in this respect that most current REG algorithms generate referring expressions that are solely intended to identify a target object and have no other (nonidentificational) communicative purposes (such as giving a warning). With that in mind, Dale and Reiter (1995, p. 240) propose the following interpretation of the Maxim of Quantity: "a referring expression should contain enough information to enable the hearer to identify the object referred to, but not more information". Many of the various REG algorithms that have been proposed so far rely on this interpretation. For example, the *Full Brevity* Algorithm (Dale, 1989; 1992) is based on a strict interpretation of the Maxim of Quantity and always seeks to find the shortest possible target description (in terms of the number of attributes included). This is not necessarily the case with the *Greedy Heuristic* algorithm (Dale, 1989; 1992), which iteratively selects the attribute that rules out most of the distractor objects at each stage of the attribute selection process. However, the most influential REG algorithm to date (as discussed by Van Deemter, Gatt, Van de Sluis, & Power, 2012a) is arguably the *Incremental Algorithm* (IA). In this chapter, we therefore base our predictions mainly on this algorithm.

The Incremental Algorithm (Dale & Reiter, 1995) generates target descriptions by using a predetermined *preference order*. This is a ranking of all attributes that can possibly occur in a given domain, where preferred attributes are ranked before less preferred attributes. Dale and Reiter (1995) argue that this preference order is fixed for every domain, and that it can typically be determined empirically. To illustrate how a preference order is usually determined, let us consider the *furniture domain*, which has often been used before in REG studies (e.g., Gatt & Belz, 2010; Gatt, Van der Sluis, & Van Deemter, 2007; Van Deemter et al., 2012a), and which we use in this chapter as well. Empirical data presented in Chapter 2 of this dissertation show that the target's type (e.g. "chair") is practically always mentioned. Therefore, *type* can be placed at the head of the preference order. The second most frequent attribute in this domain is color, while infrequent attributes such as size and orientation occur in the tail of the preference order.

How does the IA use this preference order? Consider the two chairs depicted in Fig. 1 (on the next page), and imagine that the algorithm wants to distinguish the brown chair from the green chair.

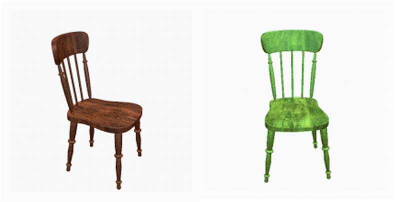


Fig. 1: A brown chair and a green chair.

In this case, the IA would first consider the *type* attribute, because it is at the head of the preference order. Since *type* does not rule out the only distractor (because both objects are chairs), the IA considers the next attribute in line (*color*) and checks whether the attribute-value pair <*color*, *brown*> rules out the distractor, which it does. The resulting description could be realized as "the brown one". However, since most descriptions tend to contain a head noun mentioning a target's type, the IA always adds the type to a description (note that it only does this when type was not selected at an earlier stage, like in our example).

The IA does not backtrack for overspecification, which means that it does not remove selected attributes that turn out to be redundant in the end. This is the case when there exists one other - less preferred - attribute (or a combination of several other attributes) that renders all higher ranked attributes obsolete. For example, one can think of a visual scene with one target object and two distractors, where color rules out one distractor and size two. Color and size are then both selected, although including size would have been sufficient. In this way, the IA is able to generate overspecified descriptions.

The current study

To what extent do algorithms like the IA produce descriptions that are psychologically realistic? This question is often raised in the field of REG, particularly when the output of algorithms is evaluated against human corpus data (e.g., Gatt & Belz, 2010). In this chapter, we compare human object descriptions with those of REG algorithms like the IA, where we focus on overspecification: to what extent are automatically generated descriptions comparable to human descriptions in terms of the redundant attributes that they contain?

We expect to find at least one important difference between automatically generated and human target descriptions, and we expect this difference to be related to the visual variation that is present in a scene. There is growing awareness that visual scene perception and speech production are closely related (e.g., Griffin & Bock, 2000; Hanna & Brennan, 2007; Meyer et al., 1998), but little is known about how scene perception influences the production of referring expressions, and, typically, existing REG algorithms such as the IA pay no attention to visual information in a scene. We hypothesize that human speakers are sensitive to *scene variation,* which we operationalize as the number of dimensions in which the objects in a scene differ. For example, reconsider the furniture domain, in which there might be scenes where objects differ in terms of only one dimension (e.g., type), but also scenes in which objects differ in terms of several dimensions (e.g., type, color, orientation and size). Arguably, describing a target in the latter case is a more difficult task, and might therefore cause speakers to include more redundant attributes in their target descriptions (resulting in more overspecification). Based on existing research discussed earlier, we expect this to be at least the case for the preferred attribute color.

The IA does not necessarily include more redundant attributes when the objects in a scene differ across more dimensions. One situation in which this problem becomes apparent is when type is sufficient to distinguish the target: given that, in our domain, type is at the head of the preference order and hence selected first, the IA would not select a (redundant) color attribute, irrespective of the visual variation in a scene.

In order to investigate whether human speakers include color more often in high variation scenes, we performed three experiments in which participants were presented with visual scenes consisting of eight objects including one target, asking them to produce distinguishing descriptions for the target objects. The experiments had two conditions (high and low visual variation), and we made sure that color was never needed for target identification. We study whether human speakers use color more frequently in the high variation condition, and conclude with contrasting our findings with the predictions of the Incremental Algorithm.

Experiment 1

Method

Participants. Participants were 42 undergraduate students who took part in pairs. Twenty-one students (11 female, mean age = 21 years and 7 months) acted as speakers, the other twenty-one as addressees. All participants were native speakers of Dutch (the language of the study) and participated for course credits.

Materials. The stimulus material consisted of artificially constructed pictures of furniture items[1], which have been used before extensively in the field of REG (e.g., Van Deemter et al., 2012a). The furniture items could vary in terms of four attributes and their corresponding values. All possible attribute-value pairs are listed in table 1.

Table 1: Attributes and possible values of the furniture items.

Attributes	Possible values
Type	Chair, sofa, fan, television, desk
Color	Red, blue, green, brown, grey
Orientation	Front, back, left, right
Size	Large, small

The critical trials all contained eight furniture items: one target object and seven distractor objects. The target objects were clearly marked by black borders so that the speakers could easily distinguish them from the distractor objects. The furniture items were randomly positioned on a computer screen in a 2 (row) by 4 (column) picture grid.

Experiment 1 had two conditions. The critical trials in the *low variation condition* were constructed in such a way that there was limited variation

[1] These objects were taken from the Object Databank: http://www.tarrlab.org/

between the target and the distractor objects: the furniture items differed only in terms of the attribute type. In the *high variation condition* the target and the distractor objects differed in terms of all four possible attributes (i.e., type, color, size and orientation). Mentioning type was sufficient to successfully distinguish the target in all critical trials in the two conditions, which implies that including color was never needed to distinguish the target. Note also that the Incremental Algorithm would not include color in either of the two conditions: since including type (which is at the head of the preference order in this domain) is sufficient for distinguishing the target, the IA would not include any further attributes in line (such as color). Fig. 2 depicts examples of critical trials in the two respective conditions.

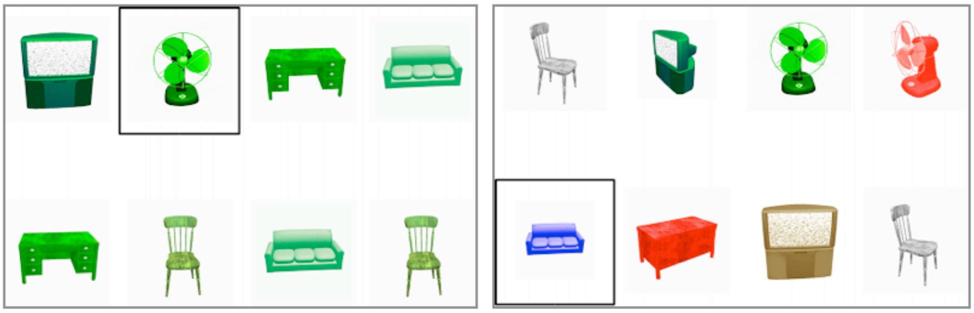


Fig. 2: Examples of critical trials in Experiment 1: for the low variation condition (left picture) and for the high variation condition (right picture).

There were twenty critical trials (ten per condition) and forty fillers. We made one block of sixty trials in a fixed random order (which was presented to one half of the speakers), and a second block containing the same trials in reverse order (which was presented to the other half of the speakers). The fillers consisted of four pictures of Greebles (Gauthier & Tarr, 1997): one clearly marked target referent and three distractor objects, all positioned in a 2 by 2 picture grid. Greebles are complex and difficult to refer to, which made them useful fillers in our experiment. The Greebles could not be distinguished in terms of their color because they were all in the same color every time (so speakers were not primed with the attribute color when describing the fillers).

Procedure. The experiment was performed in an experimental lab. After the two participants had arrived in the room, it was randomly decided who was going to act as the speaker and who as the addressee. Thereafter, the participants were seated opposite to each other. The speaker was presented with the sixty trials on a computer screen, and was asked to describe the target referents in such a way that the addressee would be able to uniquely identify them. There were two practice trials. The instructions emphasized that it would not make sense to include location information in the descriptions, since the addressee was presented with the pictures in a different order. The speaker could take as much time as needed to describe the target, and his or her target descriptions were recorded with a voice recorder. The addressee was presented with the same sixty trials as the speaker in a paper booklet, and was asked to mark the picture that he or she thought the speaker was describing on an answering form. The instructions emphasized that the addressee was – to a limited extent – allowed to ask for clarification: it was allowed to ask the speaker to give more information or to repeat information that had already been given, but not to ask for specific information (i.e., specific attributes). Because of the small number of clarification requests and thus clarifications by our speakers (asking for clarification occurred in only 1.1% of the critical trials), the data presented here should be regarded as initial reference. Once the addressee had identified a target, this was communicated to the speaker, who then went on to describe the next one. After completion of the experiment, none of the participants indicated that they had been aware of the actual goal of the study. All found it an easy task to accomplish.

*Design and statistical analysis***.** Experiment 1 had a within participants design with scene variation (levels: low, high) as the independent variable, and the proportion of descriptions containing a color attribute as the dependent variable. As described above, we made sure that speakers never needed to include color in their target descriptions in order to produce a distinguishing description of the target. Thus, if speakers did mention color anyway, this caused the expression to be overspecified.

Our statistical procedure consisted of two repeated measures ANOVAs: one on the participant means ($F1$) and one on the item means ($F2$).

Results

In total, 420 target descriptions were produced in this experiment. All of these contained a type attribute and were fully distinguishing. Fig. 3 depicts the proportion of expressions that contained a color attribute as a function of the condition in which the descriptions were uttered.

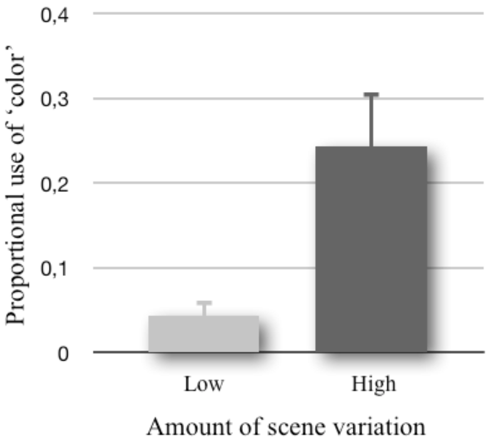


Fig. 3: Results for Experiment 1: the proportion of referring expressions (plus standard deviations) containing a 'color' attribute as a function of the variation in the visual scene.

As hypothesized, the scene variation affected the proportional use of the redundant attribute color ($F1_{(1,20)}$ = 12.537, p < .01; $F2_{(1,18)}$ = 23.416, p < .001). More specifically, speakers were more likely to include color when there was high variation in the picture grid (M = .24, SD = .07) as compared to when this variation was low (M = .04, SD = .02).

Experiment 1 confirmed our hypothesis about the role of scene variation on speakers' tendencies to redundantly include a color attribute in their target descriptions. In the next experiment, we will see whether the same applies when the difference between the low and high variation condition becomes more subtle (in terms of amount of variation).

Experiment 2

Method

Participants. Participants were again Dutch speaking undergraduate students who participated in pairs. This time, there were twenty-two students who acted as speakers (12 female, mean age = 22 years and 4 months). None of these speakers had acted as a speaker in Experiment 1. Another twenty-two students acted as addressees in this experiment. Most of these had been speakers in Experiment 1 or 3, in a few cases the addressee was a confederate.

Materials. There were twenty critical trials in two conditions, and these trials all contained one clearly marked target referent and seven distractor objects. We used the same fillers as before. Again, there was maximum variation between the target and the distractor objects in the *high variation condition* (thus, the objects again differed in terms of the attributes type, color, orientation and size). However, unlike in Experiment 1 (where the objects only had different types), the pictures in the *low variation condition* now varied in terms of three attributes (type, orientation and size) instead of one. This caused the difference between the trials in the two conditions to be more subtle than in the first experiment. Fig. 4 depicts examples of trials in the two conditions of Experiment 2.

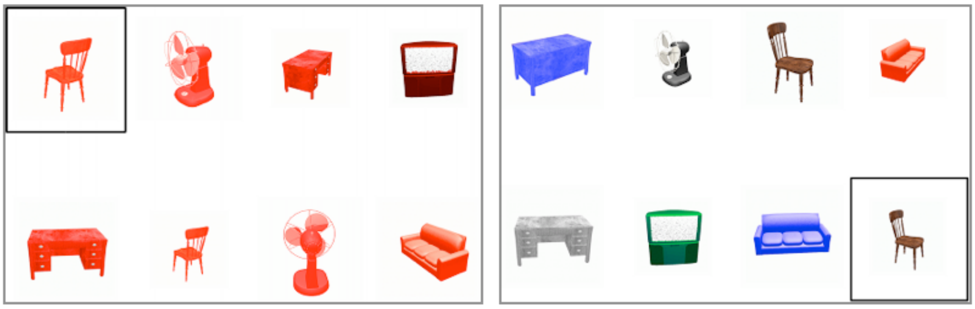


Fig. 4: Examples of critical trials in Experiment 2: for the low variation condition (left picture) and for the high variation condition (right picture).

In all critical trials, mentioning type plus one other attribute (orientation or size, but never color) was sufficient to produce a distinguishing description of the target. For example, in Fig. 4, a speaker could distinguish the target objects in both conditions by mentioning type and size. There were ten trials in each condition, with an equal number of size and orientation trials, and again, mentioning color was never needed to distinguish the target. As in Experiment 1, the trials were built in such a way that the Incremental Algorithm would not include color in its descriptions. In the low variation condition in Fig. 4, the IA would not select color because all objects have the same color. In the high variation condition in Fig. 4, the IA would first select type, because it is at the head of the preference order in this domain. Since both remaining distractors are then brown chairs, the algorithm will not include color and select size instead.

Procedure, design and statistical analysis. As above.

Results

In total, 440 target descriptions were produced in this experiment. All of these contained a type attribute, and most (99.5%) were fully distinguishing. Given the fact that underspecification was very rare in the sample and did not result in significant differences on scene variation, we decided not to further address this topic here. Fig. 5 (on the next page) depicts the proportion of expressions that contained a color attribute as a function of the condition in which the descriptions were uttered.

The general picture of the results of this experiment is comparable to that of the results of Experiment 1. We again found that the amount of variation between the target and the distractors affected the number of times that speakers included the redundant attribute color in their referring expressions ($F1_{(1,21)}$ = 7.092, p < .05; $F2_{(1,18)}$ = 10.515, p < .01). More specifically, the results showed that speakers were more likely to mention color when the variation in the picture grid was high (*M* = .18, *SD* = .06) as compared to when it was low (*M* = .09*, SD* = .05).

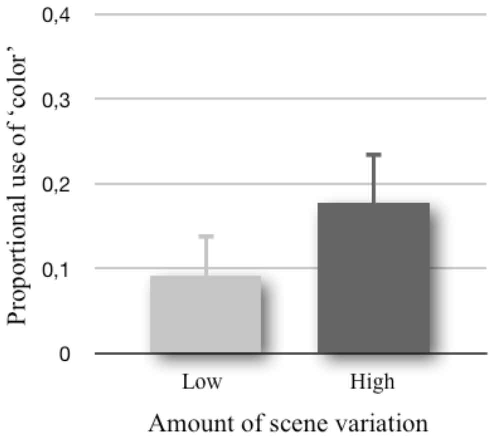


Fig. 5: Results for Experiment 2: the proportion of referring expressions (plus standard deviations) containing a 'color' attribute as a function of the variation in the visual scene.

The results of Experiment 2, like those of Experiment 1, confirmed our hypothesis and indicate that speakers more often redundantly include color in their descriptions when the variation in a visual scene is high. In the next experiment, we will take a closer look at the role of the attribute type: are speakers more likely to redundantly include color when the objects in a scene have different types as compared to when all objects are of the same type?

Experiment 3

Method

Participants. Participants were Dutch speaking undergraduate students, again participating in pairs. There were twenty participants who acted as speakers (14 female, mean age = 22 years and 2 months). None of these participants had acted as a speaker in Experiment 1 or 2. Another twenty students acted as addressees, most of whom had been speakers in Experiment 1 or 2. In a few cases the addressee was a confederate.

Materials. Critical trials and fillers were constructed as above. Here, the crucial manipulation was that all scene objects had the same type in the *low* *variation condition* (e.g., eight chairs), while they had different types in the *high variation condition*. Furthermore, in the low variation condition, the objects varied in terms of their orientation and size, while in the high variation condition, they had different types, colors, orientations and sizes. Fig. 6 depicts examples of trials in the two respective conditions.

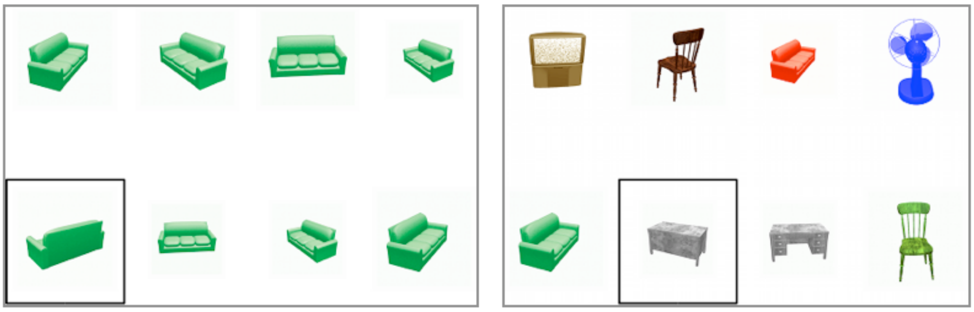


Fig. 6: Examples of critical trials in Experiment 3: for the low variation condition (left picture) and for the high variation condition (right picture).

In all critical trials, mentioning type plus one additional attribute (orientation or size, but never color) was sufficient to uniquely distinguish the target. For example, in Fig. 6, a speaker could distinguish the targets in both conditions by mentioning type and orientation. There were ten trials in each condition, with an equal number of size and orientation trials, and again, color was never needed to produce a distinguishing description. Like in the previous experiments, the critical trials were constructed in such a way that the Incremental Algorithm would not include color in its descriptions. In the low variation condition (left picture in Fig. 6), the algorithm would skip both type and color (since they do not exclude any of the distractors), would then select orientation, and finally would add type to make the description a proper noun phrase. In the high variation condition (right picture in Fig. 6), the IA would first select type. Since both remaining objects in this example are now desks of the same color and size, the algorithm would select orientation instead of color.

Procedure, design and statistical analysis. As above.

Results

In total, 400 target descriptions were produced in this experiment. Most of these contained a type attribute (94.3%), and the vast majority (97.5%) was fully distinguishing. As in Experiment 2, underspecification was rare in the sample, and did not result in significant differences on visual scene variation. Fig. 7 depicts the proportion of referring expressions that contained a color attribute as a function of the condition in which the expressions were uttered.

The results of Experiment 3 show a similar pattern as those of Experiment 1 and 2. We again found that the amount of variation in a visual scene affected the number of times that speakers redundantly included color in their target descriptions ($F1_{(1,19)} = 7.616$, $p < .05$; $F2_{(1,18)} = 20.643$, $p < .001$). Speakers were more likely to mention color when the variation in the scene was high ($M = .27$, $SD = .08$) than when it was low ($M = .10$, $SD = .04$). These results are again in line with our hypothesis that a higher amount of scene variation causes speakers to more frequently mention color redundantly in their descriptions.

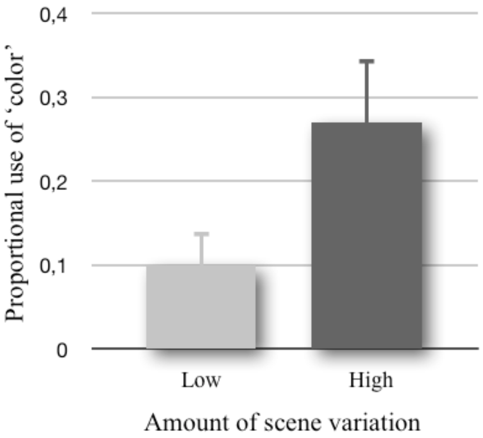


Fig. 7: Results for Experiment 3: the proportion of referring expressions (plus standard deviations) containing a 'color' attribute as a function of the variation in the visual scene.

Meta analysis and speaker variation

Meta analysis

The results of the three experiments reported so far all confirm our hypothesis that human speakers more often redundantly use color when presented with high variation scenes than with low variation scenes. The low variation scenes of our three experiments tested scene variation in different, but related ways: in Experiment 1, the objects differed only in terms of their types, in Experiment 2, there was variation in terms of type, orientation and size, and in Experiment 3, the objects were all of the same type. Did these differences lead to different use of color?

As we have seen, the proportion of referring expressions that contained a color attribute in low variation scenes was slightly lower in the first experiment (M = .04) as compared to the second (M = .09) and third experiment (M = .10). We ran a statistical meta analysis to find our whether these proportions were significantly different, combining data of the three experiments. In order to do this, we performed a 3 x 2 Repeated Measures ANOVA (using Tukey's HSD for multiple comparisons), with Experiment as a between-subjects variable (levels: Experiment 1, 2 and 3) and Condition as a within-subjects variable (levels: low and high variation). As expected, Condition had a significant influence on the redundant use of color ($F_{(1,60)}$ = 26.727, p < .001). However, interestingly, we neither found a main effect of Experiment ($F_{(2,60)}$ = .302, *ns*), nor an interaction between Experiment and Condition ($F_{(2,60)}$ = 1.373, *ns*). This suggests that the effects reported in this chapter generalize over the different manipulations of variation in the low variation conditions.

Speaker variation

Although the results of all three experiments show a similar pattern regarding the effects of condition, it might be that there are more individual differences between speakers in one experiment as compared to another. Arguably, individual differences between speakers are an interesting challenge for researchers in the field of Referring Expression Generation (Dale & Viethen, 2010). Therefore, in order to see whether there was variation between the speakers in our three experiments, we drew scatter plots showing redundant color use of all speakers. More specifically, we calculated the difference between the proportional use of color in the high and the low variation conditions for each speaker. For example, if a speaker mentioned color once when presented with low variation scenes, and five times when presented with high variation scenes, this person scored 4 in terms of the difference between the two conditions. Since there were ten trials in each condition, the scores for each speaker could range from -10 to 10. Fig. 8 depicts the individual scores for all participants that took part in our experiments.

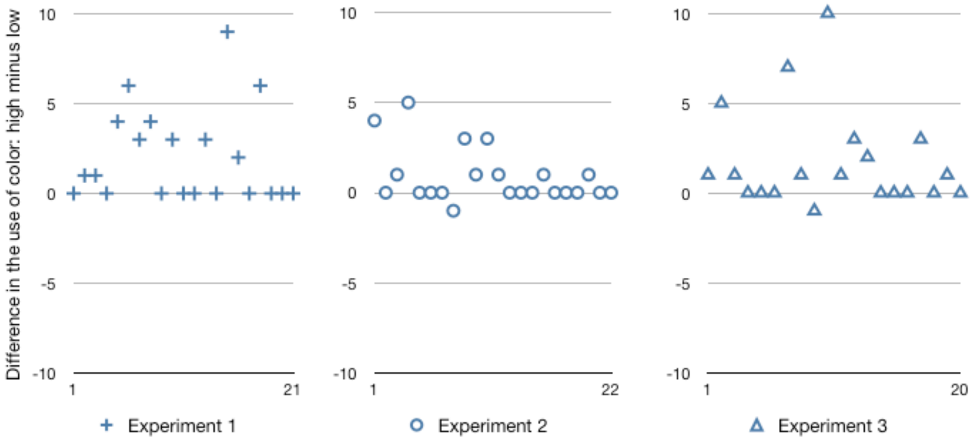


Fig. 8: Speaker variation in the three experiments: the difference in the use of color in the high and the low variation condition (y-axis) as a function of the individual speakers (x-axis). A positive value indicates that a speaker used color more frequently in the high than in the low variation condition.

Fig. 8 shows that the vast majority of the speakers scored between 0 and 5, meaning that these speakers mentioned more (or as many) redundant color attributes in the high variation condition as compared to the low variation condition. Therefore, we can be fairly certain that the effects reported in this chapter are consistent across participants, and that they were not driven by one or two individuals.

General discussion

We have shown that the amount of visual variation in a scene affects the number of times that speakers redundantly include color in their descriptions. In all three experiments, participants mentioned the color of the target more often when this object occurred in a high variation scene than when it occurred in a low variation scene. This was the case when the difference between the low and high variation condition was large (like in Experiment 1), but also when this difference was more subtle (like in Experiment 2), or when the objects in the low variation condition all had the same type (like in Experiment 3). In addition to this, a meta analysis showed that these effects generalized over the different manipulations of scene variation in the low variation conditions. Although the Incremental Algorithm (Dale & Reiter, 1995) was designed as a computational interpretation of the Gricean Maxim of Quantity and uses a preference order that causes it to generate target descriptions that might contain redundant attributes, it would never include color redundantly in any of the trials in the experiments discussed here. Our results show that speakers are sensitive to the amount of visual variation in a scene while the IA is not.

An interesting question, of course, is how an extension of the IA could account for the results presented here. We have seen that Dale and Reiter (1995) argue that type should always be included in a final description, even if it does not rule out any distractors. Thinking along similar lines, one could consider extending the IA by allowing it to always include certain other attributes as well (such as color in high variation scenes). In practice, this could be done by calculating a "variability index" for each particular scene, and deciding, based on this index, whether certain attributes should always be included or not. However, this solution seems rather ad hoc. For one thing, it would predict that speakers *always* include color in the high variation condition, a prediction that clearly is not borne out by the data. Besides that, needless to say, visual variation is likely to be only one of a number of possible factors behind speakers' tendency to overspecify.

So what *do* the speakers in our experiments do? Even though our experiments were not set up to test specific models of the human production of referring expressions, we would like to discuss a search strategy that human speakers might use to decide about which attributes to (redundantly) include in their descriptions, related to the use of heuristics. Tversky and Kahnemann (1974) argue that people rely on *quick heuristics* when making decisions, which they define as "beliefs concerning the likelihood of uncertain events (...) that reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations" (p. 1124). It could well be that speakers also rely on heuristics when producing referring expressions, a suggestion that can also be found in Viethen and Dale (2009), Dale and Viethen (2009), and Van Deemter, Gatt, Van Gompel and Krahmer (2012b). Instead of carefully scanning a visual scene in search of target attributes with high distinguishing value, speakers might simplify the attribute selection process by using other strategies. Although our findings do not allow us to define specific heuristics that might be at play during reference production, we speculate that at least two interacting criteria play a role in this respect: visual saliency and scene gist.

Several studies have shown that speakers tend to include *visually salient* attributes (such as color) in their target descriptions, irrespective of their contrastive value (e.g., Belke & Meyer, 2002; Pechmann, 1989). This seems to be in line with speakers relying on heuristics when they need to describe a target in a scene with much variation: they will select visually salient attributes that immediately grab their attention, without making sure that these attributes do indeed help ruling out distractors. The same is the case with *scene gist*. Based on early work by Friedman (1979), Potter (1976), and in line with several other papers on the psychophysics of vision (e.g., Itti & Koch, 2001; Kremer & Baroni, 2010), Oliva (2005) defines the gist of a visual scene as the representation of that scene, including perceptual levels of processing (e.g., color, spatial characteristics) and conceptual levels of processing (e.g., objects, activation of semantic information). In line with this, Oliva, Torralba, Castelhano and Henderson (2003) suggest that when people need to detect or describe a target object, they are initially guided by heuristics that result from processing the scene on a perceptual level: people are inclined to take the visual context in which the target occurs into account. In this way, heuristics assist in focusing people's attention on relevant information about the target. Similarly, since describing objects requires detecting them first, the gist of a scene could also guide speakers' decisions on which target attributes to include in their referring expressions. For example, if the objects in a scene do not vary in terms of a particular attribute (for example because they all have the same color), the speaker may quickly discard this attribute as an attribute that has any discriminatory power, and may thus not mention it. On the other hand, if a certain attribute attracts visual attention in some way in the context of the target (for example, the attribute color in a scene where objects have different colors), then it is more likely to be used as an attribute within a target description. The situation might become more complex in high variation scenes in which a speaker's attention is grasped by different attributes that vary across the objects in the visual scene. In those cases, it will be less immediately obvious which attributes help in ruling out the distractors, and speakers may therefore be more likely to quickly include a preferred property (such as color). This will, at least in cases as those studied here, lead to overspecification.

One interesting question that remains is how scene variation affects listeners, and what the role of redundant attributes would be in the target identification process. Regarding the latter, some papers claim that listeners are hindered by redundant target attributes during identification (e.g., Engelhardt et al., 2006), while others suggest the opposite and show that overspecification shortens the time needed for identification (e.g., Arts et al. 2011). It seems plausible to assume that the amount of variation in the visual scene plays a crucial role here, since various studies have claimed there to be a close connection between scene perception and comprehension (see Ferreira and Tanenhaus (2007) for an overview). For one thing, in line with Engelhardt, Demiral, and Ferreira (2011), it could be the case that a redundant color attribute facilitates identification in a high variation scene (because it could rule out one or more distractors), while it could possibly distract the listener in the case of a low variation scene. Although this connection between scene variation and comprehension goes beyond the scope of the data presented in this chapter, we assume that it might cause problems for the current REG algorithms, since these are generally designed so as to mimic human speakers as much as possible. However, this does not necessarily cause them to generate descriptions that are of optimal use for the listener (Krahmer & Van Deemter, 2012), or that are adapted to the amount of variation in the visual scene in terms of their level of redundancy. Therefore, one challenge for the future would be to design algorithms that overcome these problems, and to evaluate their output in terms of its benefits for the listener (see Gatt and Belz (2010) for an example of how this could be done).

Conclusion

This study demonstrates that speakers are more likely to include a color attribute in their target descriptions (even if this leads to overspecification) when the scene variation is high than when this variation is low. Our findings are problematic for existing algorithms (such as the Incremental Algorithm) that are often claimed to generate psychologically realistic target descriptions, since these algorithms make use of a fixed preference order per domain and do not take visual scene variation into account.

Acknowledgments

We thank Jette Viethen and two anonymous reviewers for their comments on an earlier version of this chapter, and Tessa Dwyer and Joost Driessen for assistance in collecting the data.

References

Arts, A. (2004). Overspecification in instructive texts. Dissertation, Tilburg University. Wolf Publishers, Nijmegen.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics,* 43 (1), 361-374.

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.

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. MIT Press, Cambridge.

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

Dale, R., & Viethen, J. (2009). Referring Expression Generation through attribute-based heuristics. *Proceedings of the 12th European workshop on Natural Language Generation (ENLG)* (pp. 58-65). Athens, Greece.

Dale, R., & Viethen, J. (2010). Attribute-centric referring expression generation. In Emiel Krahmer & Mariet Theune (Eds.), *Empirical Methods in Natural Language Generation, Lecture Notes in Computer Science* (Vol. 5980), Berlin and Heidelberg: Springer.

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.

Engelhardt, P., Demiral, S., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: an ERP study. *Brain and Cognition*, 77, 304-314.

Friedman, A. (1979). Framing pictures: the role of knowledge in automized encoding and memory for gist. *Journal for Experimental Psychology: General*, 108, 316-355.

Gatt, A., & Belz, A. (2010). Introducing shared task evaluation to NLG: the TUNA shared task evaluation challenges. In E. Krahmer & M. Theune (Eds.), *Empirical methods in Natural Language Generation.* Berlin and Heidelberg: Springer (LNCS 5790).

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 European Workshop on Natural Language Generation (ENLG)* (pp. 49-56). Saarbruecken, Germany.

Gauthier, I., & Tarr, M. (1997). Becoming a Greeble expert: exploring mechanisms for face recognition. *Vision research,* 37, 1673-1682.

Grice, H. P. (1975). Logic and conversation. In: P. Cole, & J. L. Morgan (Eds.), *Speech Acts*. Academic Press, New York.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11, 274-279.

Hanna, J., & Brennan, S. (2007). Speaker's eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57 (4), 596-615.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature reviews neuroscience*, 2, 194-203.

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Computational Linguistics*, 38 (1), 173-218.

Levelt, W. (1989). Speaking: From intention to articulation. MIT Press, Cambridge, London.

Mellish, C., Scott, D., Cahill, L., Evans, R., Paiva, D., & Reape, M. (2006). A reference architecture for Natural Language Generation systems. *Natural Language Engineering,* 12 (1), 1-34.

Meyer, A., Sleiderink, A., & Levelt, W. (1998). Viewing and naming objects: eye-movements during noun-phrase production. *Cognition*, 66, B25-B33.

Oliva, A., Torralba, A., Castelhano, M., & Henderson, J. (2003). Top-down control of visual attention in object detention. *Proceedings of the IEEE International Conference Image Processing, vol. 1*, 253-256.

Oliva, A. (2005). Gist of the scene. In Neurobiology of attention, L. Itti, G. Rees and J. K. Tsotsos (Eds.), Elsevier, San Diego, CA, 251-256.

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

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

Potter, M. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human learning and memory*, 2, 509-522.

Reiter, E., & Dale, R. (2000). *Building Natural Language Generation sytems*. Cambridge University Press.

Ferreira, F. & Tanenhaus, M. K. (2007). Introduction to the special issue on language-vision interactions. *Journal of Memory and Language*, 57, 455-459.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, New Series, Vol. 185, 1124-1131.

Van Deemter, K., Gatt, A., Van der Sluis, I. & Power, R. (2012a). Generation of referring expressions: assessing the Incremental Algorithm. *Cognitive Science,* 36 (5), 799-836.

Van Deemter, K., Gatt, A., Van Gompel, R., & Krahmer, E. (2012b) Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4 (2), 166-183.

Viethen, J., & Dale, R. (2009). Referring Expression Generation: what can we learn from human data? *Proceedings of the 2009 workshop on the production of* *referring expressions: bridging the gap between computational and empirical approaches to reference*. Amsterdam, The Netherlands.

5

How distractor objects affect the redundant use of color: effects of bottom-up and top-down saliency cues

Abstract

In two experiments, we investigate to what extent various visual saliency cues in realistic visual scenes cause speakers to include a redundant color attribute in their definite object descriptions, and in particular how these cues guide speakers in determining which objects in the scene are relevant distractors, and which not. Firstly, the results of our experiments demonstrate that bottom-up, perceptual factors (i.e., distractor color and type, distractor distance, and the presence of visual clutter) affect speakers' use of redundant color attributes. We were unable to observe reliable effects of top-down, conceptual cues (i.e., that speakers use color more often when a target's type is used in the instructions), although our results do hint at an impact of this type of cue as well. Taken together, we argue that our findings are problematic for algorithms that generally aim to generate human-like descriptions of objects (such as the Incremental Algorithm), since these generally select properties that help to distinguish a target from all objects that are present in a scene.

This chapter is based on:

- Koolen, R., Krahmer, E., & Swerts, M. (submitted). How distractor objects affect the redundant use of color in definite reference: Effects of bottom-up and top-down saliency cues.

Introduction

When producing definite object descriptions (such as *"the green chair"*), speakers must constantly decide on the information that they want to include in order to make a *target* object identifiable for their addressee(s). In this respect, many referential tasks require a speaker to distinguish such a target object from one or more *distractor* objects. The properties that are included to identify a target seem to be largely determined by the properties of the distractors. To illustrate this, consider the two scenes depicted in Fig. 1.

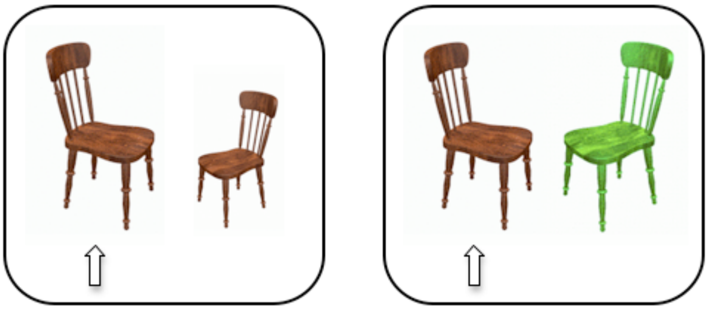


Fig. 1: Two simple visual scenes. Both scenes contain the same target object, but the distractor object is different.

Although the target is the same in both scenes (a large brown chair), the distractor will probably cause a speaker to describe it in different ways. In the left scene, where the distractor is a small brown chair, there is a high chance that a speaker produces a description like *"the large chair"*. However, in the right scene, where the distractor is a large green chair, a description like *"the brown chair"* is more likely to be uttered.

While the distractor object(s) thus appear to play a large role in the production of target descriptions in simple visual scenes such as the ones depicted in Fig. 1 (which involve comparisons of structurally different minimal pairs of objects), the question is what happens when speakers refer to objects in realistic scenes. Arguably, in such scenes, not all objects that are present will always be taken into account as relevant distractors. For example, imagine a speaker who asks a listener to hand her a plate that is lying on a table full of objects. Does the speaker then regard all objects on the table as equally relevant distractors? Or might there be reasons why certain objects are less prominent in or even excluded from the distractor set? And, most importantly, how does this affect reference production?

Algorithms in the field of Referring Expression Generation (REG) do not tend to have an explicit module to determine what the relevant distractor objects are in a visual scene (with the notable exception of Kelleher and Kruijff"s (2006) model, as we explain later on). REG algorithms (of which Dale and Reiter's (1995) Incremental Algorithm is perhaps the best known) are computational models that aim to generate definite descriptions of objects (Reiter & Dale, 2000), focusing on *content planning*: they select attributes with the goal to distinguish a target object from the objects in the distractor set. Given that the number of entities that algorithms are dealing with is generally small, the algorithms usually take all objects in a given domain or scene (except for the target as the distractor set (Dale & Reiter, 1995). This approach might be problematic from a psychological point of view: for example, earlier research has shown that the structure of a *discourse* guides a speaker's focus of attention, and that it therefore restricts the number of objects that are regarded as relevant distractors (e.g., Krahmer & Theune, 2002; Brown-Schmidt & Tanenhaus, 2008).

In this chapter, we explore the role of the *visual context* in the determination of the distractor set during *"one-shot reference"* (where no preceding discourse is involved). We use Itti and Koch's (2000) theory on selective visual attention in doing so: in two production experiments, we systematically manipulate various *perceptual* cues related to bottom-up scene processing (i.e., distractor distance, the presence of clutter objects, distractor type, and distractor color) and one *conceptual* cue related to top-down scene processing (i.e., specificity of the referential task). We take the proportional use of *redundant* color attributes (i.e., attributes that are not necessary for identification, and cause a description to be overspecified) as our dependent variable, expecting to find that speakers are more likely to overspecify with color when they cannot 'calculate' for every object how it can be distinguished from the target: in such cases, speakers might be more inclined to redundantly use color to make sure that their descriptions are distinguishing.

The automatic generation of distinguishing object descriptions

The automatic generation of object descriptions has received considerable attention in Natural Language Generation (NLG), which is a Natural Language Processing task, dedicated to the automatic production of coherent text or speech from a non-linguistic representation (e.g., a database). NLG systems typically have a component that allows them to compute distinguishing descriptions of objects. Because objects can be described in various ways (e.g., even a simple chair can be referred to as *brown*, *large* or a combination of these and other attributes), one basic problem that algorithms are faced with is a problem of choice.

How do the current algorithms solve this problem? Given that the core purpose of reference production is to distinguish a target from its distractors (e.g., Dale & Reiter, 1995), algorithms often take the notion of *discriminatory power* as a starting point, which is determined by the number of distractor objects that can be ruled out by selecting an attribute or a set of attributes. It stands to reason that the algorithms' stop criterion is usually an empty distractor set, meaning that the set of selected attributes can be realized as a fully distinguishing description.

The various REG algorithms use different strategies to generate a distinguishing description: for example, the *Greedy Algorithm* (Dale, 1989) selects the attribute with the highest discriminatory power at every given stage of the generation process, while Dale and Reiter's (1995) *Incremental Algorithm* (IA) uses a preference order (PO) that assumes that some attributes are preferred over others. In this way, the PO can be seen as a ranking of all attributes that can occur in a given domain, where the most preferred attributes (such as *color*) are ranked before less preferred ones (such as *size* or *orientation*). As Dale and Reiter point out, these preferences are typically based on empirical investigation. When generating a target description, the IA iterates through the list of attributes in the PO, selecting relevant values (e.g., brown) if they rule out at least one of the distractor objects (because they have a different value for the relevant attribute, say 'blue'). This process stops when the target object is uniquely identified. It needs to be emphasized that each domain has its own PO, although *color* is known to be preferred in general (e.g., Pechmann, 1989; Viethen, Goudbeek, & Krahmer, 2012; and Chapter 4 of this dissertation).

Determining the distractor set in Referring Expression Generation

Given that REG algorithms generally aim to distinguish a target object from one or more distractors in the context, it is important to have an appropriate formalisation of the notion *context* (Viethen, Dale, & Guhe, 2011). Early work in the field of Referring Expression Generation has used *discourse structure* as a cue to do this: for example, following Grosz and Sidner (1986), Dale (1989) defines the distractor set in terms of discourse-accessible referents. In addition to this, Dale (1992) argues that the number of entities that algorithms are dealing with is generally rather small, which makes it adequate to take the global working set (consisting of any object that can be referred to in a given domain except for the target object) as the distractor set. In line with this, Dale and Reiter (1995) write: "we define the context set to be the set of entities that the hearer is currently assumed to be attending to" (p. 236). Thus, Dale and Reiter do not point out explicitly how the set of distractors should be determined for a scene, and that it may be restricted in certain communicative situations.

From a psychological point of view, however, this may be problematic. Consider the right hand part of Fig. 1 again, and assume that a speaker just referred to the target (e.g., "I prefer the brown chair."). The target chair would be linguistically salient at this point, which means that the second description of it could be reduced (e.g., "This chair would fit nicely in our living room."). Krahmer and Theune (2002) model this computationally by arguing that a description like "the chair" is most likely to refer to the chair that is linguistically most salient in the current context. This means that the other chair in the context (which has not been mentioned before in the discourse) can be ignored as a distractor in the case of the second description, causing a speaker to describe the target chair as "the chair" rather than “the brown chair”. Related to this, earlier work by Brown-Schmidt and Tanenhaus (2008) has shown that speakers often mention attributes that distinguish a target object from recently mentioned distractors, showing that recency is a discourse-related cue that plays a role in the determination of the distractor set as well.

Given that, as mentioned before, it is not explained explicitly how current REG algorithms should determine the distractor set for a given target, Krahmer and Theune (2002) argue that the continuously changing contents of this set have repercussions for REG algorithms such as the IA, and that these algorithms would benefit from an extension that enables them to restrict the distractor set to a proper subset of relevant distractor objects. So the question is: what are the conditions that should be satisfied for an object to be part of the distractor set? While Krahmer and Theune (2002) focus on linguistic salience to answer this question, we expect *visual saliency cues* to play a role in this. For example, when speaker and addressee are only attending to the brown chair that is depicted in Fig. 1 (e.g., because the green chair is positioned far away), the speaker may produce "the chair" as an initial description as well. Put differently, if the speaker were to mention color and thus produces "the brown chair", this suggests that the speaker is taking the green distractor into account.

Although prior work in Referring Expression Generation has indeed suggested that visual saliency cues play a role in determining the distractor set (e.g., Kelleher & van Genabith, 2004; Kelleher & Kruijff, 2006; Van der Sluis, 2005), empirical research that systematically tests to what extent human speakers are driven by visual saliency cues is lacking. Therefore, in the current chapter, we take Itti and Koch’s (2000) model of selective visual attention as a starting point to study how visual saliency affects the production of definite object descriptions.

The impact of bottom-up and top-down saliency cues

In their model of visual attention, Itti and Koch (2000) argue that at least two kinds of visual cues guide a viewer’s attention when they are presented with a visual scene: *bottom-up* (perceptual) cues and *top-down* (conceptual) cues. We predict that both these cues affect the production of reference, and study how they guide speakers in determining the distractor set. We discuss our predictions in more detail below.

Predictions regarding bottom-up scene processing

Itti and Koch (2000) define *bottom-up* saliency cues as image-based cues, and state that an object will pop out of a scene if it is sufficiently salient. This is consistent with early work by Treisman and Gelade (1980) and a more recent paper by Foulsham and Underwood (2008), which showed that perceptual saliency largely predicts viewers' eye movements when looking at a scene. Bottom-up saliency cues can be processed very quickly, and are computed in a pre-attentive manner across the visual field. With regard to reference production, we hypothesize speakers to be driven by at least two kinds of bottom-up cues when determining whether a specific object is considered to be part of the distractor set or not.

Firstly, we expect *distractor distance* (that is, the distance between the target and a potential distractor) to guide speakers in determining the distractor set. For dialogue reference, Beun and Cremers (1998) suggest that a speaker's *focus of attention* limits the number of relevant distractors. Based on their analyses of conversations between participants in a construction task, they found that speakers often produce ambiguous expressions like, for example, *"the yellow one"* when they are presented with a scene containing several yellow objects. This implies that none of these yellow distractor objects were in the speakers' focus of attention, and, as Beun and Cremers argue, that only visually close objects are generally taken to be part of the distractor set. An eye-tracking study by Brown-Schmidt and Tanenhaus (2008) confirmed these findings, and revealed that objects that are positioned close to the last-mentioned target are more likely to be in the speaker's focus of attention than less proximal ones.

Interestingly, researchers in the field of Referring Expression Generation have also used this notion of proximity to compute an object's salience. For example, Van der Sluis (2005) uses proximity to build a three-dimensional notion of object salience (based on Thórisson's (1994) work on perceptual grouping), while Kelleher and Van Genabith (2004) and Kelleher and Kruijff (2006) compute a relative salience for objects in a scene based on their centrality relative to a speaker's focus of attention. However, although the above authors do conduct user evaluations of their algorithms, experimental work on the assumed effect of distractor distance is mostly lacking.

Secondly, we expect *visual clutter* to guide speakers in determining the distractor set. In this chapter, we define clutter as a collection of objects that are thematically related to the target, and assume the amount of clutter to be positively correlated to the number of objects that are present in a scene (following, for example, Bravo and Farid (2008)). In previous research, clutter has for example been shown to have an effect on the type of referring expressions used in route descriptions (Westerbeek & Maes, 2013), and on speakers' response times when they are to describe naturalistic scenes, with slower reactions for cluttered scenes (Coco & Keller, 2009). In line with this, we hypothesize that since a cluttered scene contains more objects (and may thus be more difficult to process), it is unlikely that speakers 'calculate' for every distractor how it can be distinguished from the target object. There, we expect that speakers will rely on mentioning perceptually salient attributes of the target (such as color) when they are presented with a cluttered scene, without making sure that these indeed have distinguishing value.

Predictions regarding top-down scene processing

Besides bottom-up cues, Itti and Koch (2000) argue that also *top-down* cues guide a viewer's attention when looking at a scene. Itti and Koch claim that this top-down mechanism is typically related to the task that a viewer has in looking at a scene. For example, consider the task of describing a plate: a viewer's attention is then directed to a specific object in the visual scene, which requires effort. For this reason, the top-down mechanism is more deliberate and powerful than its bottom-up counterpart: it directs the viewer's attentional focus under cognitive control.

In vision research, various studies have been conducted that measure eye-movements to study how different kinds of top-down cues might guide a viewer's attention. One example is a study by Hegarty, Canham and Fabrikant (2010) that has looked into the effect of background knowledge on participants' interpretation of weather maps. Hegarty et al. eye-tracked participants before and after the provision of declarative knowledge about meteorological principles. The results showed that as long as participants were naïve, their attention was merely guided by perceptual cues, while task-relevant areas of the maps became dominant after the provision of meteorological background knowledge. In a different vein, Einhäuser, Spain, and Perona (2008) have also found an effect of top-down scene processing. In their eye-tracking experiment, they found that viewers usually attend to the objects that are present in a scene rather than to "early" features such as color and contrast.

Previous research that has systematically tested the effect of top-down saliency on language production is scarce. One interesting exception is earlier work of Griffin and Bock (2000), who tracked speakers' eye movements when describing simple events in black-and-white drawings, and found that grammatical subjects are typically selected based on comprehended events rather than perceptual saliency cues. In line with this, Coco and Keller (2012) eye-tracked speakers during the description of photo-realistic scenes, finding similarities between sequences of fixated objects in the scan patterns and sequences of words in the sentences that were produced by the speakers. Based on these results, Coco and Keller argue that when speakers describe a scene, cross-modal coordination processes (including the retrieval of relevant context information) go beyond coordination based on low-level features of objects in the visual scene. In a similar vein, Arts, Maes, Noordman, and Jansen (2011) made speakers refer to objects in an experiment manipulating the importance of the instructed referential task. Arts et al. found that the descriptions uttered in their high-importance condition (in which participants were told that their addressee was a surgeon who fully depended on the descriptions for successful performance of the surgery) were more often overspecified than the descriptions uttered in the low-importance condition (where participants were simply asked to identify objects). As Arts and colleagues argue, these results suggest that task importance is a strong indicator for the degree of overspecification in object descriptions.

In this chapter, we introduce a manipulation of top-down processing that is related to the task that speakers need to perform when they are to describe a target in a visual scene. More specifically, we expect the *specificity of the referential task* to affect the production of definite reference, and in particular how conceptual cues guide speakers in determining the set of relevant distractors in a scene. We expect that a general task (such as "describe this object") will leave speakers with a bigger, less restricted set of distractors than a more specific task (such as "describe this plate", where the target's type is mentioned). In the latter case, speakers might leave objects other than plates unattended, while any object that is present in the scene may be regarded as a relevant distractor in the former case.

The current study

We report on the results of two reference production experiments in which we presented participants with photo-realistic scenes, asking them to describe one target in such as way that their listener could distinguish it from all the other objects that were present in that particular visual scene. Our study concerns the production of so-called *"one shot"* reference, allowing us to formulate implications for Referring Expression Generation algorithms such as the IA (Dale & Reiter, 1995).

The dependent variable in our experiment was the extent to which bottom-up and top-down saliency cues cause speakers to include a *redundant color attribute* in their descriptions (that is, a color attribute that is not needed for identification). Therefore, in our stimulus materials, we made sure that color was never needed to distinguish the target from its distractors; mentioning the target's *type* and *size* was always sufficient. This is consistent with Chapter 4 of this dissertation, where we used the redundant use of color to study differences in how speakers perceive low- and high-variation scenes. It needs to be mentioned that the IA would *never* select color in our stimuli either (with the same Preference Order), except in one specific condition (where it would *always* select color). We come back to this in the General Discussion.

So what are our predictions regarding the redundant use of color? Regarding the *bottom-up* saliency cues discussed earlier, our hypotheses are as follows. Firstly, there might be an effect of *distance*, where we hypothesize that redundant color attributes might be used more often when a potential distractor is positioned close to the target as compared to when it is distant. Furthermore, we hypothesize there to be an effect of *clutter*, with a positive association between the presence of clutter and the redundant use of color, because cluttered scenes contain more objects and might therefore be more difficult to process. Secondly, regarding *top-down* saliency, we expect to find an effect related to the *specificity of the referential task*: we hypothesize that when speakers are instructed in a general way (i.e., "Describe this object"), they are more likely to use a redundant color attribute as compared to when the instruction includes the target's type (e.g., "Describe this plate").

Experiment 1

In our first experiment, we study to what extent *distractor distance*, *visual clutter*, and the *specificity of the referential task* affect speakers' redundant color use in object descriptions, and their determination of the distractor set.

Method

Participants. Participants were 43 undergraduate students (30 female, 13 male) from Tilburg University. All participants (mean age 21 years and 1 month, range 18 - 34 years) were native speakers of Dutch (the language of the study) and participated for course credits.

Materials. The stimulus materials consisted of 80 trials, all of which were photo-realistic pictures of objects on either a kitchen table or an office desk. The participants were asked to describe one of these objects (the target, which was clearly marked by a white arrow) in such a way that the listener could distinguish it from the other objects on the table or desk. In the 40 critical trials, there were always at least three objects present on the table: one target object and two distractor objects. Crucially, one of the distractors had the same type and color as the target object (meaning that it could only be ruled out by means of its size), and was always positioned next to the target object (either left or right). The second distractor always had a different type and color as compared to the target, and its positioning varied across conditions. Besides that, we manipulated two other principal factors: one related to the presence of clutter in the scene, and one related to the specificity level of the instruction that was given to the speakers.

As mentioned above, the first manipulation (hence called *distractor distance*) was related to the distance between the target and the second distractor. This distance was manipulated as follows: in half of the trials, the distractor was positioned close to the target (with the target and its two distractors placed in the same corner of the table), whereas this distance was maximized in the other half of the trials (with the target and the first distractor in one corner of the table, and the second distractor in the opposite corner). This manipulation is illustrated in Fig. 2 on the next page, where the left pictures have a *close* distractor, and the right pictures have a *distant* distractor. In scenes with a distant distractor, this object was always positioned in the corner opposite the target. Note that mentioning the target's type and size was sufficient to identify the target in both the close and distant conditions, implying that the use of color would inevitably result in an overspecified description.

The second manipulation was related to whether or not there was clutter present in the visual scene. We call this factor *clutter presence*, where clutter is defined as a collection of all kinds of objects that are thematically related to the target and its two main distractors. These object were not systematically varied, and were unique in the sense that they all had different types. The color of these clutter objects was kept as neutral as possible; it was at least made sure that the clutter objects did not have the same color as the target and the two distractors. In the cluttered pictures in

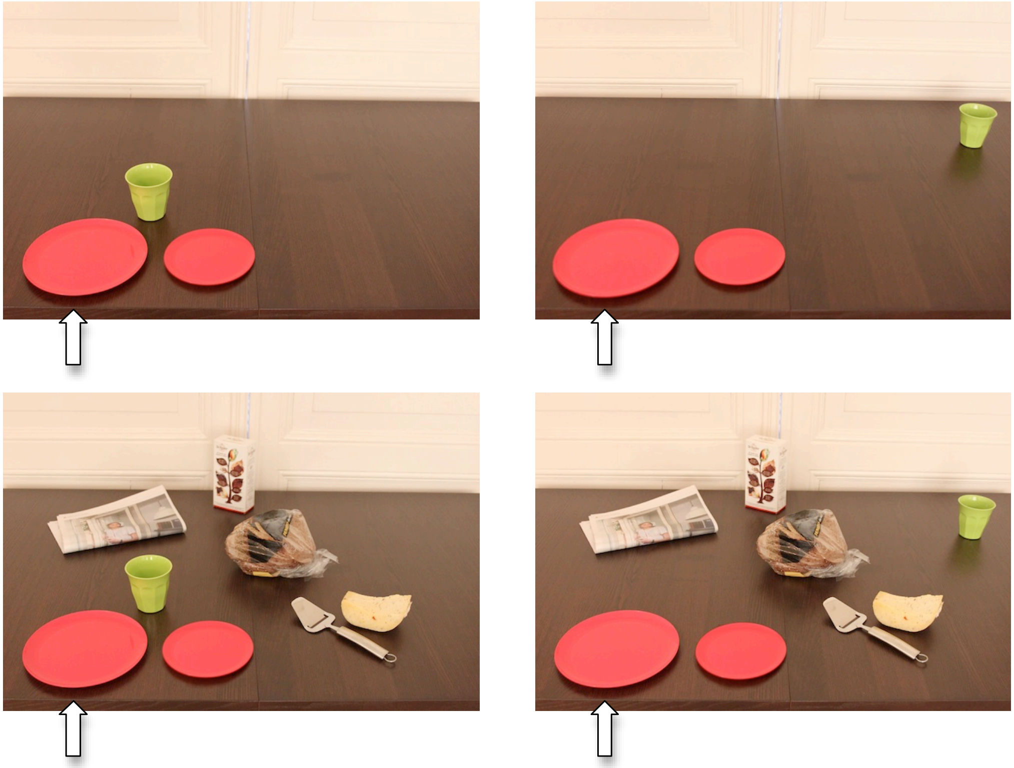


Fig. 2: Examples of critical trials in Experiment 1. The left scenes have a close distractor, whereas this distractor is distant in the right scenes. The upper scenes do not contain clutter, whereas the lower scenes do. Note that both the small and the large plate could be the target.

Fig. 2, five objects are added that one would expect to see on a breakfast table, where most do not have a salient color: a bag of bread, a newspaper, a piece of cheese, a cheese slicer, and a pack of chocolate sprinkles. Note that the green cup in the bottom scenes of Fig. 2 was a distractor, and not part of the clutter. Clutter was added in half of the critical trials, and the same clutter objects were used for the scenes with a close and a distant distractor, and the positioning of the clutter was always identical.

The experiment had eighty trials: forty critical trials and forty fillers. Regarding the critical trials, we used ten scenes: five scenes with objects on an office desk and five scenes with objects on a kitchen table. These ten scenes were all manipulated in four within-conditions: one picture with two close distractors but without clutter, one with a close and a distant second distractor without clutter, one with two close distractors and with clutter, and one with a close and a distant distractor and with clutter. Note that the target could be positioned in all four corners of the table (and not necessarily in the left bottom corner, as is the case in Fig. 2). Since there were always two similar objects in a scene (one of which being the target), we marked the small object as the target in half of the scenes, and the large object in the other half of the scenes.

Besides distractor distance and clutter presence (both being manipulated as within participants factors), the experiment also had one between participants variable (hence called *specificity of the referential task*), which was related to the specific instruction that was given to the participants. As mentioned earlier, it was the participants' task to describe each target in such a way that it could be distinguished from the other objects in the scene. All participants were presented with the same stimuli, but two kinds of instructions were manipulated by using two different pre-recorded requests that were played for every new trial. Half of the participants heard the request to "*Describe this object*" (which means that they took part in the *low specificity condition*), whereas the other half of the participants took part in the *high specificity condition* and heard a more specific request in which the target's type was mentioned. For example, for the target depicted in Fig. 2 the request would be "*Describe this plate*".

The experiment had forty fillers: twenty from the kitchen table domain and twenty from the office desk domain. These fillers were set up in the same way as the critical trials, in the sense that there were scenes containing few objects that were positioned in the same way as those in the critical trials, and scenes containing many different objects (in line with the clutter scenes that served as critical trials). Again, one of the objects was marked as the target and had to be referred to by the participants, with the crucial difference that the objects in the filler pictures did not differ in terms of their color. In this way, in order to avoid a response strategy, speakers were discouraged from using color when describing the fillers.

Procedure. The experiment was performed in an experimental laboratory, with an average running time of 10 minutes. After participants had entered the lab, they were randomly assigned to one of the two conditions: 21 participants took part in the low specificity condition, and 22 in the high specificity condition. Thereafter, they were seated opposite the listener (who was a confederate of the experimenter), and were instructed so as to describe the target objects in such a way that their listener could uniquely identify them. Speakers could take as much time as they needed to do this, and their descriptions were recorded with a voice recorder.

The trials were presented to participants on a computer screen. We made one block of eighty trials in a fixed random order (which was presented to one half of the participants), and a second block containing the same trials in the reverse order (which was presented to the other half of the speakers). There were two practice trials. The listener had a paper booklet in front of her, containing - for each trial - separate pictures of all the objects that occurred in that given scene. These pictures were taken from the pictures the speaker was presented with. Based on the speaker's descriptions, the listener marked the object that she thought was referred to on an answering form. In order to prevent speakers from including location information in their target descriptions (e.g., *'The plate in the left bottom corner'*), the instructions emphasized that the listener was presented with the same objects ranked in a different order. The listener always acted as though she understood the descriptions, and never asked clarification questions. This was done to enable a focus on content planning of initial descriptions ('first mentions'). Once the listener had identified a target, this was communicated to the speaker, who then went on to describe the next target. After they had completed the experiment, none of the participants indicated to have been aware of the actual goal of the study. All found it an easy task to accomplish.

Design and statistical analysis. The experiment had a 2 x 2 x 2 design (see table 1 on the next page) with two within participants factors: *distractor distance* (levels: close, distant) and *clutter presence* (levels: no clutter,

clutter), and one between participants factor: *instruction specificity* (levels: low, high). The experiment had one dependent variable: the proportion of descriptions containing a color attribute. As described above, we made sure that speakers never needed color in order to distinguish the target from its distractors: mentioning the target's type and size was always sufficient. Thus, when speakers mentioned color, this always resulted in an overspecified description.

Table 1: Overview of the experimental design and the number of descriptions within each cell in Experiment 1.

	No clutter		Clutter	
	Close	Distant	Close	Distant
Low	210	210	210	210
High	220	220	220	220

Our statistical procedure consisted of Repeated Measures ANOVAs: one on the participant means (F_1) and one on the item means (F_2). We only report on interactions where these are significant. To compensate for departures from normality, we applied a standard arcsin transformation to the proportions before running the ANOVAs. For the sake of readability, we report the untransformed proportions in the results section.

Results

In total, 1720 target descriptions were produced in this experiment. All of these contained a type attribute, and most (85.8%) a size attribute. In the remainder of the cases, other additional attributes were mentioned to distinguish the target (such as orientation). All descriptions were fully distinguishing, and color was mentioned in 39% of the descriptions.

Results for distractor distance. The first factor that was hypothesized to influence speakers' redundant use of color was distractor distance: whether the distractor with the different type and color was placed close to or far from the target. Fig. 3 depicts the proportion of target descriptions that contained a color attribute as a function of distractor distance.

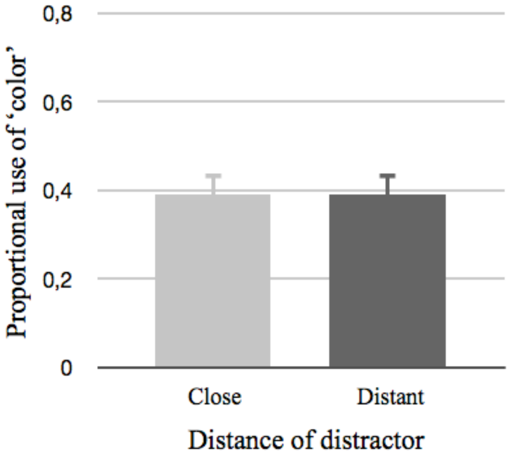


Fig. 3: The proportion of target descriptions (plus standard deviations) containing a color attribute as a function of distractor distance.

As can be seen in Fig. 3, distractor distance did not affect the proportional use of the redundant attribute color ($F1_{(1,41)}$ = .068, p = .80; $F2_{(1,36)}$ = .00, p = .99). More specifically, speakers mentioned color exactly as much when the distractor was close (M = .39, SD = .05) as compared to when it was distant (M = .39, SD = .05).

Results for clutter presence. We also looked how the presence of clutter in a scene affected the speakers' redundant use of color. Fig. 4 (on the next page) depicts the proportional use of color as a function of clutter presence. As can be seen in Fig. 4, the presence of clutter positively affected the redundant use of the attribute color ($F1_{(1,41)}$ = 13.38, p = .001; $F2_{(1,36)}$ = 3.91, p = .06). In other words, speakers were more likely to include color when presented with visual scenes containing clutter (M = .43, SD = .05) as compared to when the scene did not contain clutter (M = .35, SD = .06).

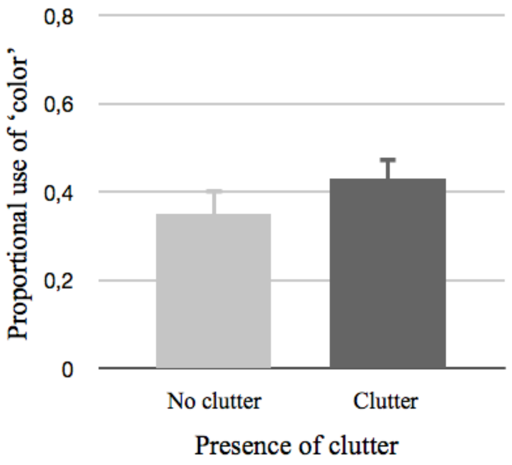


Fig. 4: The proportion of target descriptions (plus standard deviations) containing a color attribute as a function of clutter presence.

Results for specificity of the referential task. The third factor that we manipulated was related to the instructions that were given to the participants. Fig. 5 depicts the proportional use of color as a function of the specificity of the referential task.

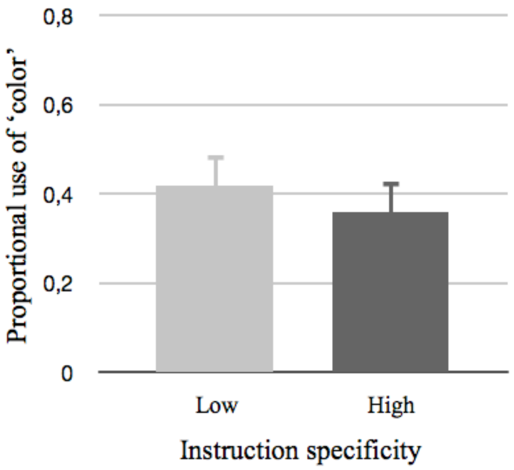


Fig. 5: The proportion of target descriptions (plus standard deviations) containing a color attribute as a function of instruction specificity.

As reflected in Fig. 5, the specificity of the referential task to some extent affected the redundant use of color, but this effect was only significant by items ($F1_{(1,41)}$ = .355, p = .55; $F2_{(1,36)}$ = 15.81, p < .001). Speakers who took part in the low specificity condition (M = .42, SD = .07) used color more frequently as compared to those taking part in the high specificity condition (M = .36, SD = .07), although we did not find a reliable difference between general and more specific instructions.

Discussion

In this first experiment, we have investigated the role of distractor objects in the production of object descriptions, and, in particular, to what extent several *bottom-up* and *top-down* saliency cues cause speakers to mention a redundant color attribute.

Firstly, related to bottom-up scene processing, we have found that - at least for the scenes used here - the distance between the target and the distractor does not affect the redundant use of color in object descriptions. This is not what we predicted based on earlier work on the production of dialogue reference (e.g., Beun & Cremers, 1998; Brown-Schmidt & Tanenhaus, 2008), and may be an artefact of the experimental set up: since our speakers knew that the addressee was presented with – for every scene – separate pictures of all objects that were depicted in that scene (meaning that the distance between the target and a distractor was always the same from the addressee's perspective), this may have caused them to ignore the distance between the target and the distractors. In Experiment 2, we aim to solve this issue.

Secondly, again related to bottom-up processing, we have found that speakers are more likely to redundantly mention color when there are clutter objects present in the scene as compared to when this is not the case. One explanation for this may be that a scene with clutter simply contains more objects than a scene without clutter. As we have suggested earlier, this may lower the chance that speakers exactly 'calculate' for each distractor how it can be distinguished from the target object in the most efficient way. Our results suggest that speakers when process cluttered scenes, they may be more inclined to rely on *heuristics* (Tversky & Kahneman, 1974): when they have to produce a unique description of a target object, they might use color just to rule out at least some of the distractor objects. We will further elaborate on this in the General Discussion.

Thirdly, regarding the specificity of the referential task, we expected to find that participants in the low specificity condition (who were instructed to "describe this object") would be more likely to redundantly mention color than participants that took part in high specificity condition (who were asked to "describe this X", for example "this plate"), since in the latter case (where the target's type was mentioned) only the distractor with the same type would remain to be ruled out (which could always be done by mentioning *type* and *size*). We indeed found a numerical difference between the two conditions, but this was not statistically reliable. We plan to further study this effect in Experiment 2.

In a follow-up experiment (which we present in the next section), we aim to further improve our manipulations of distance and task specificity. Besides that, to obtain a better understanding of visual cues that may guide speakers in determining the set of distractors for a scene, we also systematically vary two additional bottom-up factors, namely the type and color of a distractor. With regard to *type*, we expect to find that an object that has the same type as that of the target is more likely to be part of the set of relevant distractors than an object of a different type. The reasoning behind this is intuitive: if a speaker wants to refer to, for example, a plate, then other plates are more likely to serve a relevant distractors than, say, a mug or a tea cup; in the latter case, mentioning type would already be sufficient to distinguish the target. Therefore, we expect speakers to use more redundant color attributes when a specific distractor and the target have the same type as compared to when they have different types.

With regard to *color*, we expect that an object with a different color as compared to the target is more likely to be considered as a relevant distractor than an object that has the same color as compared to the target. In the latter case, the object may be less likely to catch a speaker's attention. In prior research, color has already been found to have a high perceptual power: for example, it has been shown that speakers usually use color more often when there is color variation in a scene as compared to when this is not the case (see Chapter 4 of the current dissertation), and, for the case of dialogue reference, that speakers often include color attributes irrespective of the visual scene (Brown-Schmidt and Konopka, 2011). Therefore, we expect speakers to use more redundant color attributes in situations where the target has a different color as compared to a potential distractor object.

Thus, in the second experiment, we reconsider the effect of distractor distance on the redundant use of color, and investigate in addition to what extent an object's type and color have an effect. In order to keep the design of the second experiment within reasonable limits (and also to reduce experimental time), the effect of visual clutter is not reconsidered.

Experiment 2

Method

Participants. Participants were 26 undergraduate students (20 female, 6 male) from Tilburg University. None of these had taken part in Experiment 1. All participants (mean age 22 years and 2 months, range 18 - 27 years) were native speakers of Dutch and participated for course credits.

Materials. The stimulus materials consisted of 88 trials, which were comparable to those in the first experiment. Again, participants were presented with photo-realistic pictures of objects on a kitchen table or an office desk, and were instructed to describe one target object so that an addressee could distinguish it from its distractors. Since there were no conditions manipulating clutter in this experiment, there were always three objects present in the critical trials: one target object and two distractors. Like in the previous experiment, the first distractor was always positioned next to the target, and, could always be distinguished by mentioning *type* and *size*. The within variables in this experiment (manipulating bottom-up saliency) were again related to the second distractor, this time varying the distractor's type, color, and distance.

The first manipulation (hence called *distractor type*) was related to the type of the second distractor in the scene: this could be either different or similar to the target's type. For example, consider the example trials in Fig. 6 (on page 140), where the second distractor (the mug) has a different type as compared to the target object (the plate) in the upper four trials, whereas all objects have the same type in the lower four trials.

As can be seen in Fig. 6, our second manipulation was related to the *color* of the second distractor: in half of the critical trials, this distractor had a different color as compared to the target, whereas all three objects had the same color in the other half of the trials. Third, we manipulated *distractor distance* in the same way as in the first experiment: in half of the critical trials (see for example the left pictures in Fig. 6), the second distractor was positioned close to the target (where all objects were placed in the same corner of the table), whereas this distance was maximized in the other half of the trials (where the target and the first distractor were placed in one corner of the table, with the second distractor in the opposite corner). Note that the upper two trials in the figure were taken from the first experiment, and that color was never needed to distinguish a target from its distractors (*type* and *size* were always sufficient).

The experiment had one between variable related to top-down saliency, which was again the *specificity of the referential task*. As explained in the materials section of the first experiment, this variable was related to the kinds of instructions that were used: one half of the speakers took part in the low specificity condition and was presented with the request to *"describe this object"*, whereas the other half was assigned to the high specificity condition and heard a more specific request in which the target's type was mentioned (for example, "*describe this plate"*). Like in the previous experiment, these requests were played aloud for all trials.

In this experiment, we expect to find a significant interaction between the type of a distractor and the specificity of the task rather than a main effect of task specificity. In half of the conditions of the current experiment, the target and its two distractors were of the same type (for example, three plates). In these cases, one would not expect to find an effect of specificity of the task: even if the instruction includes the target's type (e.g., *"describe this plate"*), both distractor objects (the two other plates) still remain to be ruled out. This was not the case in the first experiment, in which one of the two distractors always had a different type as compared to the target object (meaning that the instructions always ruled out one distractor object). We randomly assigned every participant to either one of the two instruction conditions. This resulted in a total of 14 participants in the low specificity condition, and 12 in the high specificity condition.

Regarding the critical trials, we used six scenes that were randomly selected from the ten scenes that were used in the first experiment: three scenes with objects on a kitchen table and three scenes with objects on an office desk. These scenes were all manipulated for the eight conditions described above (see Fig. 6 for an example). The scores for one scene were not included in the analyses[1]. Therefore, in total, there were 40 critical trials. Furthermore, there were 40 fillers, the majority of which were also used in Experiment 1. Some fillers were replaced since they were initially designed so as to share characteristics with scenes containing clutter (which was necessary in the first experiment, but not in the current experiment). We made sure that the objects in the fillers did not have a specific color, so that speakers were discouraged from using color when describing the fillers.

[1] This was done because we found that the participants' descriptions for this scene were rather different as compared to those for the remaining five scenes. Crucially, the *F*1 analyses with this scene included showed the exact same picture as compared to the *F*1 scores that we now report in the results section. However, naturally, this was not the case for the *F*2 scores. We conjecture that this is due to the fact that the target in this scene did not have a clear shape, contrary to the targets in the other five scenes.

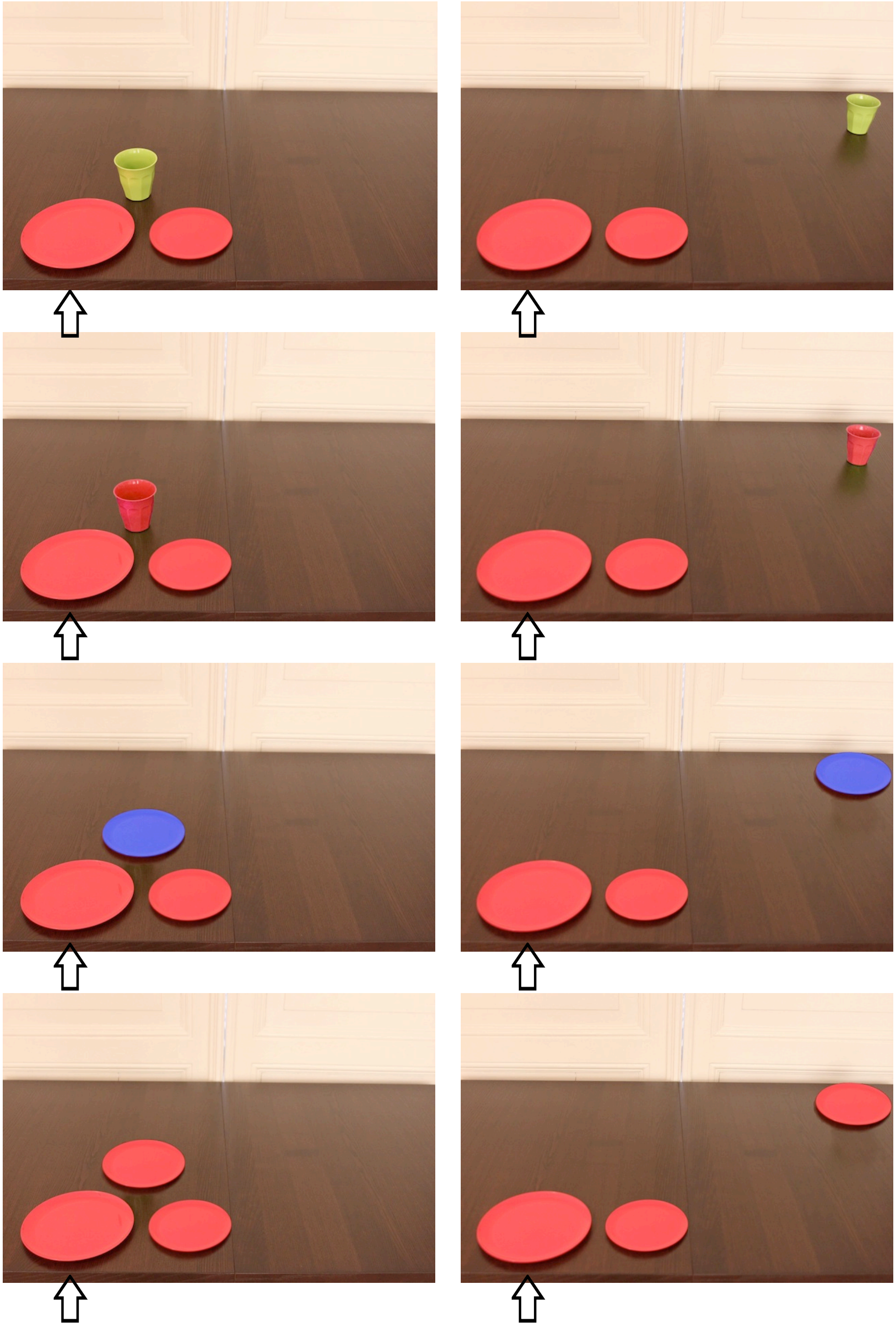


Fig. 6: Examples of critical trials in Experiment 2. Note that both the small and the large plate could be the target.

To summarize Fig. 6: the crucial distractor has a different type than the target in the upper four pictures, whereas this type is the same in the lower four pictures. The crucial distractor has a different color than the target in the first, second, fifth and sixth picture, whereas this color is the same in the third, fourth, seventh, and eighth picture. The crucial distractor is close to the target in the left scenes, and the distractor is distant in the right scenes.

Procedure. The procedure was basically the same as in the first experiment, with one important difference related to the pictures that the addressee had in front of her. In the first experiment, the addressee had a paper booklet containing separate pictures of all objects. However, as discussed earlier, this might have ruled out the effect of distractor distance. Therefore, this time, the addressee was presented with the same trials as the speaker on a screen, where she was asked to use the computer mouse to click on the object she thought the speaker was talking about. However, this time, the speaker was told that the listener was presented with the same objects depicted in the same configuration, except that the pictures were mirror images of the speakers' scene (where mirroring could either occur along the X- or the Y-axis). In this way, we changed the manipulation of distance as compared to the previous experiment; in the current experiment both speakers and addressee perceive the same distance between target and distractor, but location phrases (e.g., "The plate in the left-bottom corner") were still avoided. After the experiment, none of the participants indicated to have been aware that the listener had actually been presented with the same pictures rather than mirror images.

Design and statistical analysis. The experiment had a 2 x 2 x 2 x 2 design (see table 2 on the next page). There were three within factors: *type* (levels: different, same), *color* (levels: different, same) and *distractor distance* (levels: close, distant), and one between factor: *specificity of the referential task* (levels: low, high). As in the first experiment, our dependent variable was the proportion of descriptions containing a color attribute. It was again made sure that mentioning type and size was always sufficient, and that color was never needed to distinguish the target from its distractors. Thus, when a speaker mentioned color, this always resulted in an overspecified description.

Table 2: Overview of the experimental design and the number of descriptions within each cell in Experiment 2.

	Different type				Same type			
	Different color		Same color		Different color		Same color	
	Close	Distant	Close	Distant	Close	Distant	Close	Distant
Low	70	70	70	70	70	70	70	70
High	60	60	60	60	60	60	60	60

Again, our statistical procedure consisted of Repeated Measures ANOVAs on the participant means (F_1) and on the item means (F_2). To compensate for departures of normality, we applied an arcsin transformation to the proportions before running the ANOVAs, but report the untransformed proportions in the results section. We report on interactions where these are significant.

Results

In total, 1040 target descriptions were produced in this experiment. Again, all of these contained a type attribute, and most (99.8%) contained a size attribute. In the remaining 0.2% of the cases, other attributes were mentioned to distinguish the target referent. All descriptions were fully distinguishing, and color was used in 48% of the cases.

We first analysed whether there was an effect of the specificity of the referential task, but we did not find an effect with this factor involved: there was no main effect ($F1_{(1,24)}$ = .014, p = .91; $F2_{(1,32)}$ = 1.55, p = .26), and also the predicted interaction between instruction specificity and distractor type was not significant ($F1_{(1,24)}$ = .765, p = .39; $F2_{(1,32)}$ = 1.07, p = .31). For this reason, we do not take task specificity into account in the remainder of this results section, but we come back to it in the General Discussion.

Results for distractor type. Firstly, we studied the extent to which the type of the second distractor affected the speakers' redundant use of color, where we compared conditions in which this type was either different or similar to the target's type. Fig. 7 shows the proportion of descriptions that contained a color attribute as a function of distractor type.

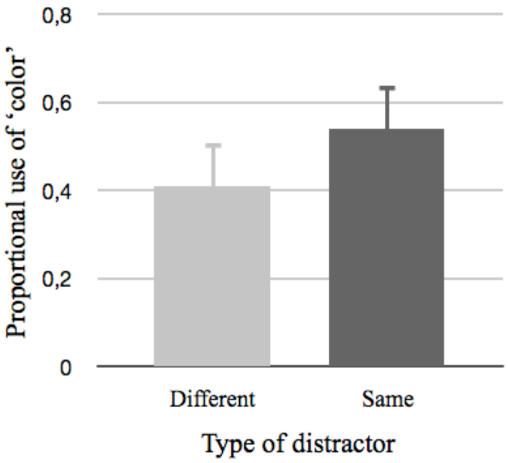


Fig. 7: The proportion of target descriptions (plus standard deviations) containing a color attribute as a function of distractor type.

As can be seen in Fig. 7, the distractor's type influenced the redundant use of the attribute color ($F1_{(1,24)}$ = 17.26, p < .001; $F2_{(1,32)}$ = 16.07, p < .001). More specifically, this means that speakers more often used color when the distractor's type was similar to the target's type (M = .54, SD = .10) as compared to when it was different (M = .41, SD = .10).

Results for distractor color. Secondly, we investigated to what extent the distractor's color influenced speakers to redundantly use color, comparing conditions where this color was either different or similar as compared to the color of the target. Fig. 8 depicts the proportional use of color as a function of distractor color.

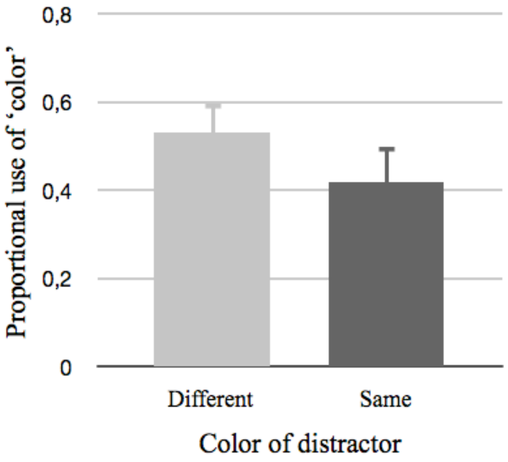


Fig. 8: The proportion of target descriptions (plus standard deviations) containing a color attribute as a function of distractor color.

As reflected in Fig. 8, the distractor's color had an effect on speakers' use of color ($F1_{(1,24)}$ = 16.91, p < .001; $F2_{(1,32)}$ = 14.11, p < .001). In other words, speakers were more likely to include color when the target and the distractor had a different color (M = .53, SD = .07) as compared to when this color was identical (M = .42, SD = .08).

This effect of distractor color was stronger when the target and the distractor both had the same type, as reflected in an interaction between distractor type and distractor color with respect to the redundant use of color ($F1_{(1,24)}$ = 27.88, p < .001; $F2_{(1,32)}$ = 15.10, p < .001). In the conditions in which the target and the distractor had the same types, but different colors, speakers were most likely to use color (M = .66, SD = .06) as compared to the conditions where the target and the distractor shared the same type and color (M = .42, SD = .07). However, the effect of color disappeared when the target and the distractor had different types: in such cases, there was no difference between conditions in which the target and the distractor had either different colors (M = .41, SD = .08) or the same color (M = .41, SD = .08).

Results for distractor distance. Last, like in the first experiment, we tested to what extent the proportional use of color was affected by distractor distance: whether this distractor was placed close to or far from the target. Fig. 9 depicts the proportion of descriptions that contained a color attribute as a function of distractor distance.

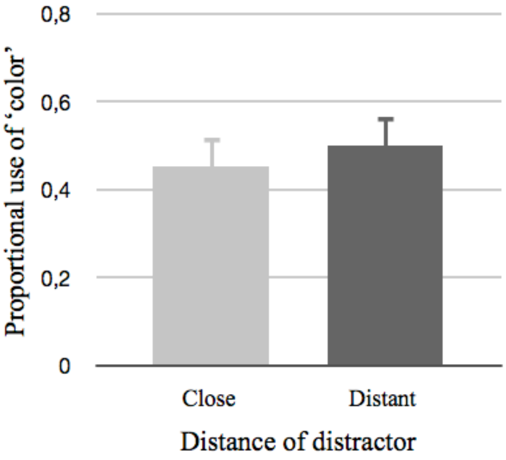


Fig. 9: The proportion of target descriptions (plus standard deviations) containing a color attribute as a function of distractor distance.

As can be seen in Fig. 9, distractor distance affected the proportional use of color, but this effect was only significant by participants ($F1_{(1,24)}$ = 7.64, p = .011; $F2_{(1,32)}$ = 2.72, p = .11). Although, somewhat surprisingly, speakers included color more often when the distractor was distant (M = .50, SD = .07) as compared to when it was close (M = .45, SD = .07), we did not find a reliable difference between close and distant distractor objects.

Discussion

The goal of this second experiment was to investigate how various manipulations related to bottom-up and top-down scene processing cause speakers to redundantly include a color attribute in their definite object descriptions. The results have revealed several interesting findings.

Firstly, we found that speakers were more likely to redundantly include color when the target and a distractor had the same *type*, as compared to when their types were different. One explanation for this could be that speakers are more likely to actually consider an object as a distractor (thus, as an object that needs to be ruled out) when this object shares its type with the target object that they have to describe. Hence, this means that the actual distractor set was bigger when all objects in the visual scene had the same type, making it more likely for speakers to include redundant attributes (such as color) in their descriptions.

Further support for the above line of reasoning comes from the interaction that we found between the *type* of the distractor and its *color*. In this respect, it needs to be mentioned first that we found a main effect of *distractor color*, meaning that speakers were more likely to mention color when the target object and its second distractor had different colors as compared to when they shared the same color. For one thing, this is in line with earlier findings saying that the amount of color variation in a visual scene positively affects the redundant use of color in speakers' descriptions (see Chapter 4 of the current dissertation). However, considering the interaction between the type of the distractor and its color, it might be the case that an additional process is going on in our data: we believe that the nature of this interaction (being that the effect of color was stronger when the target and the distractor shared the same type) suggests that the size of the distractor set might play a role here as well. As we suggested above (when discussing the effect of distractor type), an object may be more likely to be considered as an actual distractor if it shares its type with the target object. Hence, the interaction between distractor type and distractor color reported here seems to add that the effect of color was subsumed by the effect of type: in cases where the target and a potential distractor object had different types, speakers appeared to pay no attention to this object, even if the two objects had different colors.

Lastly, like in the first experiment, we did not find a statistically reliable effect of *distractor distance*: the distance between the target and the distractor did not have an effect on the redundant use of color. As said, this is not what we predicted based on previous work on the production of dialogue reference (e.g., Beun & Cremers, 1998; Brown-Schmidt & Tanenhaus, 2008). In the General Discussion, we come back to this issue.

General discussion

In the two experiments presented in this chapter, we have investigated how bottom-up and top-down saliency cues - as defined by Itti and Koch (2000) - guide speakers in determining which objects in a scene belong to the set of relevant distractor objects. In doing this, we have studied how these cues affect speakers' production of definite object descriptions, and more specifically, to what extent they cause speakers to use a redundant color attribute. On average, 44% of the descriptions contained a redundant color attribute, which is more than the proportions reported by, among others, Belke and Meyer (2002) and Pechmann (1989). Perhaps, this is due to the photo-realistic nature of our stimuli.

Following the current REG algorithms, we have studied the production of "one-shot" reference, that is, reference with no preceding discourse involved. We expected to find that bottom-up and top-down saliency would both - albeit in different ways - cause speakers to redundantly use color. Our results have partially confirmed our expectations. Regarding bottom-up cues, we found that the presence of *clutter*, as well as the *color* and *type* of a distractor object to influence speakers' redundant color use. Regarding top-down saliency, we did not find reliable effects: the effect of the *specificity of the referential task* was only significant over items in Experiment 1.

More detailed discussions of the above findings have already been provided in the discussion sections of the two separate experiments. In this general discussion, we speculate on the role of *heuristics* in the production of reference, and discuss the *implications* of our findings for the current REG algorithms such as the IA. Finally, we sketch some lines for *future research*.

The role of heuristics in the production of object descriptions

Given that our findings confirm that some visual saliency cues have an effect on speakers' object descriptions, we would like to speculate on the search strategies that speakers use when processing a visual scene. As we have suggested in Chapter 4 of the current dissertation, it is unlikely that speakers always seek for the attributes with the highest distinguishing value (see Gatt, Krahmer, Van Gompel and Van Deemter (2013) for empirical evidence). In line with this, our results add that speakers do not always regard all objects as relevant distractors: distractors might be "ignored" for several reasons. For example, factors such as color differences and clutter can cause a distractor to be salient, but also the object's type and (to a lesser extent) the intention that speakers have in scanning the scene may play a role here. Following these results, it is plausible to conclude that speakers use more clever shortcuts when processing a visual scene.

Back in 1974, Tversky and Kahneman introduced the idea that people tend to use *heuristics* when taking all kinds of decisions. In their paper, Tversky and Kahneman defined heuristics as "beliefs concerning the likelihood of uncertain events (…) that reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations" (p. 1124). In recent years, heuristics have been claimed to also affect the way in which human speakers describe objects: given that their processing capacity is limited, speakers are prevented from making exact calculations about the shortest description that is possible in a given communicative situation (Van Deemter, Gatt, Van Gompel, & Krahmer, 2012a). Viethen and Dale (2009) and Dale and Viethen (2009) have explored the role of heuristics in the production of referring expressions in detail by means of corpus research, where they studied if there are characteristics of scenes or speakers that cause certain attributes to be used in descriptions. Their results indeed reveal a reasonable correlation between the scene characteristics and particular attributes, and they also show that the decision whether or not to incorporate a certain attribute varies from speaker to speaker. However, despite these individual referring strategies, some strategies seem to be common: while we showed in this dissertation (Chapter 4) that color variation in a scene causes speakers to redundantly mention color, Dale and Viethen (2009) found speakers to use color irrespective of the visual context.

Given this assumed role of heuristics in reference production, their incorporation in Referring Expression Generation algorithms is something that needs to be addressed in the future (van Deemter et al., 2012a). We argue that one way to do this would be to make such algorithms able to dynamically restrict the distractor set depending on the visual scene: as our results suggest, speakers do not always consider all objects in a scene as relevant distractors. So how can the latter be incorporated in a REG algorithm? While, as explained earlier, Krahmer and Theune (2002) used discourse structure to do this, our results provide empirical evidence for the suggestion that also *visual saliency cues* might be relevant to take into account (raised by, among others, Kelleher and Van Genabit (2004) and Kelleher and Kruijff (2006)). So what are the implications of our findings for REG algorithms such as the IA?

Implications for Referring Expression Generation algorithms

Before we can elaborate on the implications that our findings have for current REG algorithms, we first need to look further into these algorithms' output for the visual scenes used in this chapter. Given that the Incremental Algorithm (Dale & Reiter, 1995) is generally regarded as the most influential REG algorithm to date (as discussed by Van Deemter, Gatt, Van der Sluis, & Power, 2012b), our focus is on this algorithm in the remainder of this chapter. So what kind of descriptions would the IA generate for the target objects in our two experiments? And, in particular, to what extent would it include redundant color attributes there?

As we have explained in the beginning of this chapter, the IA uses a predetermined *preference order* (PO) in the generation of its object descriptions. This PO is fixed for every given domain, and is typically determined by empirical investigation. But how to determine a PO for the visual scenes used here if these scenes have not been used before in the field of Referring Expression Generation? Following the findings presented in Chapter 3 of this dissertation, showing that even a few references can be sufficient to come up with a 'good' PO for a previously unstudied domain, we can make an educated guess about what the PO for our scenes could look like.

Our scenes all had the same basic structure: there was one target, one distractor object with same type and color as the target but with a different size, and a second distractor whose type and color could be either the same or different as compared to the target object's type and color. In practice, this means that the PO for our scenes should contain at least three attributes: *type*, *color*, and *size*. Following Levelt (1989), who claims that speakers tend to always include a head noun in their references (meaning that *type* is practically always mentioned), we argue that *type* should be placed at the head of the PO in our scenes. The second attribute in line is then *color*, where we base ourselves on previous empirical research saying that the color attribute is known to be preferred in general, even in case it causes the description to be overspecified (e.g., Pechmann, 1989). The last attribute in the PO for our domains is then *size*, which is a relative property and therefore less frequent (see again Pechmann (1989) for empirical evidence).

To illustrate how the IA uses this PO (*type* > *color* > *size*) in the scenes presented to our participants, let us consider the upper left scene in Fig. 2. In this picture, the IA would first include *type*, since this attribute is at the head of the preference order and rules out at least one distractor object (the green mug). The next attribute in the PO (*color*) is not selected, since the only distractor that is left (a red plate) shares its color with the target. However, given that the distractor and the target do not have the same size, the last attribute in the PO (*size*) does the trick. In the end, the IA thus selects *type* and *size* and generates a description that can be realized as "the large plate".

In this way, the IA can generate distinguishing descriptions for all target objects in all scenes in Fig. 2 (representing Experiment 1) and Fig. 6 (representing Experiment 2). Generally speaking, it will *not* include color in most of the scenes depicted in these two figures. For example, in Fig. 2, the algorithm would produce the same description ("the large plate") in the non-clutter and the clutter conditions, not selecting color in any of the cases.

Obviously, this is not consistent with the effects of clutter that we have found in our first experiment (showing that speakers use color more often when clutter objects are present in the visual scene).

However, regarding the interaction between *type* and *color* in Experiment 2, the predictions of the IA seem to be – at least to some extent - psychologically realistic. As said earlier, this interaction implies that in scenes where all objects have the same type, speakers are more likely to mention a redundant color attribute when there is color variation in a scene as compared to when this is not the case. This seems to be consistent with the output of the IA: for example, it would select color in the fifth and sixth scene in Fig. 6 (terminating a set of attribute-value pairs that can be realized as "the large red plate"), whereas color would not be used in the scenes containing three red plates (the IA would then terminate "the large plate")[2]. Even though the IA makes predictions for these two scenes that are generally compatible with our results, it is interesting to note that the IA *always* selects color when a blue plate is present, and *never* when only red plates are present. In other words, like most other algorithms, the IA is *deterministic*: it always generates the same description for a given condition (Van Deemter et al., 2012a). For example, in Fig. 6, the IA would *always* use color in the scenes containing a blue plate, but *never* for the scenes containing only red plates. This contrasts with our participants' object descriptions: in the former case, color was mentioned in 65% of the descriptions, while this proportion was 42% in the latter case[3].

[2] Since all objects are plates in these pictures (and thus have the same type), the IA would initially not select type. Instead, type is added afterwards, following the intuition that objects descriptions should always contain a head noun (Levelt, 1989).

[3] As Van Deemter et al. (2012b) stress, the IA is dependent on the exact PO selected, but no other PO seems to match our findings well. Consider the PO with color first and type last (which Mitchell, Van Deemter and Reiter (2013) found to perform well): for our scenes this would imply that the IA would *always* select color in experiment 1 and in experiment 2 always when the designated distractor had a different color from the target, irrespective of its type. Clearly, this does not match our data either.

To wrap up, the above shows that there are many situations where human speakers often mention color, but where the Incremental Algorithm would not do this, or where it would be hindered by its deterministic nature. So the question is: how to solve these issues? Although we do not aim to present an extension of the IA in the current chapter, we would like to elaborate shortly on possible solutions. One straightforward option would be to dynamically adapt the preference order to the *visual saliency cues* that can be derived from a given visual scene. For example, for the specific case of *visual clutter* (which is one the factors that delivered us with a convincing effect), this could be done as follows: when clutter is present, color could be placed at the head of the preference order (causing it to be selected if there is any color variation in the scene, which is more likely in the case of clutter), whereas the preference order can remain as we assume it is now (with type before color) for scenes without clutter. In this way, the IA could also (at least partly) account for other visual cues that we have found to affect redundant color use (such as the *color* and *type* of a potential distractor): it can use these cues to determine which objects in a given scene should be included in the distractor set, and make color more or less preferred depending on the amount of color variation that is present between the target and the distractors that are left.

Although the above solution is a step in the right direction, it would not make the IA less deterministic: the inclusion of color would remain a matter of all or nothing. As Van Deemter et al. (2012a) argue, one way to incorporate some non-determinism in the IA would be to enrich the algorithm with a probabilistic module that allows it to adapt the preference order in some cases, and to leave this order unattended in the rest of the cases. For example, for the cluttered scenes in our first experiment, this implies that the IA would have to check color before type in 43% of the cases and type before color in 57% of the cases (both across speakers and within a single speaker). Along similar lines, also other (recently proposed) algorithms have aimed to approximate the non-deterministic nature of reference production, most notably the *Visible Objects algorithm* (Mitchell, Van Deemter, & Reiter, 2013), and also the *Probabilistic Referential*

Overspecification algorithm (Van Gompel, Gatt, Krahmer & Van Deemter, 2012). The latter algorithm, for example, assumes that speakers first aim for a distinguishing description, and might subsequently overspecify with a certain probability when they add another attribute (depending on their preference).

Directions for future research

Since our manipulations of *distance* and *task specificity* have not delivered us with reliable effects, we believe that these are worth studying in future research. In this last section, we speculate on how this can be done.

Regarding *distractor distance*, we expected to find that objects that are visually close to the target object would be more likely to be considered as relevant distractors than objects that are distant. However, contradictory to our intuitions and to previous research on the effect of distance in dialogue reference (e.g., Beun & Cremers, 1998), we did not find this to be the case. One explanation for this might be that the effect of distance was subsumed by the (strong) effect of *distractor type*: since the target and the distractor had different types in the majority of trials, it may have been the case that speakers saw no reason to consider the distractor, even if it was close to the target. Besides that, also the fact that we showed 2D representations of 3D scenes to our participants may have reduced the effect of distance. Therefore, based on earlier research showing that people have a stronger feeling of depth in 3D scenes rather than 2D scenes (e.g., Seuntiens et al., 2005), we are currently planning to study the impact of distractor distance by using realistic 3D scenes.

Secondly, regarding the *specificity of the referential task*, we expected to find that instructions that contained the target's type would cause speakers to overspecify less frequently as compared to more general instructions, since in the first case only those objects that shared their type with the target would remain to be ruled out. Again, the differences between the conditions were not significant, in neither of the experiments. One explanation for this could be related to the fact that we took the redundant use of color as our dependent variable. Our argumentation was that the specific instruction (which included the target's type) would do some work for the speaker, since it had already selected the 'type' part of the description. It might be that the manipulation of the referential task was not sufficiently strong, or perhaps the expected effect simply does not exist. In any case, this should be addressed in future research, perhaps with a somewhat different task (e.g., akin to the one employed by Arts et al., 2011).

Conclusion

The two experiments reported in this chapter show that bottom-up, perceptual factors such as *distractor color*, *distractor type*, and *visual clutter* affect the redundant use of color in object descriptions, implying that these factors guide speakers in determining the set of relevant distractors when they are presented with a photo-realistic scene. However, similar effects of *distractor distance* and of top-down, conceptual saliency cues (i.e., *specificity of the referential task*) were not convincingly borne out by the data. Our results regarding bottom-up saliency cues are problematic for existing REG algorithms (such as the Incremental Algorithm) that aim to generate humanlike target descriptions.

Acknowledgments

We thank Anouk van Helvoort for assistance in creating the stimuli, Elsa Jonkers and Kristel Bartels for running the experiments and transcribing and annotating the data, and Martijn Goudbeek, Jorrig Vogels, Marieke Hoetjes, Albert Gatt, and Jette Viethen for their comments on an earlier version of the chapter.

References

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification in written instruction. *Linguistics*, 49 (3), 555-574.

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.

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

Bravo, M., & Farid, H. (2008). A scale invariant measure of clutter. *Journal of Vision*, 8 (1), 1-9.

Brown-Schmidt, S., & Tanenhaus, M. (2008). Real-time investigation of referential domains in unscripted conversation: a targeted language game approach. *Cognitive Science*, 32 (4), 643-684.

Brown-Schmidt S., & Konopka, A. (2011). Experimental approaches to referential domains and the online-processing of referring expressions in unscripted conversation. *Information*, 2 (2), 302-326.

Coco, M., & Keller, F. (2009). The impact of visual information on reference assignment in sentence production. In *Proceedings of the 31st annual conference of the Cognitive Science Society (CogSci)*, 274-279. Amsterdam, The Netherlands.

Coco, M., & Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, 36 (7), 1207-1223.

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

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

Dale, R., and Viethen, J. (2009). Referring Expression Generation through attribute-based heuristics. *Proceedings of the 12th European workshop on Natural Language Generation (ENLG)*, 58-65.

Einhäuser, W., Spain, M., and Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, 8 (14), 1-26.

Foulsham, T. & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of vision*, 8 (2), 1-17.

Gatt, A., Krahmer, E., Van Gompel, R., and Van Deemter, K. (2013). Production of referring expressions: Preference trumps discrimination. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society (CogSci)*.

Griffin, Z., and Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11, 274-279.

Grosz, B., & Sidner, C. (1986). Attention, intentions and the structure of discourse. *Computational Linguistics*, 12, 175-204.

Hegarty, M., Canham, M., & Fabrikant, S. (2010). Thinking about the weather: how display salience and knowledge affect performance in a graphic inference task. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 36 (1), 37-53.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40, 1489-1506.

Kelleher, J. & Kruijff, G.J. (2006). Incremental generation of spatial referring expressions in situated dialogue. In *Proceedings of COLING/ACL '06*. Sydney, Australia.

Kelleher, J. & Van Genabith, J. (2004). Visual salience and reference resolution in simulated 3-D environments. *Artificial Intelligence Review*, 21 (3), 1-14.

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In: K. van Deemter & R. Kibble (Eds.). *Information sharing: Givenness and newness in language processing* (pp. 223-264). CSLI publications, Stanford.

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

Mitchell, M., Van Deemter, K., & Reiter, E. (2013). Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).* Atlanta, USA.

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

Reiter, E., and Dale, R. (2000). *Building Natural Language Generation systems*. Cambridge, Cambridge University Press.

Seuntiens, P., Heynderickx, I., IJsselstijn, W., van den Avoort, P., & Berentsen, J., Dalm, I., Lambooij, M., & Oosting, W. (2005). Viewing experience and naturalness of 3D images. In *Proceedings of SPIE Optics East, three dimensional TV, video, and display.* Boston, USA.

Thórisson, K. (1994). Simulated perceptual grouping: an application to human-computer interaction. In *Proceedings of the 16th annual conference of the Cognitive Science Society (CogSci)*, 876-881. Atlanta, Georgia.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12, 97-136.

Tversky. B., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, New Series, Vol. 185, 1124-1131.

Van Deemter, K., Gatt, A., Van Gompel, R., & Krahmer, E. (2012a). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4 (2), 166-183.

Van Deemter, K., Gatt, A., Van der Sluis, I., & Power, R. (2012b). Generation of referring expressions: assessing the Incremental Algorithm. *Cognitive Science*, 36, 799-836.

Van der Sluis, I. (2005). Multimodal reference. *PhD thesis*, Tilburg University, The

Netherlands.

Van Gompel, R., Gatt, A., Krahmer, E., Van Deemter, K. (2012). PRO: a computational model of referential overspecification. Submitted to *Architectures and Mechanisms for Language Processing (AMLaP)*, to be held in Marseille, France.

Viethen, J., & Dale, R. (2009). Referring Expression Generation: what can we learn from human data? *Proceedings of the 2009 workshop on the production of referring expressions: bridging the gap between computational and empirical approaches to reference.* Amsterdam, the Netherlands.

Viethen, J., Dale, R., & Guhe, M. (2011). The impact of visual context on the content of referring expressions. *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*. Nancy, France.

Viethen, J., Goudbeek, M., & Krahmer, E. (2012). The impact of colour difference and colour codability on reference production. In *Proceedings of the 34th annual meeting of the Cognitive Science society (CogSci)*. Sapporo, Japan, 1066-1071.

Westerbeek, H., & Maes, A. (2013). Route-external and route-internal landmarks in route descriptions: effects of route length and map design. *Applied Cognitive Psychology*, 27 (3), 297-305.

Need I say more?

6

Developmental changes in children’s processing of redundant information

Abstract

This chapter studies developmental changes in children’s processing of redundant information in definite object descriptions. In two experiments, six- to seven- and nine- to ten-year-old children were presented with pictures of sweets. In the first experiment (pairwise comparison), two identical sweets were shown, and one of these was described with a description containing a redundant modifier. After this, the children had to indicate the sweet they preferred most in a forced-choice task. In the second experiment (absolute rating), only one sweet was shown, which was in half of the cases described with a redundant modifier and in the other half of the cases simply as “the sweet”. This time, the children were asked to what extent they preferred the sweets on a 5-point rating scale. In both experiments, the results showed that young children had a preference for the sweets described with redundant information, while this was not the case for the older ones. These results suggest that between six or seven and nine or ten years of age, children learn how to avoid drawing false conversational implicatures triggered by the inclusion of redundant information in object descriptions.

This chapter is based on:

- Koolen, R., Krahmer, E., & Swerts, M. (submitted). Developmental changes in children’s processing of redundant information in definite object descriptions.

Introduction

In referential communication, speakers often have communicative intentions that go beyond the identification of a target object. For example, imagine a mother producing the following utterance to her young son: *"Be careful with the big wine gum!"*, in a setting where there is only one wine gum present. Obviously, by producing this utterance, the mother has the intention to communicate to her son that he has to be careful with the wine gum in order to prevent him from choking. In doing this, she uses two cues: first, she explicitly uses the imperative verb phrase 'be careful', and second, interestingly, there is an implicit cue, being the adjective 'big'. In the literal sense, this adjective is not necessary for unique identification of the wine gum, which makes it *redundant*: there is no other wine gum at play, and the child will probably be able to estimate the size of the wine gum himself. Thus, at first sight, a description such as "the big wine gum" is *overspecified* in this example context.

Of course, in this particular situation, emphasizing the big size of the wine gum is driven by the mother's communicative intentions: it will (hopefully) cause the child to reason that his mother includes this information to warn him against the hazards of choking on large objects. Following Grice (1975), this kind of implicit reasoning is triggered by a *conversational implicature*, resulting from a violation of the Gricean Maxim of Quantity. This Maxim states that a contribution to a dialogue should be as informative as required, but not more informative. In our example, the mother's use of the adjective 'big' appears to violate this maxim, thereby triggering the implicit reasoning sketched above.

However, it is also known from many previous studies on adult reference that speakers - for various reasons - routinely produce object descriptions that contain one or more redundant modifiers (e.g., Pechmann, 1989; Engelhardt, Bailey, & Ferreira, 2006), referring, say, to a single wine gum as "the big wine gum", without any further communicative intentions. This happens - among others - when speakers add perceptually salient attributes (such as color) to their descriptions, which is something they often do (Belke & Meyer, 2002). To make this concrete, imagine a situation with two red wine gums, one of which being described as “the *red* wine gum”. Arguably, the redundant modifier ‘red’ could trigger a conversational implicature here, since, so an addressee might reason, why else would the speaker mention color information that is not necessary to identify the wine gum? Crucially, such an implicature would be a *false* one: the addressee may - for example - conclude that the speaker wanted to assign special status to the wine gum’s color, while this was not something that the speaker intended to communicate.

Although - as explained above - redundant properties are omnipresent in everyday object descriptions, false conversational implicatures are rather unlikely to occur among adult speakers: these are conversationally experienced language partners who are generally well able to monitor each other’s implicit communicative intentions (e.g., Brennan & Clark, 1996; Levinson, 2000). Among children, however, this might be different: young children in particular may have a higher chance of deriving (false) implicatures, and must learn to understand the implications of redundant information that is provided to them in conversation (Siegal & Surian, 2004). Related to this, some earlier work has been done on children’s development in the production (e.g., Deutsch & Pechmann, 1982; Matthews et al., 2012) and processing (e.g., Sonnenschein, 1982; Davies & Katsos, 2010) of descriptions containing one or more redundant modifiers, and on children’s ability to derive scalar quantity implicatures (e.g., Geurts, Katsos, Cummins, Moons, & Noordman, 2010; Katsos & Davies, 2011; Musolino, 2004).

The papers mentioned above study redundancy in the light of reference resolution, or have a focus on how children derive scalar implicatures. However, as far as we are aware, (developmental) research is lacking on children’s comprehension of redundant information in situations where the redundant modifier does not serve any specific goal (such as is the case with ‘red’ in the example given above)? Are children then, like adults, able to derive that the modifier is purely redundant? Or do such modifiers cause them to derive false conversational implicatures? These are the questions that we address in this chapter. We do this by presenting the results of two comprehension experiments, using forced-choice tasks (Exp. 1) and graded rating scales (Exp. 2) to investigate the effect of redundant modifiers (such as 'red') on children's preferences for objects.

Firstly, we provide an overview of the language acquisition literature showing that both adults and children (roughly from the age of seven) often overspecify their target descriptions. We then discuss what this implies for (child) listeners: to what extent are these able to judge the relevance of the redundant information they are provided with? Thereafter, we provide a detailed discussion of our research questions and hypotheses. After giving an outline of the method used, we report and discuss the results of our experiments, focusing on developmental differences between six- to seven-year-old children on the one hand, and nine to ten-year-old children on the other hand.

Overspecification in children's production of target descriptions

Previous research in language acquisition has shown that children start referring to physical objects in the world when they are around twelve months old: they then do this by means of pointing gestures (Tomasello et al., 2007), and with their first words (Fenson et al., 1994).

When children start to produce more complex referring expressions, between two and four years of age, children generally *underspecify*, meaning that they then tend to utter ambiguous target descriptions that do not contain enough information for unique identification of the target referent (Matthews et al., 2007). Matthews and colleagues conducted an experiment in which two- to four-year-old children were tested for their ability to request stickers from a dense array. The results showed that two- and three-year-old children mainly pointed at or named the target sticker they wanted (where naming was defined as referring to the character involved without using any disambiguating predicate), whereas four-year-old children do sometimes utter uniquely identifying descriptions. Deutsch and Pechmann (1982) emphasize the crucial role of feedback for children around this age: they observe that three-year-olds sometimes produce contrastive descriptions, but only when repeatedly asked for clarification. A more recent experiment by Matthews et al. (2012) confirms these results, and adds that larger visual arrays also cause four-year-old children to produce more complex descriptions.

As suggested in early experimental work by Deutsch and Pechmann (1982), and by Sonnenschein and Whitehurst (1982), the general tendency in children's spontaneous production of target descriptions is that they continue to produce underspecified and ambiguous descriptions until they are seven or eight years old. Further evidence for this pattern comes from Davies and Katsos (2010), who found five-year-old children to underspecify in 55% of their descriptions, which is line with the findings reported by Sonnenschein (1982) showing that four- to six-year-old children are more likely to underspecify than their seven- to nine-year-old counterparts. However, it still seems that five-year-olds have developed somehow: around this age, they are aware of the information they share with their listener (Nadig & Sedivy, 2002).

Also the inclusion of redundant information starts around the age of five, although children at this age only sometimes overspecify their descriptions (Davies & Katsos, 2010; Ford & Olson, 1975). Children start to *overspecify* their target descriptions on a regular basis around the age of seven (Ford and Olson, 1975). Since Ford and Olson found that seven-year-old children often emphasized the redundant attributes that they mentioned, they concluded that these children actually made comparisons between the target and its distractors. This conclusion is advocated by Whitehurst (1976) as well, showing that fully informative descriptions become more common with age, and that overspecified descriptions occur from the age of seven. Many studies have shown that speakers keep routinely including redundant modifiers in their descriptions as adults (e.g., Arts, 2004; Pechmann, 1989; Engelhardt et al., 2006; and chapter 2, 4 and 5 of this dissertation).

Regarding object descriptions that are uttered primarily to distinguish a target from one or more distractors, we have now seen that children initially tend to underspecify such descriptions, that they start to include redundant attributes roughly from the age of seven, and that they keep doing this as adult speakers. For the current study, this has at least one important implication: *addressees* will have to judge the relevance of the redundant information that they are presented with. Are these always able to do this in a successful way?

Developmental changes in the derivation of quantity implicatures

The Cooperative Principle

For the case of adult listeners, it can be argued that these have little difficulty in interpreting redundant information in speakers' object descriptions. Following Grice's Cooperative Principle (1975), this is because speakers and listeners tend to co-operate when they are in a conversation (this is echoed in, among others, Brennan and Clark, 1996). With this Cooperative Principle, Grice (1975) argues that the expectations of language partners in conversation are characterized by four maxims, which hold that speakers should not say less or more than is required (Maxim of Quantity), that they should tell the truth and avoid unfounded statements (Maxim of Quality), that their contribution should be relevant (Maxim of Relation), and they should avoid obscurity and ambiguity (Maxim of Manner).

Although the Gricean maxims intuitively seem to focus on the role of the speaker in a conversation, Grice (1989) himself emphasizes that they also apply to listeners (p. 31). This means that when speakers violate one of the maxims without any specific purpose, their addressees might fail to understand their communicative intention. This is what happens in the example with the two red wine gums: the speaker mentions a color attribute without any further communicative intentions, but the addressee draws a *false conversational implicature*, meaning that he expects there to be a reason why the speaker mentions color (which is actually not the case).

Given that adult speakers are generally cooperative, it is plausible to assume that speakers aim to prevent their listeners from deriving false implicatures, and that they normally make sure that listeners are able to assess the relevance of all information included in a description. Empirical findings from Engelhardt et al. (2006) provide support for this assumption, showing that adult listeners whom are asked to judge the quality of instructions do not rate overspecified target descriptions as any worse than minimally specified ones. However, arguably, this might be different for children, because these may not understand why, when and how conversational maxims can be violated, and therefore may fail to interpret speakers' "meanings" (Siegal & Surian, 2004). In other words, they might have a higher chance of deriving false implicatures in the sense of Grice (1975).

In the next sections, we discuss acquisition literature on children's development in the derivation of two kinds of conversational implicatures related to Grice's Maxim of Quantity. First, we report on existing research on how children in different age groups process redundant attributes during *reference resolution*, and we do the same for the comprehension of *scalar implicatures*.

Quantity implicatures related to reference resolution

It is traditionally assumed that until the age of seven or eight, children are not good at evaluating and editing the communicative content of the expressions that they are presented with. Some early studies (e.g., Ackermann, 1981; Ackermann et al., 1990; Bonitatibus et al., 1988) argue that young children generally find it hard to distinguish between ambiguous and informative descriptions when selecting a target, in the sense that they have difficulty indicating whether their selected object is the "right one", the one the speaker "meant", or the one the speaker "could have meant". Ackerman et al. explain this by claiming that these young children often fail to derive the speaker's communicative intentions from an expression, and relate this to the common ground that is shared between the speaker and the listener: children under eight years old find it difficult to infer relevant information from shared knowledge.

Some studies directly measure the impact of redundant attributes on child listeners. These studies investigate overspecification in the light of identification: to what extent do redundant attributes help or inhibit a child listener to select a target referent? In her experiments, Sonnenschein (1982) presented five- and nine-year-old children with overspecified and minimally specified target descriptions, and studied how redundant information assisted these children in identifying a target group of objects. The results indicated that when the visual array in which the target objects occurred was complex (due to the length of the exposure or the size of the stimuli), the presence of redundant information improved the older children's performance. The redundant attributes did not help for the five-year-old children, which Sonnenschein explains by arguing that the memory capacity of these younger children is not sufficient to process redundant information in referring expressions.

Davies and Katsos (2010) studied the comprehension of redundant information in arrays that consisted of four everyday objects, and were simpler than the ones used by Sonnenschein (1982). In their first comprehension study, a binary judgement task was used to do this, where five-year-old children and adults heard minimally specified and overspecified descriptions and were asked to indicate whether these were natural or not. This was done by asking the children whether they found that certain descriptions were produced in a good, sensible way, or in a bad, silly way. Although the results revealed that the children did not reject overspecified target descriptions more than minimal ones, Davies and Katsos argued that this result might be due to the nature of the children's task. Therefore, in their second experiment, Davies and Katsos (2010) used magnitude estimation ratings, and these showed a different picture: this time, the children rated the overspecified descriptions lower than the minimal ones, implying that children are already sensitive to violations of the Gricean maxims from the age of five.

Another paper that used graded ratings to study the impact of redundant modifiers on children of different ages was conducted by Krahmer, Noordewier, Goudbeek and Koolen (2013). In this study, both six- and nine-year-old children were presented with picture grids containing two toys of different types. One of these toys was the target, which was referred to in a description that could either be minimal or overspecified in the context of the second object. For example, in a scene where the target is a football and the distractor is a teddy bear, "the football" would be minimal, while "the large football" would be overspecified. In all trials, type was distinguishing, while size was always used as a redundant modifier in the overspecification condition. The children were asked to estimate the size of every target on a magnitude estimation scale of 100 millimeters (based on the descriptions they had heard). The results showed an effect of age on children's processing of the redundant size modifiers: the younger children made larger size estimates than the older ones in the overspecification condition, but not in the condition where the toys were minimally referred to. According to Krahmer and colleagues (2013), these results suggest that the nine-year-old children were less sensitive to redundant size modifiers than their six-year-old counterparts.

Children's ability to derive scalar implicatures

Besides Quantity Implicatures in the context of reference resolution, children must also learn to derive *scalar implicatures*. These are relevant for the current study, since – as we discuss below – they raise some interesting suggestions regarding the *task* that is used to investigate children's capabilities in deriving such implicatures. Scalar implicatures follow from sentences in which scalar modifiers are used, such as "Some toys are green" (which has the semantic meaning that *at least two* toys are green, and elicits the implicature that *not all* toys are green). These scalar modifiers can come in two types: *superlative* quantifiers (such as "at least" or "at most") and *comparative* ones (such as "more than" or "fewer than"). The traditional view regarding the derivation of scalar implicatures is that children are not able to comprehend scalar quantifiers at adult-like levels until they are at least seven years old (Noveck & Reboul, 2009). However, recently, this view has been nuanced in the sense that children's ability to successfully derive scalar implicatures seems to depend on the *type* of quantifying term that they have to comprehend (Geurts et al., 2010), and, interestingly, on the *task* they are faced with (Katsos & Bishop, 2011).

To start with the latter, Katsos and Bishop (2011) show that five-year-old children are to a large extent able to derive scalar implicatures related to 'some', and that their performance largely depends the nature of the *experimental task*. In their experiments, Katsos and Bishop (2011) presented five-year-old children with animated Powerpoint displays in which a protagonist performed some course of action with a number of objects (e.g., a mouse picks up five carrots, but leaves five pumpkins unattended). After every display, the children heard a statement that they had to judge (exp. 1) or reward (exp. 2). In half of the trials, the statements contained the scalar modifier 'some', and were pragmatically underinformative in the context of the performed action (e.g., "the mouse moves some of the carrots"). The other half of the stimuli consisted of statements that were fully informative. In the first experiment, Katsos and Bishop used a binary judgment task to have the participants indicate whether the statements were right or wrong. The results showed that children rejected underinformative scalar expressions in only 26% of the cases, suggesting that the children were not competent in deriving scalar implicatures related to 'some' at this age. However, when Katsos and Bishop (2011) used three-point Likert scales to reward the statements (which they did in the second experiment, reusing the stimuli of the first experiment), the children's performance improved: this time, the underinformative statements were rated lower than fully informative ones. Katsos and Bishop (2011) explain these results by suggesting that the children's poor performance in their first experiment was due to the binary task, and that children are already sensitive to violations of the Gricean maxims at the age of five.

There is also some prior research suggesting that five-year-old children are indeed able to derive scalar implicatures related to superlative quantifiers, but that they keep having difficulties with (some) comparative quantifiers until they are at least eleven years old. This is in line with Geurts and Nouwen (2007), who claim that superlative quantifiers are inherently more complex than comparative ones. Based on this claim, Geurts et al. (2010) argue that the assumed difference in complexity between the two kinds of quantifiers should also be reflected in language acquisition. Musolino (2004) provides empirical evidence for the case of five-year-old children, whom he shows to have difficulties in processing superlative quantifiers, but not in the comprehension of comparative quantifiers. Following these findings, Geurts et al. (2010) obtained more fine-grained results with eleven-year-old children. In an acquisition experiment, they asked participants to match arrays of boxes and small toys with quantifying sentences. The sentences contained superlative quantifiers that could either be upward ("At least three boxes have a toy") or downward entailing ("At most three boxes have a toy"), or comparative quantifiers, again with the distinction between upward ("More than three boxes have a toy") and downward entailment ("Fewer than three boxes have a toy"). The results reveal that eleven-year-old children are generally well able to process the two comparative quantifiers (downward and upward) and the superlative quantifier 'at least', while they still have serious problems with 'at most'. Generally speaking, these results presented by Geurts et al. (2010) imply that even at the age of eleven, children are not fully competent in deriving scalar implicatures.

The current study

In the previous sections, we have described how children of different ages produce overspecified object descriptions, and how they develop in their ability to successfully derive implicatures related to the Maxim of Quantity (Grice, 1975). With regard to the latter, the existing literature has so far mainly focused on implicatures concerning the processing of redundant information during *reference resolution*, and on the extent to which children are competent in deriving *scalar implicatures*.

The general picture that emerges from the literature discussed above is ambiguous, and raises two interesting observations. Firstly, as related to children's development, some prior work has shown that children are not yet fully pragmatically competent at the age of five: they often produce underspecified descriptions (e.g., Davies & Katsos, 2010), they are not aware that redundant modifiers may be beneficial during the target identification process (Davies & Katsos, 2010; Sonnenschein, 1982), and they find it difficult to process superlative quantifiers (Musolino, 2004). Interestingly, the above studies also show that nine-year-old children do not have such problems. However, on the other hand, there are also studies showing that five-year-olds are actually capable of deriving quantity implicatures that include 'some' (Katsos & Bishop, 2011), and, contrastively, that eleven-year-olds have serious problems with downward superlative quantifiers (Geurts et al., 2010).

Secondly, previous work has emphasized the impact of the experimental task that is used to investigate children's capabilities in the derivation of quantity implicatures. In particular, Davies and Katsos (2010) have shown that five-year-old children do not reject overspecified descriptions more often than minimal ones in a binary judgment task, but that they do rate descriptions that contain redundant information lower when magnitude estimation scales are used. Related to this, Katsos and Bishop (2011) have shown that in sentences where scalar modifiers are used, five-year-old children do not reject underinformative statements in a binary task, but rate such underinformative statements lower than fully informative ones when graded rating scales are used.

The current study aims to gain further insight in the above issues by addressing the question how children develop in communicative situations where a speaker includes a redundant attribute in an object description, but *not* with the purpose to distinguish one object from one or more distractors. This is an important question, since children (roughly from the age of seven) and adult speakers often include redundant attributes in their object descriptions (e.g., Pechmann, 1989; Engelhardt et al., 2006), implying that young children must learn to judge whether this information is relevant or not in a given communicative context. In particular, we focus on situations in which a child is presented with overspecified descriptions of sweets, where we measure the extent to which a child is guided in its preferences for sweets by the redundant information. We have conducted two experiments to investigate this: one using a binary, forced-choice task, and one using 5-point rating scales. We were particularly interested in the impact of *objective* redundant information on children's preferences for sweets, expecting to find differences between children of six to seven and nine to ten years old.

Experiment 1: The effect of redundancy in a forced-choice task

In our first experiment, we presented children in two age groups with pictures of two sweets. In all critical trials, these two sweets were identical, and one of them was described with an objective redundant modifier (including the sweet's color or shape). In a forced-choice task, the children were asked to indicate which of the two sweets they preferred (based on the descriptions they had heard).

Method

Participants. Participants were 49 children in two age groups. The population in the younger age group consisted of 22 children (10 males, 12 females) with a mean age of 6;7 years (ranging between 6;0 and 7;3). The population in the older age group consisted of 27 children (12 males, 15 females) with a mean age of 9;8 years (ranging between 9;2 and 10;2). In the Dutch primary school system, the younger children were all in 'Group 3', while older ones were in 'Group 6'. All children were recruited at the same primary school, and came from predominantly Dutch-speaking families. Before the experiment, all parents had given their children permission to participate via a signed consent form.

Materials. The materials consisted of pictures of two sweets that were placed next to each other. In the critical trials, the two sweets were of the same kind and colour (see the left picture in Fig. 1), and the participating children therefore had no reason to have an a priori preference for one of the two.

Pre-recorded descriptions of the two sweets were played while the pictures were shown to the children. These spoken descriptions were presented as questions that always had the following structure: *"Would you like this (...) sweet or this (...) sweet?"* First, a picture of the two sweets was shown (see the left picture of Fig. 1), during which the first part of the question was played: *"Would you like..."* After that, the description of the left sweet was played. During this description, the corresponding sweet was highlighted with a red arrow (see the middle picture of Fig. 1). Once the left sweet had been described, a description of the right sweet followed, again highlighted with an arrow (see the right picture of Fig. 1). After having heard the two descriptions, the child had to indicate which sweet she preferred. The pre-recorded descriptions were produced by a female voice with a neutral intonation, avoiding prominent contrastive accents to highlight the redundant information. Note that the combination of the arrow with the description "*this sweet*" was always sufficient to unambiguously identify a sweet, so that any inclusion of a modifier would result in an overspecified description.

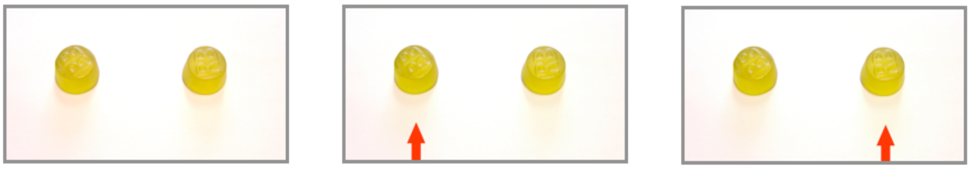


Fig. 1: Example of a criticial trial. The trial started and ended with the left picture. In between, the middle and right pictures were presented (highlighting the left and right sweet).

One sweet per critical trial was described with a redundant modifier; the other sweet was simply described as "this sweet". There were six critical trials in which the redundant modifier provided *objective* information about one of the two sweets, which was always information about perceptual characteristics of the sweets. In three critical trials, this was information about the colour of the sweets, while shape information was provided in another three critical trials. For example, in the critical trial depicted in Fig. 1, this led to the following question: *"Would you like this sweet or this green sweet?"* Based on the descriptions and the objective information, the children were asked to indicate which sweet they preferred. Whether the left or right sweet was redundantly described was counterbalanced over trials.

We performed a pre-test in order to make sure that the children were aware that the two sweets in the pictures used for the critical trials looked in fact identical. To do this, we presented six pictures of identical looking sweets (the same as the ones used in our experiment) and six pictures of different sweets (randomly selected from the fillers) to twenty children (ten in each age group). None of these children had participated in the main experiment. For each picture, we asked the children the following question: *"Do these two sweets taste the same?"*. For the pictures of identical sweets, the results showed that the children expected the two sweets to taste identically in 100% of the cases, and this percentage decreased to 2% for the pictures of different sweets. We can conclude from this pre-test that all information that was provided in speech about the sweets in the critical trials could be considered redundant. Hence, if a child preferred the redundantly described sweet, the redundant modifier arguably guided the choice for this sweet.

Sixty-six trials were included as fillers, the majority of which (fourty-eight) consisted of pictures of two different kinds of sweets, both described with various kinds of information (a combination of objective and affective information). A small minority of the fillers (eighteen) consisted of pictures of two similar sweets. In the majority of these trials, one of the two sweets was described with positive or negative information (such as 'delicious' or 'disgusting'). The remaining six fillers served as a baseline, allowing us to check whether the children were biased in favour of either the left or the right sweet. In these cases, neither of the two sweets was described with a modifier, and the following question was asked: "Would you like this sweet or this sweet?". The results showed that the children did not have a bias for left or right: the younger children chose for the left sweet in 50% of the baseline trials, while the older ones did this in 51.8% of the cases. These proportions did not differ significantly from chance level (i.e., 50%), as revealed by one-sample *t* tests that were applied on the participants' mean scores on the six baseline trials (younger children: *t* (21) = .00, *ns*; older children: *t* (26) = .35, *ns*). Based on these results, we ignore order of presentation in the analyses of the data.

Procedure. The procedure was identical for the children in the two age groups. The experiments had a running time of approximately fifteen minutes, and were individually performed in a quiet room inside the school building. The experiment was conducted in Dutch. We made one block of seventy-two trials in a fixed random order (this was presented to one half of the participants), and a second block containing the same trials in reverse order (presented to the other half of the participants). Our analyses did not show a significant difference between these groups, meaning that there was no habituation effect over the course of the experiment.

After a child had entered the experimental room, he or she was asked to take place in front of a computer screen. The experimenter was seated next to the child for the whole experiment. The instructions (which were provided orally by the experimenter) explained that the children first had to listen carefully to the pre-recorded descriptions that were given about the two sweets, and that they then had to indicate which of the two sweets they preferred. They could either do this by pointing at the sweet of choice on the screen, or by telling the experimenter which choice they had made. It was emphasized that the children had to base their choices on the descriptions that were provided about the sweets. Each time a child had completed a trial, the experimenter marked his or her choice on an answering form. In each trial, the children had four seconds to indicate the sweet they preferred, but they were given more time in case this was necessary. The experimental procedure started with three practice trials to acquaint the children with the procedure. After the completion of the experiment, the children confirmed that they had understood the experimental task, and all of them had enjoyed participating. When explicitly asked, none of the children indicated to have been aware of the actual goal of the study.

Design and statistical analysis. The experiment had a between-participants design with *age* (levels: six to seven years and nine to ten years) as the independent variable, and the proportion of choices for the redundantly described sweet as the dependent variable. We ran a Repeated Measures ANOVA[1] to test for significant differences between the age groups.

[1] In order to control for departures of normality, we also ran the ANOVA on transformated proportions (using standard arcsin transformations). Because the results for the transformed proportions showed exactly the same picture as those for the untransformed data, we stick to the untransformed data here.

We applied one-sample *t* tests on the participants' mean scores to see whether these were significantly different from chance level (i.e., 50%) in any of the two age groups.

Results and discussion

Fig. 2 depicts the proportion of choices for the sweets described with redundant information as a function of the two different age groups.

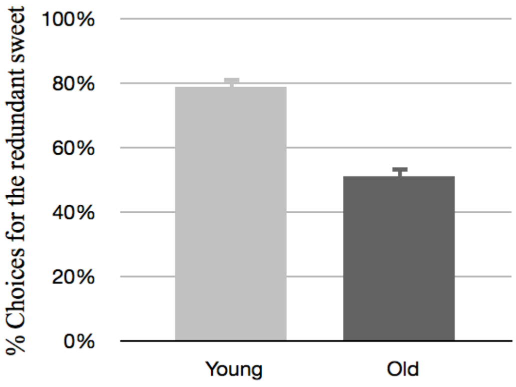


Fig. 2: The proportion of choices for the redundantly described sweets as a function of age group.

Fig. 2 shows a clear difference between the younger and older children in terms of the proportion of choices for the sweets that were described with a redundant modifier. As reflected in an effect of *age group* ($F_{(1,47)} = 35.218$, $p < .001$, $\eta^2 = .43$), the young children were far more sensitive to these redundant modifiers. More specifically, we found that the young children were more likely to choose the sweet that was described with redundant information ($M = .79$, $SD = .03$)[2], and this proportion was significantly higher

[2] Following Sedivy (2003), who found that redundant color terms do not necessarily trigger contrastive inference, we also checked whether the two types of modifiers used in this study (color and shape) had a different impact. The results were contradictory to Sedivy's claim: we found that the color modifiers (M = .75) had a bigger impact as compared to the shape modifiers (M = .55).

than 50% chance level: *t* (21) = 9.80, *p* < .001. For the older children, however, the results showed a different pattern, in the sense that they only chose the redundantly described sweet in around half of the cases (*M* = .51, *SD* = .03). This proportion was not significantly different from 50% chance level: *t* (26) = .36, *ns*.

Discussion. The above results indicate that while six- to seven-year-old children are guided in their choices by redundant information, nine to ten-year-old children are not. For the latter children, we found that they were not sensitive to redundant modifiers related to the colour or shape of the sweets: in the cases where they had to choose between for example "the green sweet or the sweet", they scored at chance level. The latter was not the case for the young children, who chose for the redundantly described sweet in almost 80% of the cases.

As we have seen in the introduction section, earlier work has emphasized the role of the experimental task if one wants to study children's ability to successfully derive quantity implicatures. Particularly, in some situations, five-year-olds have been shown to perform better when magnitude estimation scales rather than binary judgement tasks are used (e.g., Davies & Katsos, 2010; Katsos & Bishop, 2011). Therefore, in the next experiment, we replicate our first experiment with pictures of single sweets. Instead of binary judgments, we use 5-point rating scales to measure to what extent children are affected by redundant information in their preferences.

Experiment 2: The effect of redundancy using graded rating scales

Method

Participants. Participants were 60 children in roughly the same age groups as those used in the first experiment. The population of the younger age group consisted of 30 children (15 males, 15 females) with a mean age of 7;1 years (ranging between 6;6 and 8;3). The population in the older age group consisted of 30 children as well (17 males, 13 females) with a mean age of 10;2 years (ranging between 8;7 and 11;5). As in the first experiment, the younger children were in 'Group 3' of the Dutch primary school system, while the older children were in 'Group 6'. The children had not participated in the previous experiment, came from pre-dominantly Dutch-speaking families, and had been given permission to participate by their parents via signed agreement.

Materials. Experiment 2 was a partial replication of Experiment 1, in the sense that we again measured children's preferences for sweets, but that this time only one sweet was depicted in every trial (see Fig. 3). In all trials, a pre-recorded description of this sweet was played while it was shown. The descriptions were produced by a male voice with a neutral intonation, again avoiding prominent accents. Because there was only one sweet in every picture, the basic description *"The sweet"* was always sufficient for unambiguous identification, causing any modifier to be redundant.

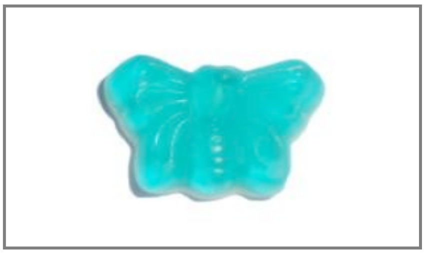


Fig. 3: Example of a criticial trial in the second experiment. While a trial was shown, a pre-recorded description of the sweet was played.

This second experiment had twelve critical trials in two conditions. In the *baseline condition* (represented by six trials), the depicted sweet was simply described as *"The sweet"*. The other six critical trials formed the *objective information condition*. Like in Experiment 1, sweets in this condition were described with objective information, but now only color modifiers were used to do this. For example, in Fig. 3, this would lead to a description such as *"The blue sweet"*. In order to avoid that a priori preferences for sweets would interfere with our results, the same six pictures of sweets were used in both conditions.

The crucial difference between Experiment 1 and 2 was related to the way in which the children marked their preferences for sweets in every trial: while a forced-choice task was used in the first experiment, we now used 5-point graded rating scales. This went as follows: after having been presented with a picture of a sweet and having listened to the corresponding description, the children were asked to indicate to what extent they liked the sweet. This was done via a pre-recorded question *("How much do you like this sweet?")*, which was played automatically after the description of a sweet, and was the same for all trials. The children could indicate the extent to which they liked the sweet by pointing at one of the cardboard smiley faces that were lying in front of them. Such facial representations of graded rating scales are commonly used in studies with children (e.g., Lockl & Schneider, 2002; Visser, Krahmer, & Swerts, to appear). As Fig. 4 shows, the smiley faces represented five categories, ranging from "not at all" to "very much".

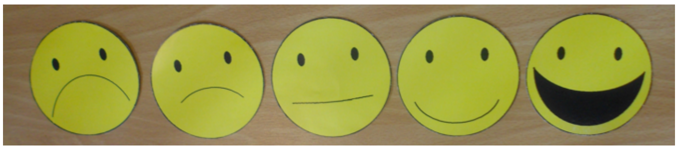


Fig. 4: The cardboard smiley faces representing the 5-point rating scale.

As fillers, we used pictures of sweets and other kinds of food (such as bananas and eggs). As in the critical trials, always one food item was depicted per trial. These items could be described by means of an affective modifier (such as delicious or disgusting), or via a combination of affective and objective information. Thirty-six fillers were included, resulting in a total of forty-eight trials for the entire experiment.

Procedure. The procedure of the current experiment was essentially similar to the one followed in the first experiment, and was the same for the children in the two age groups. The experiment was conducted in a quiet room in the school building, and had an average running time of approximately ten minutes. The language of the study was Dutch. The trials were presented in a fixed random order, which was the same for all children in both age groups.

The children were seated in front of a computer screen depicting the trials, with the experimenter seated next to the child for the course of the whole experiment. As in the first experiment, the instructions were provided orally, and explained that the children had to listen carefully to the descriptions that were given about the sweets. Moreover, the 5-point graded rating scale was introduced, and the children were asked explicitly to base their choices on the pre-recorded descriptions of the sweets. Two practice trials were included to acquaint the children with the rating scale.

After every trial, the experimenter marked on an answering form which of the five smiley faces the child had pointed at. Like in the previous experiment, the children had four seconds to make their choice, but more time was given if necessary. Afterwards, the children confirmed that they had understood the experimental task. Most of them indicated to think that the experiment was about food in general.

Design and statistical analysis. The experiment had one between-variable (*age* – levels: six to seven and nine to ten years old), and one within-variable (*condition* – levels: baseline and objective information), with as dependent variable the extent to which the children liked the sweets, with scores ranging from 1 to 5. We performed a Repeated Measures ANOVA to test for significance.

Results and discussion

Fig. 5 (on the next page) depicts the average scores of the younger and older children as a function of the two conditions.

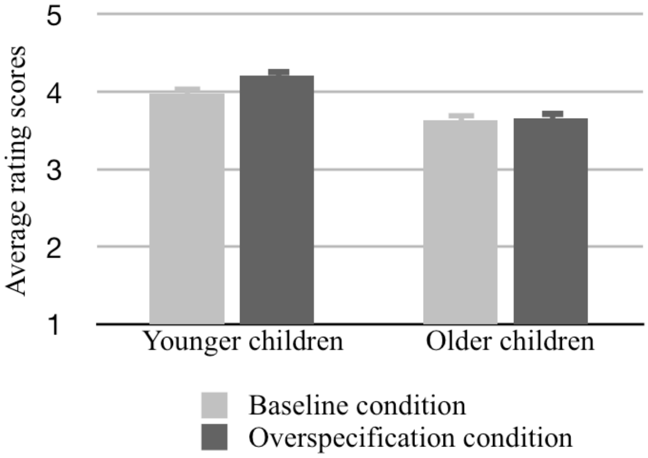


Fig. 5: The average scores of the children in the two age groups, as a function of the baseline and the overspecification conditions.

Firstly, the results revealed a main effect of *age* ($F_{(1,58)}$ = 6.065, p < .02, η^2 = .10), showing that the younger children (M = 4.08, SD = .13) generally appreciated the sweets better than the older ones (M = 3.64, SD = .13). Furthermore, there was a main effect of *condition* ($F_{(1,58)}$ = 3.977, p = .05, η^2 = .06), indicating that the scores in the trials where a sweet was described with objective redundant information (M = 3.91, SD = .09) were significantly higher than the scores in the baseline condition (M = 3.81, SD = .09). As Fig. 5 suggests, we also found a significant interaction between *age group* and *condition* ($F_{(1,58)}$ = 6.003, p < .02, η^2 = .09), showing that the effect of condition was due to the performance of the younger children: while the older children gave practically equal scores to the sweets in the baseline (M = 3.63, SD = .13) and the objective information (M = 3.65, SD = .13) conditions, the younger ones generally appreciated a sweet better when it was described with an objective modifier (M = 4.19, SD = .13) as compared to when this was not the case (M = 3.97, SD = .13).

The above interaction between age group and condition follows a similar pattern as compared to the results of the first experiment (in which we also found differences between the baseline condition and the overspecification condition for the younger children, but not for the older ones). However, this time, the effect sizes that we measured were generally smaller. Therefore, we zoomed in on the individual scores for all participants, to see whether the effects of redundancy were consistent across participants in either of the age groups. In doing this, we calculated for each child the difference between the scores in the objective information condition and the baseline condition. For example, if a child scored 4.15 on average in the baseline condition and 3.95 on average in the objective information condition, the individual score for this child was 0.20. Fig. 6 depicts the individual scores for all children that took part in this second experiment.

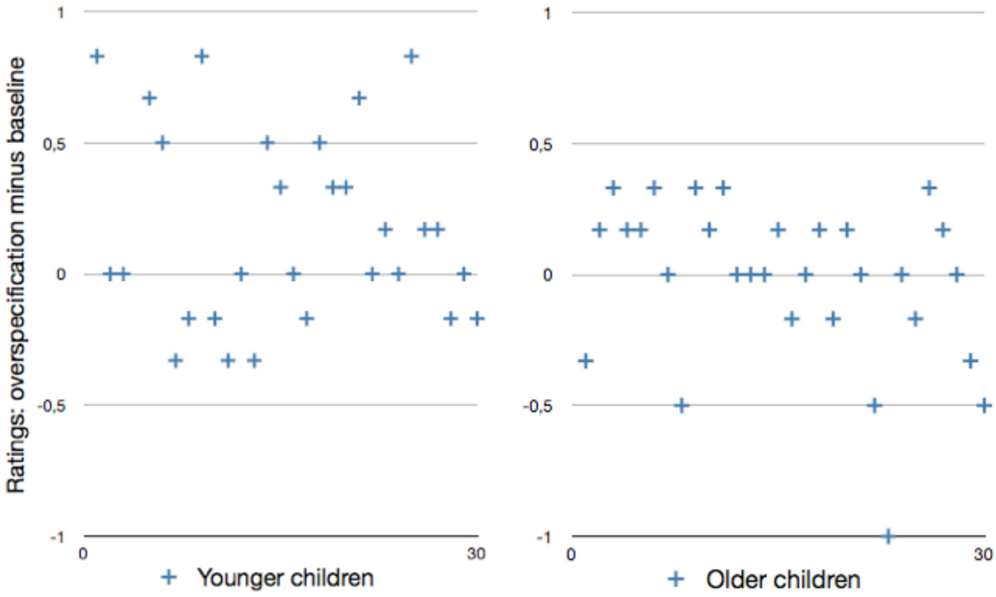


Fig. 6: Speaker variation in the two age groups: the difference in the scores for in the baseline and the overspecification conditions (y-axis) as a function of the individual children in the two age groups (x-axis). A positive value indicates that a child rated the overspecified descriptions higher than the minimal ones.

As can be seen in Fig. 6, the most important difference between the plots for the two age groups is that the younger children more often scored above 0.5. The number of children who scored zero (young = seven - older = eight) or below zero (young = eight – older = nine) was practically equal for the two age groups. These observations are in line with the fact that the effect sizes in this second experiment were generally rather small.

Discussion. The results of this second experiment again indicate that six- to seven-year-old children are guided in their preferences by objective redundant information to a larger extent than nine to ten-year-old ones. In particular, the younger children liked a sweet significantly better if it was described with a (redundant) color attribute as compared to when this was not the case, while the scores of the older children were practically identical in these conditions. The results of this second experiment (which were obtained using graded ratings) show the same pattern as those found in the first experiment (where forced-choice tasks were used), although the effect sizes that were measured were generally smaller.

General discussion

The results of the two experiments presented in this chapter have revealed that six- to seven-year-old children are influenced by objective redundant information in their preferences for sweets, while older, nine to ten-year-old children generally ignore such information in a similar setting. We have shown this to be true for two experimental tasks: one in which we used a forced-choice task (Exp. 1, in which children had to choose between two identical sweets, one of which was described with a redundant color of shape modifier), and one in which we used graded rating scales (Exp. 2, in which children were asked to indicate to what extent they preferred sweets that were either or not described with a redundant color modifier). The observed age differences at least suggest that the younger and older children comprehend redundant information in a different way. Is this due to pragmatic development?

Based on previous acquisition literature on *reference resolution*, the answer to this question should be yes. As we have discussed in the introduction of the current chapter, earlier papers on redundancy and overspecification in children's production of definite object descriptions report on differences between young children (five or six years old) and older ones (nine or ten years old) that are similar to the results of our experiments (e.g., Deutsch & Pechmann, 1982; Ford & Olson, 1975; Matthews et al. 2007; 2012). Furthermore, such age differences can be found in previous work on comprehension, showing that six-year-olds often have difficulties in comprehending redundant descriptions (Sonnenschein, 1982), while older children (nine years old) do not seem to have such problems. Siegal and Surian (2004) argued that these kinds of age differences are the result of children's pragmatic development, and explained them by suggesting that children must learn to understand the pragmatic implications of the (redundant) information that is provided to them in a conversation. Although the above papers generally used identification tasks (in which children had to comprehend distinguishing descriptions), the claim that pragmatic development takes place may also apply to our experiments: like in an identification experiment, the children in our experiments had to determine whether the information that they were presented with was relevant or not. In any case, it seems plausible to assume that the redundant modifiers somehow attracted the young children's attention in the first experiment (causing the redundantly described sweet to be more interesting than the other sweet), and made sweets more preferred in the second experiment (in which single sweets were described, either or not with a color modifier). Clearly, this was not the case for the older children that participated. If the differences observed here between the age groups indicate some form of pragmatic development, then what is the nature of this development?

One may argue that nine to ten-year-old children are usually better at understanding the implications that speakers have in mentioning redundant information than six- to seven-year-old children. In terms of Grice (1975), this implies that older children are better able to prevent themselves from deriving false implicatures: for example, in Experiment 1, our participants seem to have been aware that - since they assumed that the two sweets were similar (which was shown by our pre-experiment) - the objective information was redundant and therefore pragmatically irrelevant. Arguably, this was not the case for the young children, following their tendency to opt for the redundantly described sweets, and their tendency to rate these sweets higher than the sweets that did not have a redundant attribute in their description (as shown in Experiment 2). Thus, in general, it might be the case that the younger children are to a higher extent guided by pragmatically irrelevant information than the older ones, and that the older ones have learned how to judge to what extent information is relevant in a particular communicative context. One model of pragmatic development could then be that the younger children that took part in our experiments must still learn that a violation of the Gricean Maxim of Quantity (in our case due to redundancy) may sometimes evoke a conversational implicature, and that the older ones have learned that such violations are not necessarily intended as such by the speaker.

However, one could argue that it might be a bit over-simplified to assume that the children's preference for the sweets that were described with redundant information is solely the result of the derivation of false conversational implicatures. For example, in Experiment 1, it could be that the children realized that the two sweets were in fact identical, but that they were guided by the redundant information under the guise of being forced to choose. In other words, it could be that the children thought: "If I have to make a choice anyway, then why not choose for the sweet that is highlighted with extra information?" In this way, the results of our first experiment could be influenced by the experimental task. Arguably, the latter seems related to the Pragmatic Tolerance Hypothesis, as proposed by Katsos and Bishop (2011). As we have discussed earlier, Katsos and Bishop noticed that five-year-old children seem – at first sight – not fully pragmatically competent in deriving scalar implicatures, and claimed this to be due to the nature of the experimental task. Particularly, when five-year-old children were asked to indicate whether an underinformative description is optimal or not (that is, "good" or "bad" in child language), and they did this in a binary judgment task, they appeared not to reject violations of the Maxim of Quantity. However, the children's performance increased substantially when Katsos and Bishop (2011) replicated their experiment with graded rating scales instead of a binary judgment task, causing them to conclude that the low performance in their first experiment could be an artefact of the experimental task under which quantity implicatures were studied.

It is interesting to speculate on what the Pragmatic Tolerance Hypothesis would predict for the quantity implicatures studied here. Of course, some qualifications should be made while doing this. Firstly, Katsos and Bishop (2011) did not explicitly present their hypothesis as a proposal against the use of binary judgments or selection from binary options per se, but merely proposed that binary judgments on the felicity of a sentence might conceal a child's true pragmatic capabilities. Secondly, the binary judgment task used by Katsos and Bishop (2011) differed from ours in the sense that the children in our paradigm had to choose among a pair of sweets, and a choice for one sweet did not necessarily imply that a child 'rejected' the other sweet. However, as we see it, the results from our first experiment (in which we used forced-choice tasks) at least raise the suggestion that the younger children 'rejected' the sweets that were *not* described with a redundant modifier, because they opted for the redundantly described sweets in most of the cases. So what would the Pragmatic Tolerance Hypothesis predict for our data? With these caveats in mind, it may predict that, as compared to our first experiment (in which we used forced-choice tasks), the young children's performance should have improved in the second experiment (were we used graded ratings), and that even these younger children would ignore the redundant information there. However, our findings indicate that this is not the case: the pattern that we have found in this experiment was similar (albeit somewhat less pronounced) to that in the first experiment.

Does this mean that the Pragmatic Tolerance Hypothesis needs to be rejected for implicatures related to the effect of redundant modifiers on children's preferences for objects? Not necessarily, we believe. For example, one might also think of a language development model predicting that the young children in our experiments were aware that (referential) communication involves making pragmatic inferences, but that they "overgeneralized" in the sense that they expected all information in the descriptions of sweets to be relevant. For example, they might have thought that our speaker had a communicative purpose in uttering the redundant modifiers. In general, this would imply that six- to seven-year-old children are already capable of making pragmatic inferences, but are not aware that there are also communicative situations in which such inferences should not be made (i.e., where a speaker does not necessarily have any communicative intentions at all). Hence, because the nine to ten-year-old children in our experiments were not guided by the redundant information, it might be that pragmatic development has taught these children to determine that the objective modifiers in our stimuli were in fact redundant.

Conclusion

In the current chapter, we studied, for the first time, how children in different age groups are guided in their preferences by objective redundant information in referring expressions. Crucially, we found that only the six- to seven-year-old children rely on redundant information in their preferences for objects, while this was not the case for nine- to ten-year-old children. In particular, the young children preferred a sweet that was redundantly referred to as, say, "green", better than the older ones, irrespective of the task that was used to study these preferences. These results suggest that children learn how to avoid deriving false implicatures triggered by the inclusion of redundant information in descriptions.

Acknowledgments

We gratefully acknowledge the cooperation of the children, their parents and the teaching staff of St. Gerardus Primary School in Deurne and Mgr. Zwijssen Primary School in Kerkdriel (both in the Netherlands). We thank Hanneke Schoormans for assistance in creating the stimuli of Experiment 1, and Maud Broekmeulen for collecting the data of Experiment 2.

References

Ackermann, B. (1981). Performative bias in children's interpretations of ambiguous referential communications. *Child Development,* 52, 1224-1230.

Ackermann, B., Szymanski, J. & Silver, D. (1990). Children's use of common ground in interpreting ambiguous referential utterances. *Developmental Psychology,* 26, 234-245.

Arts, A. (2004). Overspecification in instructive texts. Dissertation, Tilburg University. Wolf Publishers, Nijmegen.

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.

Bonitatibus, G., Godshall, S., Kelley, M., Levering, T. & Lynch, E. (1988). The role of social cognition in comprehension monitoring. *First Language,* 8, 287-298.

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

Davies, C. & Katsos, N. (2010). Over-informative children: Production / comprehension asymmetry or tolerance to pragmatic violations? *Lingua,* 120, 1956-1972.

Deutsch, W. & Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition,* 11, 159-184.

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.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D. & Pethick, S. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, 59, 1-185.

Ford, W. & Olson, D. (1975). The elaboration of the noun phrase in children's description of objects. *Journal of Experimental Child Psychology,* 19, 371-382.

Geurts, B. & Nouwen, R. (2007). At least et al: The semantics of scalar modifiers. *Language*, 83, 533-559.

Geurts, B., Katsos, N., Cummins, C., Moons, J., & Noordman, L. (2010). Scalar quantifiers: Logic, acquisition, and processing. *Language and Cognitive Processes*, 25 (1), 130-148.

Grice, H. (1975). Logic and conversation. In: Cole, P. & Morgan, J. L. (Eds.), *Speech Acts*. Academic Press, New York, 41-58.

Grice, H. (1989). *Studies in the way of words*. Harvard University Press, Cambridge, MA.

Katsos, N. & Bishop, D. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition,* 120, 67-81.

Krahmer, E., Noordewier, M., Goudbeek, M., & Koolen, R. (2013). How big is the BFG? The impact of redundant size adjective on size perception. In *Proceedings of the 35th annual meeting of the Cognitive Science Society (CogSci 2013)*, Berlin, Germany.

Levinson, S. (2000). *Presumptive meaning. The theory of generalized conversational* *implicature*. MIT Press, Cambridge MA.

Lockl, K., & Schneider, W. (2002). Developmental trends in children's feeling-of-knowing judgments. *International journal of Behavioral Development*, 26, 327-333.

Matthews, D., Lieven, E. & Tomasello, M. (2007). How toddlers and preschoolers learn to uniquely identify objects for others: a training study. *Child Development,* 78, 1744-1759.

Matthews, D., Butcher, J., Lieven, E. & Tomasello, M. (2012). Two- and four-year-olds learn to adapt referring expressions in context: Effects of distracters and feedback on referential communication. *TopiCS in Cognitive Science,* 4, 184-210.

Musolino, J. (2004). The semantics and acquisition of number words: integrating linguistic and developmental perspectives. *Cognition*, 93, 1-41.

Nadig, A. & Sedivy, J. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science,* 13, 329-336.

Noveck, I. & Reboul, A. (2009). Experimental pragmatics: a Gricean turn in the study of language. *Trends in Cognitive Sciences,* 12, 425-431.

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

Sedivy, J. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32 (1), 3-23.

Siegal M. & Surian, L. (2004). Conceptual development and conversational understanding. *Trends in Cognitive Sciences,* 8, 534-538.

Sonnenschein, S. (1982). The effects of redundant communication on listeners – when more is less. *Child Development,* 53, 717-729.

Sonnenschein, S. & Whitehurst, G. (1982). The effects of redundant communications on the behavior of listeners: Does a picture need a thousand words? *Journal of Psycholinguistic Research,* 11, 115-125.

Tomasello, M., Carpenter, M. & Liszkowski, U. (2007). A new look at infant pointing. *Child Development,* 78, 705-722.

Visser, M., Krahmer, E., & Swerts, M. Children's expression of uncertainty in collaborative and competitive contexts. To appear in *Language and Speech*. DOI: 10.1177/0023830913479117.

Whitehurst, G. (1976). Development of communication-changes with age and modeling. *Child Development,* 47, 473-482.

7

General discussion and conclusion

In this dissertation, we have reported on the results of a series of experiments to find patterns in the human production and comprehension of overspecified definite object descriptions. In five studies, we have addressed the following research question:

"Which factors cause human speakers to overspecify their target descriptions, and to what extent can these be modeled by existing REG algorithms?"

Regarding reference production, we have found empirical support for the majority of the factors that we manipulated, showing that these indeed have an effect on the extent to which speakers overspecify. These factors are mainly related to *visual scene perception*. However, regarding the *referential task* and the *communicative setting* in which object descriptions are uttered, our results did not show reliable effects. Lastly, we have found developmental changes in children's *compehension* of redundant attributes in the object descriptions they are presented with. In this final chapter, we summarize our findings, formulate implications for the current REG algorithms, and elaborate on directions for future research.

Overspecification and visual scene perception

Throughout this dissertation, our results have revealed that various factors related to *visual scene perception* affect speakers' tendency to overspecify. Firstly, in Chapter 2, we found that the *referential domain* in which a target referent occurs has an effect on overspecification. In this study (which was inspired by the TUNA study by Van Deemter, Gatt, Van der Sluis, & Power, 2012a), we had participants describe objects that could occur in two domains: half of the scenes contained artificial pictures of furniture items, and the other half consisted of photo-realistic pictures of people. We expected to find that the descriptions of the people pictures (which gave speakers many referential possibilities to describe the target, since they differed on a relatively high number of attributes) would be more likely to contain more redundant attributes than the descriptions of the furniture items (where speakers had fewer attributes to choose from). Our findings confirmed this hypothesis, implying that the specification level of human-produced descriptions is influenced by the referential possibilities that speakers have in a given domain, and – given that the people pictures were arguably more perceptually similar to each other than the furniture items – that differences between domains in the level of perceptual similarity can influence the occurrence of overspecification.

In order to study the number of choices available to a speaker and the assumed role of perceptual similarity more accurately, we also manipulated the *amount of variation within a single domain* (Chapter 4), where we opted for the furniture domain. In three experiments, we systematically tested how speakers describe objects occurring in low and high variation scenes within the furniture domain, expecting to find that scenes of the latter kind would cause speakers to overspecify more (with a color attribute) than scenes of the former kind. We indeed found that this was the case when the difference between the two conditions was large, but also when this difference was subtler (i.e., when the scenes only differed in terms of color variation), or when the objects in the low variation condition were all of the same type.

Thirdly, on the level of specific objects that are present in a visual scene, we found an effect of *cardinality*, showing that plural references (in which speakers refer to two target referents at once, as in "The brown chair and the green desk") are more likely to be overspecified than singular ones. We found this at least to be true for the object descriptions in our Dutch TUNA corpus, and our explanation for this effect of cardinality is twofold. On the one hand, as Arnold and Griffin (2007) propose, there might have been competition between the two targets, which forces speakers to divide attention, causing the referring task to become more difficult. On the other hand, in line with the reasoning by Gatt and Van Deemter (2007), one could reason that when speakers are confronted with two targets, they tend to conceptualize these in parallel fashion, meaning that if one attribute is used to refer to the one target, it will probably also be used in the description of the second one.

In our fourth study (Chapter 5), our focus was not on characteristics of a target, but on the extent to which the *distractor objects* that are part of a scene shape a speaker's object descriptions in terms of overspecification. Firstly, we found an effect of *visual clutter* here: in scenes where clutter objects were present (which we define as objects that are thematically related to the target), speakers were more likely to overspecify (with a color attribute) than in scenes in which this was not the case. Secondly, we found an effect of *distractor type*, showing that speakers more often included a redundant color attribute when the target and its distractors were all of the same type as compared to when they had different types. This effect interacted with the effect of *distractor color*, revealing that descriptions were more likely to contain a redundant color attribute when there was color variation between the target and its distractors (which is consistent with our results on visual scene variation presented in Chapter 4). The effect of color was most convincing in the scenes in which all objects had the same type. We explain these effects by means of the probability that objects are actually considered as a relevant distractor. More specifically, we conclude that a given object is most likely to be part of the distractor set if there is no visual clutter present in a scene, and if the object has the target's type, but not its color.

One manipulation in Chapter 5 that we hypothesized to have a similar impact, was *distractor distance*, which we defined as the physical distance between target and distractor. In our study, objects could be either close or distant, and we expected to find that close objects would be more likely to be considered as a relevant distractor than distant ones. However, contradictory to our intuitions and to previous literature on this topic (e.g., Beun & Cremers, 1998), this was not found to be the case, or only weakly so. One explanation could be that the assumed effect of distractor distance was subsumed by the effect of type: since the type of distractor was different in the majority of trials, there was no point considering the distractor, even in situations where it was close to the target. Another possible explanation is related to the fact that we showed speakers a 2D representation of a 3D visual scene, which may have limited the effect of distance. An interesting line for future research could therefore be to investigate the effect of distance by using realistic 3D scenes that are probably more difficult to scan as a whole than our 2D scenes.

Visual scene perception in REG. The above findings regarding the impact of scene perception on speakers' tendency to overspecify their object descriptions have several interesting implications for the current REG algorithms. The first one is related to the distractor set that algorithms work with when generating a description. As explained in the introduction of this dissertation, most current REG algorithms (with the notable exception of Mitchell et al.'s (2012) *Midge* algorithm) use a database that is basically a semantic representation of all objects (including their attributes) that are present in a particular scene. Algorithms such as the Incremental Algorithm (Dale & Reiter, 1995) generally consider any object that is present in the direct visual context of the target object to be part of the distractor set. In practice, this usually means that the distractor set is taken to exist of all objects that are visible, except for the target itself. Along the lines of Krahmer and Theune (2002), who used linguistic salience to enable the IA to restrict the distractor set, we follow Kelleher and Kruijff (2006) in proposing that *visual saliency cues* can be used to do this as well.

Secondly, our findings regarding visual scene perception suggest that speakers rely on *heuristics* when processing a visual scene. In others words, speakers seem to use shortcuts rather than compute exact calculations when deciding on the content of their object descriptions. The original idea behind people relying on heuristics (in decision making) comes from Tversky and Kahneman (1982), while Van Deemter, Gatt, Van Gompel and Krahmer (2012b) suggest that heuristics guide speakers in the production of referring expressions as well. However, Van Deemter et al. do not point out which specific heuristics speakers might use. Based on our experiments, we can suggest that human speakers are guided by perceptual cues related to similarities and differences between and within different *domains*, by the presence of *visual clutter*, and by visual saliency cues on the level of specific objects that are present in a scene, including the *plurality* of a target and the *type* and *color* of the distractors. All these factors have been shown to affect the redundant use of color, and the occurrence of overspecification in general.

If one assumes that REG research aims to build algorithms that are able to generate object descriptions that are comparable to those produced by human speakers, the question is how the use of heuristics can be incorporated in the existing systems. For example, how can the IA learn that it should overspecify more when a target object occurs in a high rather than a low variation scene? Or that descriptions are more often overspecified in the case of cluttered scenes? As we have discussed, one option may be to dynamically adapt the preference order for every scene and description. One can for example make sure that color is always at the head of the Preference Order (PO) when there is color variation in a visual scene, causing it to be the first attribute that is considered for selection. However, this might not be the best solution, because of the algorithm's deterministic nature: for example, in this scenario, it would *always* select color in visual scenes with color variation, and *never* in scenes where all objects have the same color. As our results (and those from many previous studies on referential overspecification) indicate, this is not what humans do. In future research, it would thus be interesting to develop REG algorithms with a probabilistic nature, such as the Probabilistic Referential Overspecification algorithm (as recently introduced by Van Gompel, Gatt, Krahmer, and Van Deemter (2012)).

Van Deemter et al. (2012a) raise another limitation of the IA that is related to the use of a PO, namely that systematically searching for the best performing PO in a previously unstudied domain often becomes impractical, because of the large amount of possibilities that need to be considered. Furthermore, collecting large, semantically transparent corpora to learn what attributes speakers actually prefer in a new domain is often time consuming. However, the results of the learning curve experiments that we present in Chapter 3 show that such corpora do not necessarily have to be big: training on a handful of descriptions can already lead to a performance that does not significantly differ from training on a substantially larger training set. We have shown this to be the case in two domains (pictures of furniture and people), in two languages (English and Dutch), and for two algorithms (Dale and Reiter's (1995) IA and the Graph algorithm by Krahmer, Van Erk, and Verleg (2003)). An interesting direction for future research would be to replicate these experiments by using more open-ended domains and more complex object descriptions, such as the plural references studied by Van Deemter et al. (2012a).

Overspecification as related to tasks and communicative settings

For some of the factors we tested, we were not able to find reliable effects. Firstly, in Chapter 5, we looked into the way in which the *specificity of the referential task* causes speakers to overspecify their descriptions (with color). We presented one half of the participants in this study with a general instruction (i.e., "Describe this object"), while the other half heard a more specific instruction that contained the target's type (e.g., "Describe this plate"). We hypothesized speakers with the former, more general, task to overspecify more often, since in this case any distractor object that was present in the scene had to be ruled out. Along the same lines, in the more specific task, we expected the distractor set to be restricted to only those objects that shared their type with the target referent. Although we indeed found numerical differences between the conditions in this direction, these were not statistically reliable. Another direction for future research would therefore be to come up with more accurate ways to manipulate a participant's instruction, where a distinction can be made between what speakers are focusing on (or have been instructed to focus on) when describing an object, and the communicative intention that they have in doing this (for example, describing versus instruction giving; see also Arts, Maes, Noordman, and Jansen (2011)).

Secondly, in Chapter 2, we studied how different *communicative settings* influence speakers' tendency to overspecify. We had two main hypotheses in this direction: we expected spoken descriptions to be more frequently overspecified than written ones, and hypothesized that descriptions are more likely to be overspecified when speakers cannot see their addressee as compared to when they can. Again, we found numerical differences in the expected directions, with speakers using most redundant attributes in the setting in which they could openly see and speak to their addressee, and least in the setting where they produced written object descriptions to an imaginary addressee. However, again, these differences were not statistically reliable, meaning that we did not find empirical evidence to support our hypotheses at these points. In retrospect, we believe that this might have been due to the fact that the communication in this experiment was rather one-sided and thus not very interactive. This may have caused the role of the addressee to be equal in the two spoken conditions, and comparable to the (marginal) role of the imaginary addressee in the written condition.

Interactivity in REG. Although the above manipulations regarding the specificity of the referential task and the communicative setting have only resulted in numerical differences between the respective conditions (which prevents us from formulating direct implications for current REG algorithms), they lead us to shortly discuss one general limitation that goes for all our production experiments, related to interactivity. As we have explained in the General Introduction, the focus in this dissertation has been on "one-shot" descriptions, implying that we have not addressed the common assumption that the *discourse context* (between speaker and addressee) influences the way in which objects are described. For example, Brennan and Clark (1996) have shown that language partners form conceptual pacts when repeatedly referring to the same object (reducing the number of words and attributes that are mentioned), while others have added that repeated object descriptions are generally reduced in the number of gestures they contain as compared to initial ones (e.g., Hoetjes, Koolen, Goudbeek, Krahmer, & Swerts, 2011; Holler & Stevens, 2007; De Ruiter, Bangerter, & Dings, 2012). Similarly, the discourse context may also affect overspecification.

The reason why we have focused on initial, one-shot descriptions was motivated by the fact that many classic REG algorithms tend to do this as well. In their survey, Krahmer and Van Deemter (2012) explain that "it is still a largely open question to what extent the classical approaches to REG can be used in interactions" (p. 206), and that sophisticated addressee modelling is likely to be computationally expensive. Nevertheless, there is some relevant previous work on how dialogue-related processes can be incorporated in Natural Language Generation systems. For example, in line with observations showing that language partners tend to 'align' with each other when they are in a conversation (e.g., Pickering & Garrod, 2004; Goudbeek & Krahmer, 2012), Viethen, Dale and Guhe (2011) combined the traditional REG perspective (that aims to select distinguishing attributes) with a perspective where earlier descriptions of a given object are also taken into account. Similarly, as discussed earlier, Krahmer and Theune (2002) presented an extension of the IA that can dynamically restrict the distractor set based on liguistic saliency cues. Moreover, another example of a paper studying REG in interactive settings comes from Stoia, Shockley, Byron and Foster-Lussier (2006), who took both the dialogue history and the spatial visual context into account. Furthermore, in the GIVE challenges (e.g., Koller et al., 2010; Striegnitz et al., 2011), NLG systems had to guide human addressees in solving an instruction task in a 3D environment. However, the above work does not necessarily address how discourse-related factors may affect the occurrence of referential overspecification. In line with Jordan and Walker (2005), who present a computational system that enables the generation of descriptions that are overspecified in dialogue (using information that has already been used earlier in the discourse), we believe that an interesting line for future research would be to focus on reduction in repeated reference as a function of overspecification.

The comprehension of overspecified reference

In our last study (Chapter 6), we investigated the comprehension of overspecified reference, and in particular how children develop in processing redundant information in object descriptions. Our participants were six- to seven and nine- to ten-year-old children. We hypothesized the younger children to more often derive false implicatures (Grice, 1975) than the older ones, in the sense that these young children would be more likely to be guided in their preferences by redundant information than the older ones. Following Katsos and Bishop (2011), who emphasized the importance of the experimental task in studying children's pragmatic capabilities, we used two different methods to investigate the extent to which children are guided by redundant attributes in their preferences for sweets. In our first experiment, children saw pictures of two similar-looking sweets, one of which was described with a redundant color or shape modifier. In a forced-choice task, the children were asked which sweet they preferred. The results indeed showed an effect of age: the young children more often preferred the sweets described with a redundant modifier, whereas the older ones performed at chance level. In order to make sure that this effect was not due to the forced-choice task, we replicated the first experiment using graded rating scales, using pictures of one sweet, half of which were described with a redundant color or shape attribute. We again found an effect of age, showing that the younger children liked the redundantly described sweets significantly better than the older ones. These findings suggest that children develop their pragmatic capabilities between the age of roughly six and nine.

Implications for REG. Although the above comprehension study was not intended to formulate direct implications for REG, it has at least one important implication that needs to be emphasized here: redundant attributes may affect addressees in ways that are not necessarily related to target identification. As we have pointed out throughout the dissertation, redundant information might help (e.g., Arts, Maes, Noordman, & Jansen, 2011) or hinder (e.g., Engelhardt, Bailey, & Ferreira, 2006) addressees when identifying a target. The results of our two comprehension experiments add to this that redundant attributes can also affect an addressee's preferences for objects (at least when it concerns children). Related to this, Mooney (2004) argues that it depends on the context in which an utterance takes place whether all information that is part of an object description is relevant. Thus, in an ideal world, also NLG systems would have to cope with other (non-linguistic) goals that involve more than the communication of information alone.

Conclusion

This dissertation has shown various factors to affect the occurrence of referential overspecification in the *production* of definite object descriptions, which are mainly related to visual scene perception. We have also seen that redundant information has an effect on *comprehension*, in the sense that it guides children in their preferences for objects. Throughout the dissertation, we have formulated implications of our findings for the classic algorithms in the field of Referring Expression Generation. As we have discussed above, interesting lines for future research involve using more realistic 3D scenes, studying more complex referring expressions in open-ended domains, taking discourse-related factors into account, and addressing other communicative purposes than target identification alone.

References

Arnold, J. & Griffin, Z. (2007). The effect of additional characters on choice of referring expressions: everyone competes. *Journal of Memory and Language*, 56, 521-536.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, 43 (1), 361-374.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification in written instruction. *Linguistics*, 49 (3), 555-574.

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

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

De Ruiter, J.P., Bangerter, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: investigating the Tradeoff Hypothesis. *Topics in Cognitive Science*, 4, 232-248.

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

Gatt, A., & Van Deemter, K. (2007). Incremental generation of plural descriptions: Similarity and partitioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2007)*. Prague, Czech Republic.

Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: content planning, modifier ordering, and referential overspecification. *Topics in Cognitive Science*, 4 (2), 269-289.

Grice, H. (1975). *Logic and conversation*. In: Cole, P. & Morgan, J. L. (Eds.), Speech Acts. Academic Press, New York, pp. 41-58.

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2011). GREEBLES, Greeble, greeb: On reduction in speech and gesture in repeated reference. In *Proceedings of the 33rd annual meeting of the Cognitive Science Society (CogSci 2011)*. Boston Massachusetts. 3250-3255.

Holler, J., & Stevens, R. (2007). An experimental investigation into the effect of common ground on how speakers use gesture and speech to represent size information in referential communication. *Journal of Language and Social Psychology*, 26, 4-27.

Jordan, P., & Walker, M. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24, 157-194.

Katsos, N., & Bishop, D. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition,* 120, 67-81.

Kelleher, J. & Kruijff, G.J. (2006). Incremental generation of spatial referring expressions in situated dialogue. In *Proceedings of COLING/ACL '06*. Sydney, Australia.

Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Moore, J., & Oberlander, J. (2010). The first challenge on generating instructions in virtual environments. In E. Krahmer & M. Theune (Eds.), *Empirical methods in Natural Language Generation.* Berlin and Heidelberg: Springer (LNCS 5790).

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

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In: K. van Deemter & R. Kibble (Eds.). *Information sharing: Givenness and newness in language processing* (pp. 223-264). CSLI publications, Stanford.

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Computational Linguistics*, 38 (1), 173-218.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A, Berg, T., Daumé III, H. (2012). Midge: generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European chapter of the Association for Computational Linguistics (EACL 2012)*. Avignon, France.

Mooney, A. (2004). Co-operation, violations and making sense. *Journal of Pragmatics*, 36 (5), 899-920.

Pickering, M., & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27, 694-718.

Stoia, L., Shockley, D., Byron, D., & Foster-Lussier, E. (2006). Noun phrase generation for situated dialogs. In *Proceedings of the 4th International conference on Natural Language Generation (INLG 2006)*, Morristown, NJ, USA. 81-88.

Striegnitz, K., Denis, A., Gargett, A., Garoufi, K., Koller, A., & Theune, M. (2011). Report on the *second* second challenge on generating instructions in virtual environments (GIVE-2.5). In *Proceedings of the Generation Challenges Session at the 13th European workshop on Natural Language Generation (ENLG 2011).* Nancy, France.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, New Series, Vol. 185, 1124-1131.

Van Deemter, K., Gatt, A., Van der Sluis, I., & Power, R. (2012a). Generation of referring expressions: assessing the Incremental Algorithm. *Cognitive Science*, 36, 799-836.

Van Deemter, K., Gatt, A., Van Gompel, R., & Krahmer, E. (2012b). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, 4 (2), 166-183.

Van Gompel, R., Gatt, A., Krahmer, E., & Van Deemter, K. (2012). PRO: a computational model of referential overspecification. Submitted to *Architectures and Mechanisms for Language Processing (AMLaP)*, to be held in Marseille, France.

Viethen, J., Dale, R., & Guhe, M. (2011). Generating subsequent reference in shared visual scenes: computation vs. re-use. In *Proceedings of the Conference on empirical methods in Natural Language Processing.* Edinburgh, Scotland.

Need I say more?

Summary

In everyday life, people often routinely produce descriptions of physical objects in the world around them. Examples of such descriptions are “the brown chair” or “the guy wearing leather pants”. One obvious reason why they might do this is that they want to distinguish their target referent from any other object in the direct visual context that could serve as a relevant distractor. Given the ubiquity of target descriptions in speakers’ everyday language, they are often the subject of study in scientific research, mainly in the fields of psycholinguistics and computational linguistics. One interesting observation in psycholinguistics is that speakers generally tend to provide their listener with redundant attributes that are not necessary for unique identification of the target. In other words: speakers often overspecify their descriptions. Overspecification has implications for computational linguists who build Referring Expression Generation (REG) algorithms that are able to automatically generate definite target descriptions.

Although it is debated whether REG algorithms are built to mimic human reference production, they do make interesting predictions that are relevant for psycholinguistics. For example, some REG algorithms predict that target descriptions can actually be overspecified. Nevertheless, factors that cause human speakers to overspecify are still largely unknown. Therefore, in this dissertation, the following research question has been addressed:

“Which factors cause human speakers to overspecify their target descriptions, and to what extent can these be modeled by existing REG algorithms?”

This dissertation has aimed to answer this question by reporting on the results of psycholinguistic experiments, searching for patterns in the human production (study 1, 2, 3, 4, as described in chapter 2, 3, 4, 5) and processing (study 5, as described in chapter 6) of overspecified target descriptions. In doing this, we have aimed to formulate implications for the performance of current REG algorithms.

Study 1

The first study presented the D-TUNA corpus, which is a large corpus of semantically annotated Dutch target descriptions; that is, all its expressions (that were produced by participants in a large-scale production experiment) were annotated with information regarding attributes of both the target and distractor objects. The data collection and annotation was inspired by the English TUNA corpus.

In this first study, we used the D-TUNA corpus to explore various factors that might cause speakers to overspecify their target descriptions, related to the properties of the target referent and to properties of the communicative setting. Firstly, regarding properties of the target, we found an effect of the *domain* in which a description is uttered (i.e., artificial pictures of furniture items vs. realistic pictures of people): descriptions of people contained more redundant attributes than descriptions of furniture items. This implies that perceptual similarity among objects in a visual scene affects the amount of redundant information that speakers provide about a target. Secondly, again related to properties of the target, we found the *number of target referents* that a visual scene contains (one or two) to affect overspecification, where we observed a higher level of redundancy in plural references. These results suggest that referring to two objects in a parallel fashion is more difficult for the speaker, and attributes that are used to refer to one target are likely to be used for the other target as well.

With regard to properties of the *communicative setting*, our results did not show reliable effects. Manipulations involved both the modality (speech vs. writing) and the interactivity (monologue vs. dialogue) of the referential setting. Although we found numerical differences in the expected directions, with speakers using most redundant attributes in the setting in which they could openly see and speak to their addressee, and least in the setting where they produced written object descriptions to an imaginary addressee, these differences were not statistically reliable. In retrospect, we believe that this might have been due to the fact that the communication in this experiment was rather one-sided and thus not very interactive.

Study 2

The second study investigated *learning curves* for REG algorithms: how many human-produced references are needed to make a good estimate of which attributes are preferred in a given domain? For previously unstudied domains, recent scientific work has argued that systematically considering all possible preference orders for a certain domain is impractical, because there are often too many possibilities to test. Given that this problem exists for various REG algorithms (e.g., the Incremental Algorithm and the Graph-based algorithm, which both in their own way use attribute preferences), this second study explored how difficult it actually is to determine which attributes are preferred in a new domain. Are hundreds of human-produced instances needed, or would a few of them do?

We answered this question for two algorithms (IA and Graph), for two domains (furniture and people), and for two languages (English and Dutch; we used the TUNA and D-TUNA corpora). The results or our learning curve experiments showed that corpora do not necessarily have to be big: training on a handful of target descriptions already led to a performance that did not significantly differ from training on a substantially larger training set. These findings were generally consistent across the various algorithms, domains, and languages. This suggests that collecting large, semantically transparent corpora is not necessarily needed to find the best preference order in rather simple, limited domains (such as the furniture and people domains).

Study 3

The third study explored the link between visual scene perception and referential overspecification, and in particular how the amount of *variation* in a scene relates to speakers' tendency to overspecify. The motivation for this study came from one of the findings presented in the first study, being that speakers were more likely to include one or more redundant attributes when they are to refer to photo-realistic pictures of people rather than to artificial pictures of furniture items. In Study 1, we argued that this could be due to the difference in the extent to which the pictures in the two domains were realistic, but also to the fact that the pictures in the people domain left speakers with more possible attributes to distinguish the target. This raised the suggestion that the amount of visual variation may have played a role in the amount of redundant attributes that speakers included.

In order to test this suggestion directly, the third study reported on three experiments that in a different (but related) way manipulated the amount of variation within a single domain. More specifically, we systematically tested whether speakers were more likely to overspecify when they are presented with high variation scenes as compared to low variation scenes. The results showed that this was indeed the case when the difference between the low and high variation condition was large (Experiment 1), but also when this difference was more subtle (Experiment 2), or when the objects in the low variation condition all had the same type (Experiment 3). Our findings raise the suggestion that speakers use quick heuristics when processing a scene (e.g., 'use color if there is color variation in the scene').

The trials of the three experiments of this study were designed in such a way that the Incremental Algorithm would – under certain conditions – not include color redundantly in a description of a target. This implies that there are referential situations in which speakers are influenced by the amount of visual variation in a scene, while REG algorithms such as the IA are not.

Study 4

The fourth study went further into the connection between visual scene perception and overspecification by exploring factors that might determine whether or not any object in a scene is regarded as a relevant distractor of the target. Since it is often not explained explicitly how the set of distractors should be determined for REG algorithms, these algorithms usually take all objects that are present in a scene into account as relevant distractors. In line with observations that linguistic salience may play a role in limiting the distractor set, this fourth study focused on the impact of visual salience. In particular, we tested how various bottom-up, perceptual saliency cues, and one top-down, conceptual saliency cue guided speakers in restricting the set of relevant distractors, and in their redundant use of color.

Firstly, with regard to bottom-up saliency, our results showed an effect of visual clutter on the redundant use of color: in cluttered scenes, speakers were more likely to overspecify with a color attribute than in non-cluttered scenes. Secondly, we found effects of distractor type, showing that speakers more often overspecified with color when the target and its distractors were all of the same type as compared to when they all had different types. This effect interacted with the effect of color, showing that descriptions were more likely to be overspecified with color if there was color variation in the scene. Overall, these findings again raise the suggestion that speakers rely on heuristics when processing a scene and referring to target objects. Given that the stimuli in this study were again built in such as way that algorithms such as the IA would not include color (with a few exceptions), this suggests that the current REG algorithms should be learned how to dynamically limit the distractor set in certain referential situations.

Our manipulations regarding distractor distance (varying the physical distance between the target and a distractor, which could either be close or distant) did not result in reliable differences between conditions. It might have the case that the effect of distance was subsumed by the effect of type: since the type of distractor was different in the majority of trials, there was no point considering the distractor, even in situations where it was close to the target. Another explanation could be that we presented speakers with a 2D representation of a 3D visual scene, which may have limited the effect of distance as well. With regard to top-down saliency, we tested the effect of task specificity, where half of the participants had a specific task in which the target's type was mentioned ("describe this X"), while the other half had a more general task ("describe this object"). However, again, we did not find reliable differences between conditions.

Study 5

The fifth and final study investigated how listeners *process* redundant attributes. This study broadened the scope of this dissertation in the sense that it did not focus on target identification as the purpose of reference, but on the extent to which objective redundant modifiers (providing color and shape information) guided children in their preferences for objects. Since young children's pragmatic capabilities are naturally under development, they must learn to understand the implications of (redundant) information that is provided to them when they are in a conversation. Therefore, in this last study, we studied to what extent there are developmental differences between six- to seven-year-old children and nine- to ten-year-old children in how they were guided by redundant information in their preferences for sweets. We used two methodologies to investigate this: forced-choice tasks (Experiment 1) and graded rating scales (Experiment 2).

The results of our first experiment indeed showed an effect of age: the young children more often preferred the sweets described with a redundant modifier, whereas the older ones performed at chance level. This effect was observed in the second experiment as well, where we found that the young children liked the redundantly described sweets significantly better than the older ones. These findings suggest that children develop their pragmatic capabilities between the age of roughly six and nine.

Conclusion

This dissertation has shown various factors to affect the occurrence of overspecification in the production of definite object descriptions, which are mainly related to visual scene perception. In this respect, we found effects related to the referential domain, the plurality of the target, the amount of visual variation in a scene, the presence of visual clutter in a scene, and to various characteristics of the distractor objects that are present in a scene (i.e., the color and type of a distractor). All these factors have been shown to affect the extent to which speakers overspecify their target descriptions. In addition to this, we have found that objective redundant information has an effect on comprehension, in the sense that such information guides children in their preferences for objects.

Throughout the dissertation, we have aimed to formulate implications of our findings for algorithms in the field of Referring Expression Generation that automatically generate definite target descriptions. With respect to this, our conclusions mainly stress the important role of visual saliency cues and heuristics in human reference production, and the way in which these cause speakers to overspecify.

Need I say more?

Acknowledgments

Gekscherend wordt wel eens gezegd dat het dankwoord het meest gelezen onderdeel is van een dissertatie. Zoals ik inmiddels heb ondervonden, doet deze uitspraak geen recht aan de hoeveelheid kruim die het kost om een proefschrift te vullen met hoogwaardige wetenschappelijke inhoud. Toch ben ik blij dat het moment van bedanken is gekomen. En dan niet alleen omdat dit betekent dat die wetenschappelijke inhoud inmiddels in hoogwaardige vorm op papier staat, maar toch vooral omdat er ook echt iets te bedanken valt.

Bovenal gaat mijn grote dank uit naar mijn twee promotoren, Emiel Krahmer en Marc Swerts. Ik leerde hen reeds kennen tijdens het schrijven van mijn masterscriptie aan de Universiteit van Tilburg, waarvoor ik (in een zeer prettige samenwerking met Iris Pfrommer) onderzoek deed naar non-verbaal gedrag van nieuwslezers. Toen dit project ten einde liep, en ik (wanhopig) op zoek was naar een verlenging van mijn studentenbestaan, attendeerde Emiel mij op de relatief nieuwe onderzoeksmaster Language and Communication. Of dat niks voor mij was. Ja, dat was het zeker! Deze master heeft voor mij namelijk het pad geplaveid voor een carrière als promovendus: ik kon mijn Research Training volgen in het destijds door Emiel gelanceerde VICI-project over verwijzende expressies. Toen hij daarvoor ook nog eens op zoek was naar twee promovendi was de keuze snel gemaakt. Ik werd onderzoeker!

Een uitstekende keuze, kan ik nu zeggen. En dat komt niet in de laatste plaats door de ontzettend fijne begeleiding van Emiel en Marc. Een complementair duo, kan ik melden. "Gaan ze ook samen op vakantie?", vroeg iemand eens. Volgens mij niet, al zou het wel grappig zijn: Emiel die het tentdoek vastmaakt omdat Marc er *net* niet bij kan. Let wel: dit had ik uiteraard niet durven schrijven als Marc zelf niet zo'n fan was geweest van grapjes over lichaamslengte. Maar goed, ik dwaal af.

Emiel, om met jou te beginnen, ik wil je bedanken voor alle goede hulp die je me de afgelopen jaren hebt geboden. Vanaf dag 1 heb ik ons contact als zeer prettig ervaren. Altijd professioneel en doelgericht, maar ook met veel ruimte voor een persoonlijke noot. Niet alleen bewonder ik je om het feit dat je altijd in staat bent om inhoudelijk de juiste twist te geven aan wetenschappelijk teksten (en dus ook aan die van mij), je bent ook nog eens altijd bereikbaar voor welke vraag dan ook. Meer kan een promovendus zich niet wensen. Ik waardeer ook ons contact buiten werktijd: de treinritjes richting Eindhoven iedere donderdag, en het feit dat je kwam kijken toen ik een album presenteerde met mijn band (en dat je toen ook een exemplaar kocht). Sowieso hebben we samen allerlei bands van dichtbij horen spelen, samen met je zoon Daan en soms ook met andere collega's (Joost! Janneke!). Daar kunnen bijvoorbeeld Deerhunter, Alamo Race Track, Wilco, Spinvis en Pavement over meepraten. Opdat er nog vele concerten mogen volgen!

Marc, ik weet niet of je ooit *De Wereld Draait Door* op de Nederlandse TV kijkt, maar als je dit programma kent, dan ken je ook het fenomeen *tafelheer*. Jouw rol in mijn project viel het best te omschrijven als die van tafelheer: altijd stond je klaar om bij te springen met creatieve inzichten en ideeën, altijd was je op zoek naar de invalshoek waar eerder nog niemand aan gedacht had, en altijd stelde je de juiste vragen om een probleemstelling helder te krijgen. Dat vind ik knap, en ik ben dankbaar voor je inbreng. Als een echte tafelheer bracht jij ook humor: de grappen over de Braziliaanse voetballer Kaká waren niet van de lucht. Die schijnt een grappige naam te hebben, al laat ik de invulling van je grappen graag aan de verbeelding van de lezer over.

Wat ik als groot voordeel heb ervaren, is dat ik heb mogen functioneren binnen het VICI-project. We hadden op tweewekelijkse basis bijeenkomsten met alle projectleden om elkaars werk te bespreken, vaak onder het genot van een zak heerlijke Maltesers. Deze bijeenkomsten heetten simpelweg 'VICI-meetings', al werden ze in de wandelgangen (lees: de kamers rondom D401) om onbegrijpelijke redenen al snel omgedoopt tot 'haha-meetings'. Deze meetings hebben mij niet alleen inhoudelijk vooruit geholpen, maar ook nieuwe inzichten verschaft over het 'academische wereldje'. Zo is mij bijvoorbeeld verteld dat als je een paper indient bij een conferentie, je net zo blij moet zijn met een poster als met een praatje. Of zelfs blijer. Behalve dan wanneer je Martijn Goudbeek heet, want dan raak je je posters altijd kwijt in treinen en bussen.

De naam Martijn Goudbeek brengt me op de leden van het VICI-project; met deze collega's heb ik de afgelopen jaren het intensiefst samengewerkt. Ten eerste uiteraard met Martijn Goudbeek zelf. Martijn, bedankt dat je er bijna eigenhandig voor hebt gezorgd dat de VICI-meetings uitgroeiden tot haha-meetings (hoewel we de rol van Marc ook niet mogen uitvlakken), en voor het feit dat ik je tijdens een gastcollege mocht omschrijven als "De man met het kale(nde) hoofd". Overigens: inhoudelijk vind ik je stiekem ook heel goed! Albert and Jette: thank you for being my 'REG algorithmic' conscience; you really helped me a lot in positioning my research into the computational linguistic literature. En tot slot, Marieke, jij was de andere promovendus in het VICI-project, en tevens mijn kamergenootje en (ex-)stadgenoot. Dat was allemaal erg fijn: dankzij jou weet ik nu dat sprekers gebaren kunnen maken met "één komma drie hand". Je moet me toch nog eens leren hoe dat moet. Dank je wel voor je dagelijkse gezelligheid in D404!

Dan zijn er natuurlijk nog heel veel andere collega's die ik hier even wil noemen, omdat ze het leven in gebouw D altijd zoveel aangenamer hebben gemaakt. Allereerst natuurlijk mijn mede-aio's: Mandy (omdat je één van de grootste Amsterdamse fans van mijn bandje bent), Lisanne (omdat je altijd zo vrolijk over de gang huppelt), Constantijn ("Ecuador!"), Jorrig (omdat de Malle-band zonder jou een zinkend schip zou zijn), Martijn B. (omdat je het woord 'frikadellenvijverke' permanent in je vocabulaire hebt opgenomen), Ruud M. (omdat je een geweldige brug tussen de 3e en 4e verdieping bent), en Lisette (omdat je altijd *bijna* een huis koopt, maar intussen wel erg goed onderzoek doet). En dan zijn er natuurlijk nog Adriana, Phoebe, Karin, Lieke, Bart, Emmelyn, Karin, Suleman, Alain, Sander, en iedereen die ik vergeet... Jullie zijn fijne mensen! En tot slot: Hans. Met jou als paranimf kan het niet meer misgaan, daar ben ik van overtuigd. Na onze avonturen op Lowlands, bij Radiohead in Berlijn en op een Belgische 90's party zullen we nu de Aula tot ons grondgebied gaan verklaren. Fijn dat je me daarbij wilt helpen!

Naast de aio's zijn er ook nog andere collega's die ik niet ongenoemd wil laten. Ten eerste zijn dat Lauraine en Jacintha, die tezamen het kloppende hart vormen van de 4e verdieping. Verder nog Joost (hij heeft nou eenmaal een goede muzieksmaak), Maria (echte Ajax-fans moet je koesteren), Anja (zelden zo'n oprecht enthousiast en aardig persoon meegemaakt), Janneke (ik slaap niet in een Alamo Race Track-pyjama, echt niet!), Per (omdat hij me altijd van alles vraagt over z'n MacBook, en ik daar nooit een antwoord op weet), Rein (omdat hij me ooit kwalificeerde als een Beatle, maar nu zelf langere haren heeft dan ik), Harold (de enige echte huisfotograaf), Fons (de sympathieke Belgische leider), Carel (het statistische geweten van gebouw D), en Leonoor (samen met jou scripties begeleiden is leuk!). Ook hier geldt: het lijstje is vast niet compleet.

Het leven is niet alleen mooi binnen de muren van de universiteit, maar ook erbuiten. Ik heb heel veel mensen om me heen die daaraan bijdragen, en die ik graag even de revue wil laten passeren. Allereerst Pauwke: dank voor je jarenlange fijne vriendschap, je vaderlijke adviezen op het gebied van de wetenschap, je puberale humor, en natuurlijk niet te vergeten je steun als paranimf. Verder noem ik Mira, Sebastiaan, Akko, Jeroen, Wouter en Jona: met jullie muziek maken is iedere keer weer een genot. Dan zijn er nog Bart, Koen, Clarck, Simone, Gertjan, Pieter, Teun, Sofie, Dorus, Yinga (a.k.a. 'brown sugar'), en Iris uiteraard: weet dat ik onze vriendschappen zeer waardeer. Opdat er nog maar veel festivals, etentjes en concertjes mogen volgen!

Ook wil ik via deze weg graag mijn waardering uitspreken voor Joop en Harriët. Vanaf het eerste moment dat ik bij jullie kwam heb ik me welkom gevoeld, met als ultieme bewijs dat ik me inmiddels gekleed in een geleende beertjespyjama aan het ontbijt durf te melden. Heel erg bedankt voor jullie steun en interesse tijdens de afgelopen jaren.

Gijs, aan jou wil ik de moeilijkste alinea van dit hele proefschrift wijden. Het is krankzinnig wat er vorig jaar met je gebeurd is, en dat je niet meer bij ons bent. We missen je ontzettend.

Inge, grote zus die eigenlijk kleiner is, ik wil jou bedanken voor de leuke en gezellige band die we samen hebben. Dankzij jou - en Twan natuurlijk! - kan ik als Ome Ruudie tegenwoordig al mijn partytricks uit de kast halen. Dat voelt als een verrijking! Papa en mama, voor jullie heb ik een cliché in petto, maar het mooie aan clichés is dat ze altijd zo heel erg ontzettend waar zijn. Ik wil jullie bedanken voor alles wat jullie voor me hebben gedaan, voor de kansen die jullie me hebben gegeven om de dingen te ontdekken en doen die ik leuk vind, en voor het feit dat jullie er altijd voor me zijn. En natuurlijk ook voor het feit dat ik vroeger altijd zo veel goede muziek (dan wel herrie) mocht maken op mijn drumstel!

Tot slot wil ik nog mijn allerliefste woorden richten tot Anouk. Vanaf het eerste moment dat je me vroeg of ik je vriendje wilde zijn, voel ik me daar heel erg blij en gelukkig over. Wat een geluk dat ik destijds “ja” heb gezegd... Dank je wel voor al je steun, en voor de vele fijne momenten en knuffels. Ik vind jou echt heel lief!

Publication list

Journal papers

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, 37 (2), 395-411.

Krahmer, E., Koolen, R., & Theune, M. (2012). Is it that difficult to find a good preference order for the Incremental Algorithm? *Cognitive Science*, 36 (5), 837-841.

Koolen, R., Gatt, A., Goudbeek, M., Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics* 43 (13), 3231-3250.

Working papers

Koolen, R., Krahmer, E., & Swerts, M. (submitted). How distractor objects affect the redundant use of color in definite reference: Effects of bottom-up and top-down saliency cues.

Koolen, R., Krahmer, E., & Swerts, M. (submitted). Developmental changes in children's processing of redundant information in definite object descriptions.

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (submitted). Reduction in speech and gestures in repeated references.

Papers in conference proceedings (peer-reviewed)

Koolen, R., Krahmer, E., & Swerts, M. (2013). The impact of bottom-up and top-down saliency cues on reference production. In *Proceedings of the 35th annual meeting of the Cognitive Science Society (CogSci)*. Berlin, Germany.

Krahmer, E., Noordewier, M., Goudbeek, M., & Koolen, R. (2013). How big is the BFG? The impact of redundant size adjectives on size perception. In *Proceedings of the 35th annual meeting of the Cognitive Science Society (CogSci)*. Berlin, Germany.

Westerbeek, H., Koolen, R., & Maes, A. (2013). Color typicality and content planning in definite reference. In *Proceedings of the workshop on the Production of Referring Expressions: bridging the gap between cognitive* *and computational approaches to reference (PRE-CogSci 2013).* Berlin, Germany.

Koolen, R., Krahmer, E., Theune, M. (2012). Learning preferences for Referring Expression Generation: effects of domain, language and algorithm. In *Proceedings of the 8th International conference on Natural Language Generation (INLG)*. Chicago, USA.

Koolen, R., Goudbeek, M. & Krahmer, E. (2011). Effects of scene variation on referential overspecification. In *Proceedings of the 33rd annual meeting of the Cognitive Science Society (CogSci).* Boston, USA.

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E. & Swerts, M. (2011). GREEBLES, Greeble, greeb: on reduction in speech and gesture in repeated reference. In *Proceedings of the 33rd annual conference of the Cognitive Science Society (CogSci). Boston, USA.*

Theune, M., Koolen, R., Krahmer, E., & Wubben, S. (2011). Does size matter – how much data is required to train a REG algorithm? In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics (ACL).* Portland, Oregon, USA.

Theune, M., Koolen, R. & Krahmer, E. (2010). Cross-linguistic attribute selection for REG: comparing Dutch and English. In *Proceedings of the 6th International Natural Language Generation Conference (INLG 2010).* Dublin, Ireland.

Koolen, R. & Krahmer, E. (2010). The D-TUNA corpus: a Dutch dataset for the evaluation of REG algorithms. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010).* Valletta, Malta.

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2009). Need I say more? On factors causing referential overspecification. In *Proceedings of the workshop on the Production of Referring Expressions: bridging the gap between computational and empirical approaches to reference (PRE-CogSci 2009).* Amsterdam, The Netherlands.

Abstracts of conference presentations (peer reviewed)

Koolen, R., Krahmer, E. & Swerts, M. (2011). On children's perception of

overspecification in referring expressions. Poster presented at the biennial meeting of Experimental Pragmatics (x-prag). Barcelona, Spain, May 2011.

Hoetjes, M., Schmit, A., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2011). Talk presented at the 9th International Gesture Workshop 2011. National and Kapodistrian University of Athens, May 2011.

TiCC Ph.D. series

1. Pashiera Barkhuysen. *Audiovisual Prosody in Interaction.* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 3 October 2008.
2. Ben Torben-Nielsen. *Dendritic Morphology: Function Shapes Structure.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotor: K.P. Tuyls. Tilburg, 3 December 2008.
3. Hans Stol. *A Framework for Evidence-based Policy Making Using IT.* Promotor: H.J. van den Herik. Tilburg, 21 January 2009.
4. Jeroen Geertzen. *Dialogue Act Recognition and Prediction.* Promotor: H. Bunt. Co-promotor: J.M.B. Terken. Tilburg, 11 February 2009.
5. Sander Canisius. *Structured Prediction for Natural Language Processing.* Promotores: A.P.J. van den Bosch, W. Daelemans. Tilburg, 13 February 2009.
6. Fritz Reul. *New Architectures in Computer Chess.* Promotor: H.J. van den Herik. Co-promotor: J.W.H.M. Uiterwijk. Tilburg, 17 June 2009.
7. Laurens van der Maaten. *Feature Extraction from Visual Data.* Promotores: E.O. Postma, H.J. van den Herik. Co-promotor: A.G. Lange. Tilburg, 23 June 2009 (cum laude).
8. Stephan Raaijmakers. *Multinomial Language Learning.* Promotores: W. Daelemans, A.P.J. van den Bosch. Tilburg, 1 December 2009.
9. Igor Berezhnoy. *Digital Analysis of Paintings.* Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 7 December 2009.
10. Toine Bogers. *Recommender Systems for Social Bookmarking.* Promotor: A.P.J. van den Bosch. Tilburg, 8 December 2009.
11. Sander Bakkes. *Rapid Adaptation of Video Game AI.* Promotor: H.J. van den Herik. Co-promotor: P. Spronck. Tilburg, 3 March 2010.
12. Maria Mos. *Complex Lexical Items.* Promotor: A.P.J. van den Bosch. Co-promotores: A. Vermeer, A. Backus. Tilburg, 12 May 2010 (in collaboration with the Department of Language and Culture Studies).
13. Marieke van Erp. *Accessing Natural History. Discoveries in data cleaning, structuring, and retrieval*. Promotor: A.P.J. van den Bosch. Co-promotor: P.K. Lendvai. Tilburg, 30 June 2010.

14. Edwin Commandeur. *Implicit Causality and Implicit Consequentiality in Language Comprehension*. Promotores: L.G.M. Noordman, W. Vonk. Co-promotor: R. Cozijn. Tilburg, 30 June 2010.
15. Bart Bogaert. *Cloud Content Contention*. Promotores: H.J. van den Herik, E.O. Postma. Tilburg, 30 March 2011.
16. Xiaoyu Mao. *Airport under Control.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotores: N. Roos, A. Salden. Tilburg, 25 May 2011.
17. Olga Petukhova. *Multidimensional Dialogue Modelling*. Promotor: H. Bunt. Tilburg, 1 September 2011.
18. Lisette Mol. *Language in the hands.* Promotores: E.J. Krahmer, F. Maes, M.G.J. Swerts. Tilburg, 7 November 2011 (*cum laude*).
19. Herman Stehouwer. *Statistical Language Models for Alternative Sequence Selection*. Promotores: A.P.J. van den Bosch, H.J. van den Herik. Co-promotor: M.M. van Zaanen. Tilburg, 7 December 2011.
20. Terry Kakeeto-Aelen. *Relationship Marketing for SMEs in Uganda*. Promotores: J. Chr. van Dalen, H.J. van den Herik. Co-promotor: B.A. Van de Walle. Tilburg, 1 February 2012.
21. Suleman Shahid. *Fun & Face: Exploring non-verbal expressions of emotion during playful interactions*. Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 25 May 2012.
22. Thijs Vis. *Intelligence, Politie en Veiligheidsdienst: Verenigbare Grootheden?* Promotores: T.A. de Roos, H.J. van den Herik, A.C.M. Spapens. Tilburg, 6 June 2012 (in collaboration with the Tilburg School of Law).
23. Nancy Pascall. *Engendering Technology Empowering Women.* Promotores: H.J. van den Herik, M. Diocaretz. Tilburg, 19 November, 2012.
24. Agus Gunawan. *Information Access for SMEs in Indonesia*. Promotor: H.J. van den Herik. Co-promotores: M. Wahdan, B.A. Van de Walle. Tilburg, 19 December 2012.

25. Giel van Lankveld. *Quantifying Individual Player Differences.* Promotores: H.J. van den Herik, A.R. Arntz. Co-promotor: P. Spronck. Tilburg, 27 February 2013.
26. Sander Wubben. *Text-to-text Generation Using Monolingual Machine Translation.* Promotores: E.J. Krahmer, A.P.J. van den Bosch, H. Bunt. Tilburg, 5 June 2013.
27. Jeroen Janssens. *Outlier Selection and One-Class Classification*. Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 11 June 2013.
28. Martijn Balsters. *Expression and Perception of Emotions: The Case of Depression, Sadness and Fear*. Promotores: E.J. Krahmer, M.G.J. Swerts, A.J.J.M. Vingerhoets. Tilburg, 25 June 2013.
29. Lisanne van Weelden. *Metaphor in Good Shape*. Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 28 June 2013.
30. Ruud Koolen. *Need I Say More? On Overspecification in Definite Reference.* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 20 September 2013.