TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science (2012) 1–21

ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2012.01186.x

Alignment in Interactive Reference Production: Content Planning, Modifier Ordering, and Referential Overspecification

Martijn Goudbeek, Emiel Krahmer

Tilburg Centre for Cognition and Communication (TiCC), Tilburg University

Received 13 March 2010; received in revised form 28 January 2011; accepted 22 April 2011

Abstract

Psycholinguistic studies often look at the production of referring expressions in interactive settings, but so far few referring expression generation algorithms have been developed that are sensitive to earlier references in an interaction. Rather, such algorithms tend to rely on domain-dependent preferences for both content selection and linguistic realization. We present three experiments showing that humans may opt for dispreferred attributes and dispreferred modifier orderings when these were primed in a preceding interaction (without speakers being consciously aware of this). In addition, we show that speakers are more likely to produce overspecified references, including dispreferred attributes (although minimal descriptions with preferred attributes would suffice), when these were similarly primed.

Keywords: Referring expressions; Algorithms; Alignment; Content determination; Modifier orderings; Overspecification

1. Introduction

Referring expressions (*the large red chair*, *the man with the beard*) play an important role in communication, and hence it is not surprising that their production has been studied extensively both from a computational linguistic and from an empirical, psycholinguistic perspective. One striking difference between these two perspectives is that computational linguists often concentrate on developing algorithms for the production of referring

Correspondence should be sent to Martijn Goudbeek, Tilburg Centre for Cognition and Communication (TiCC), School of Humanities, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands.
E-mail: m.b.goudbeek@uvt.nl

expressions in isolation, while psycholinguists mostly study reference production in interaction. An important finding of the psycholinguistic studies is that speakers tend to adapt to each other. If one speaker uses, say, a certain lexical item or syntactic construction, the other speaker is more likely to use that representation as well, as a result making the interaction easier for both of them. It therefore seems that one way to bridge the gap between empirical and computational approaches to reference would be to apply algorithms to reference production in interaction, and extending them in such a way that algorithms can adapt to their communication partners in a natural way.

This is not straightforward, however: it is not obvious whether existing algorithms can be extended in this way, and moreover, although many psycholinguistic studies have concentrated on alignment and adaptation during interaction, empirical data are lacking for a number of important ingredients of reference production. This article studies reference production using an interactive reference production paradigm and collecting new data for three important aspects of reference production: content planning, modifier ordering, and referential overspecification.

1.1. Computational approaches to referring expression generation

The computational production (or ''generation'') of referring expressions has been studied primarily in the context of natural language generation, the process of automatically converting non-linguistic information (e.g., from a database) into coherent natural language text. This is useful for many practical applications, ranging from automatically generated weather forecasts to summarizing medical information in a patient-friendly way (Reiter & Dale, 2000). Most generation systems have a separate component dedicated to the generation of referring expressions (Mellish et al., 2006; Reiter & Dale, 2000). Although the details may differ, these systems often approach referring expression generation (REG) as a two-step procedure, where first it is decided which properties to include in a referring expression (content planning), after which the selected properties are turned into a natural language expression (linguistic realization). It is interesting to observe that psycholinguistic models of speech production make a similar distinction; in Levelt's blueprint for the speaker, for example, ''deciding what to say'' is done by the Conceptualizer and ''deciding how to say it'' by the Formulator (Levelt, 1989, 1999). The basic problem that algorithms need to solve in both stages is one of choice; there are potentially many ways in which one could refer to some object (a chair may be referred to as *large or red or seen from the front*, or some combination of these and other properties), and once a set of properties has been selected, there are potentially multiple ways in which they can be expressed in natural language (*the large red chair*, *the red large chair*, *the red chair that is also large*, . . .).

Typically, these choice problems are tackled by assuming that some solutions are more preferred than others. Consider, for example, the Incremental Algorithm, due to Dale and Reiter (1995) and one of the most widely used algorithms for determining the contents of referring expressions. It rests on the assumption that some attributes are preferred over others (partly based on evidence provided by Pechmann, 1989); a chair would first be described in terms of its color, and only if this does not succeed in uniquely characterizing the target

object, other attributes such as size and orientation are tried, until a description has been found which distinguishes the target from its distractors. In the Incremental Algorithm, this is modeled by defining a complete ordering of Preferred Attributes (color, size, orientation, etc.), and letting the algorithm iterate through this list, selecting the relevant value (e.g., red) of an attribute (e.g., color) if it helps to distinguish the target from one of the distractor objects (in other words, if some of the other chairs are not red). This iteration continues until all distractors have been ruled out, and a set of attribute–value pairs is collected (e.g., (type, chair), (color, red), (size, large)) that can be turned into natural language by a linguistic realization algorithm (*the large red chair*).[1]

An interesting aspect of the Incremental Algorithm is that once an attribute and value are selected, it will always be included in the final generated referring expression although a later attribute–value pair may render it redundant. This is a direct consequence of the incremental nature of speech production that the algorithm emulates, and in computational terms it is one of the properties that makes the Incremental Algorithm computationally attractive (it never engages in backtracking). In this way, the Incremental Algorithm offers a computational model of overspecification, the phenomenon where speakers include information in referring expressions that is not strictly speaking necessary for identification of the target. Various psycholinguistic studies have confirmed that human speakers frequently overspecify their references (e.g., Arts, 2004; Engelhardt, Bailey, & Ferreira, 2006), but the Incremental Algorithm makes specific predictions about overspecification, related to preferred and dispreferred attributes. In particular, the algorithm predicts that redundant attributes will typically be preferred ones: The algorithm may produce an overspecified expression such as *the large red chair* when the *the large chair* would be sufficient to uniquely characterize the target, but it would never produce *the large red chair* when *the red chair* would be distinguishing (after all, the algorithm stops as soon as a distinguishing description has been found; hence, less preferred attributes need not be tried anymore).

Finding the correct preference order is crucial; evaluation studies have revealed that the Incremental Algorithm may perform well with one order, and poorly with another (Gatt, van der Sluis, & van Deemter, 2007; van der Sluis, Gatt, & van Deemter, 2007). So how do we determine which attributes are preferred over others? Psycholinguistic studies (e.g., Belke & Meyer, 2002; Pechmann, 1989) give some general guidance, suggesting for instance that absolute attributes (of which color is an example) are generally preferred over relative ones (such as size). Pechmann explains this by pointing out that a speaker only has to look at the target itself to see what color it is, whereas the speaker has to look at the distractors before it can be determined whether the target is large or small. However, such deliberations are insufficient to completely order all relevant attributes, and Dale and Reiter (1995) observe that the preference order ''will vary with the domain, and will typically be determined by empirical investigation.''

It is important to stress, that while the Incremental Algorithm is arguably unique in assuming a complete preference order of attributes, most other REG algorithms also rely on distinguishing preferred and less preferred solutions. The graph-based algorithm (Krahmer, van Erk, & Verleg, 2003), for example, searches for the cheapest description for a target,

and it makes a distinction between cheap attributes (such as color) and more expensive ones (such as orientation).

Linguistic realization of referring expressions has traditionally received less attention than content planning, and often it is assumed that this task can be performed by a separate realization algorithm, in a second phase (but see, e.g., Horacek, 1997; Krahmer & Theune, 2002; Stone & Webber, 1998, who argue for a tighter coupling between the two). An important complication for realizing referring expressions is that selected attribute–value pairs will often be realized as linguistic, prenominal modifiers, and ordering these may be difficult. In recent years, this problem has been addressed in various studies (Malouf, 2000; Mitchell, 2010; Shaw & Hatzivassiloglou, 1999). These all assume that some orderings (e.g., *large red*) are preferred over others (*red large*) and that finding the correct ordering is an empirical matter, which can be addressed by looking at large data sets and the relative orderings of modifiers encountered in them.

1.2. Adaptation and alignment in reference production

In the previous paragraphs, discussing computational approaches to reference and referential preferences, the role of the addressee was not mentioned. This is not a coincidence, because the majority of current computational REG research ignores the role of the addressee. This obviously limits the applicability of existing REG algorithms in the context of dialog applications, as was shown by Gupta and Stent (2005) and Jordan and Walker (2005). More recently, various researchers have started exploring the generation of referring expressions in interactive settings, such as Stoia, Byron, Shockley, and Fosler-Lussier (2006) and Janarthanam and Lemon (2009). The latter, for example, present an algorithm that adapts its referring expression to the expertise of the user, referring to the same device as *the router* for an expert and as *the black box with lights* for a novice user. Various researchers also started developing computational models of linguistic realization that can handle lexical and syntactic alignment (Buschmeier, Bergmann, & StefanKopp, 2010; Reitter, 2008; Reitter, Moore, & Keller, 2006). However, so far the impact of this kind of work on REG in general has been minimal.

In contrast to computational approaches to reference, psycholinguistic studies have often looked at reference production in interaction (Clark & Bangerter, 2004). Clark and Murphy (1983), to give one example, describe an experiment using the director–matcher paradigm, in which one participant (the director) refers to figures which the other participant (the matcher) has to order based on the director's references. The task was to match 12 complex, difficult-to-describe tangram figures, and to do this six times. Clark and Murphy found that the number of words and the number of turns needed per figure before identification decreased as a function of trial, and they argued that this is because the speaker and addressee tend to converge on a shared and agreed-upon set of referring expressions. Brennan and Clark (1996) argue this kind of lexical entrainment (Garrod & Anderson, 1987) in converging descriptions derives from ''conceptual pacts,'' in which speaker and addressee have formed an implicit agreement on how to refer to an object. Brennan and Clark found that when speakers have to refer to a target in the context of a distractor from

the same basic-level category (e.g., a shoe), they consistently use less preferred, more specific expressions (such as *the sneaker*). Interestingly, they continue relying on such pacts with the same addressee, even when later on *the shoe* would be perfectly distinguishing, but not when they start interacting with a different addressee (Brennan & Clark, 1996; Brown-Schmidt, 2009; Metzing & Brennan, 2003).

A more general account of adaptation in dialog is the Interactive Alignment Model (Garrod & Pickering, 2004; Pickering & Garrod, 2004). Garrod and Pickering argue that conversation is easy, because speakers align their linguistic representations. They may do so on all levels of interaction, ranging from alignment of phonetic categories up to syntactic alignment, where alignment at lower levels is argued to facilitate alignment at higher levels. A difference with the work of Brennan, Clark, and colleagues is that alignment is not assumed to be a collaborative process but is seen as the result of a mechanistic, largely automatic process where speakers produce expressions that are easy to comprehend for their addressee because they rely on representations that were used earlier on in the interaction; alignment of representations is a consequence of priming them (e.g., Chartrand & Bargh, 1999; Dijksterhuis & Bargh, 2001).

Before referring expression generation algorithms can be applied in interactive settings (as would be required, for instance, for applications such as spoken dialog systems or embodied conversational agents), they should be able to take their addressee into account. Explicit reasoning about the knowledge state of the addressee (as Heeman & Hirst, 1995, argue for) is hard from a computational point of view, and it may also not be what humans do (Horton & Keysar, 1996; Keysar, Lin, & Barr, 2003; Lane, Groisman, & Ferreira, 2006). However, some amount of sensitivity to the referring expressions that were produced in the prior interaction seems a sine qua non for interactive applications. So far, few attempts have been made to develop algorithms that have this kind of sensitivity, partly because data are lacking for various aspects of the generation process. In this article, we describe three experiments, filling in some of these blanks.

1.3. The current experiments

''Deciding what to say'' and ''deciding how to say it'' can be seen as two consecutive choice problems. Most REG algorithms solve these choice problems by assuming that some solutions are preferred over others. Psycholinguistic studies of reference, on the other hand, suggest that speakers adapt to each other during communication, aligning representations on all levels of communication. So what do we do when REG algorithms are applied in interactive settings? Should they stick to domain dependent preferences, or should they align with referring expressions that were produced earlier in the interaction? In addition, would the domain in which they are referring matter, that is, would it matter whether they are referring to furniture items, or to persons, where preferences for certain attributes may be stronger or weaker? These questions are addressed using a new interactive reference production paradigm, where participants are asked to refer to one target object (a piece of furniture or a person) in the context of two distractors from the same domain. In earlier interactions, participants have been unknowingly exposed to either preferred or dispreferred alternatives,

which allows us to see to what extent participants employ preferred expressions or align with dispreferred ones. This paradigm is somewhat reminiscent of methods used in earlier work on syntactic priming (e.g., Bock, 1986; Bock & Griffin, 2000; Bock, Dell, Chang, & Onishi, 2007), but its interactive nature mimics dialog more closely and the fact that attributes rather than syntactic structures can be primed makes it particularly relevant for REG research. The Furniture and People domains have been used in earlier data gathering efforts (Gatt et al., 2007; Koolen, Gatt, Goudbeek, & Krahmer, 2009), which allows us to determine beforehand what are the preferred and dispreferred ways of referring to targets in these domains.

Experiment I is devoted to content determination, and studies what speakers do when referring to a target that can be distinguished in a preferred (*the red chair*) or a dispreferred way (*the left-facing chair*), when in the prior context either the preferred or the dispreferred variant was primed. Experiment II uses the same experimental paradigm but applies it to linguistic realization, and more specifically, to modifier orderings. Here, in the critical trials, participants have to refer to a target using two attributes (e.g., color and size), while in the preceding context either preferred modifier orderings (*large red*) or dispreferred ones (*red large*) were primed. In Experiment III, finally, overspecification is studied, where participants are once more asked to refer to a target, which can be distinguished using a minimal referring expression, only including a preferred attribute (*the red chair*), while earlier overspecified references (*the red large chair*) were primed. In Section 5 of this article, we compare the outcomes of these experiments with the predictions of current generation algorithms, such as the Incremental Algorithm, and discuss ways in which these can be made more ''psychologically plausible'' and suitable for interactive applications.

2. Experiment I

Experiment I studies what speakers do when referring to a target that can be distinguished in a preferred (*the blue fan*) or a dispreferred (*the left-facing fan*) way, when in the preceding context either the first or the second variant was primed.[2]

2.1. Method

2.1.1. Participants

The participants ($n = 26$) were all students of the Radboud University Nijmegen (two males, mean age = 20 years 11 months). All were native speakers of Dutch and did not report any hearing or speech problems. They participated in the experiment in exchange for course credits.

2.1.2. Materials

The stimulus pictures were taken from the TUNA corpus (Gatt et al., 2007) that has been extensively used for the evaluation of REG algorithms. This corpus consists of two domains:

one containing pictures of people (all famous mathematicians), the other containing furniture items in different colors depicted from different orientations[3] (e.g., *The red left-facing desk*). From previous studies (Gatt et al., 2007; Koolen et al., 2009), it is known that participants show a preference for (i.e., more frequently use) certain attributes (color in the Furniture domain and wearing glasses in the People domain) and disprefer other attributes (orientation of a furniture piece and wearing a tie, respectively).

2.1.3. Procedure

Trials consisted of four turns in an interactive reference understanding and production experiment: a prime, two fillers, and the experimental description (see Fig. 1 for an overview of experimental trials). First, participants listened to a pre-recorded female voice referring to one of three objects and had to indicate which one was being referred to. In this subtask, references either used a preferred or a dispreferred attribute; both were distinguishing. Second, participants themselves described a filler picture, after which, third, they had to indicate which filler picture was being referred to (as in the first subtask). The two filler turns always concerned stimuli from the alternative domain and were intended to prevent a

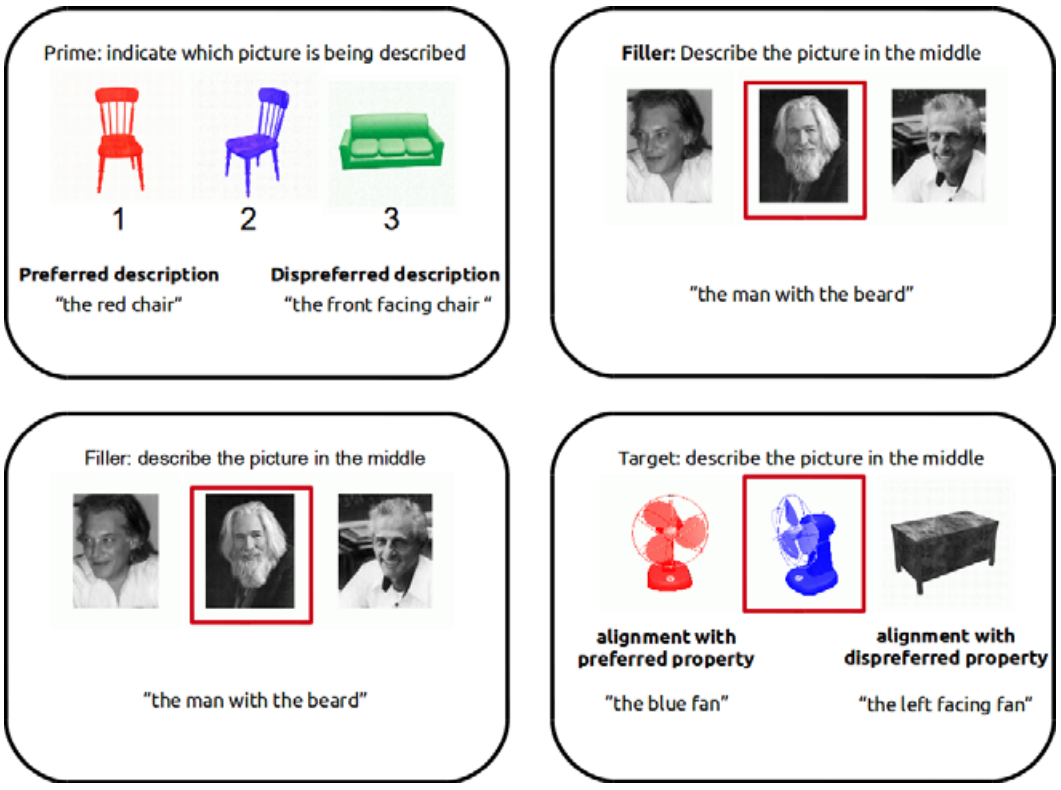

Fig. 1. The four tasks that constitute a trial. This figure shows a furniture trial; people trials have an identical structure.

too direct connection between the prime and the target. Fourth, participants described the target object, which could always be distinguished from its distractors in a preferred (*The blue fan*) or a dispreferred (*The left-facing fan*) way. Note that *attributes* are primed, not values; a participant may have heard *front-facing* in the prime turn, while the target has a different value for the orientation attribute (cf. Fig. 1). In addition, in the Furniture domain, the types could also differ; when primed with a (preferred or dispreferred) description of a chair, participants not necessarily had to describe a chair in the critical trial.

For the two domains, there were 20 preferred and 20 dispreferred trials, resulting in $2 \times (20 + 20) = 80$ critical trials. These were presented in counter-balanced blocks, and within blocks each participant received a different random order. In addition, there were 80 filler trials (each following the same structure as outlined in Fig. 1); filler trials never involved the attributes of interest. During debriefing, none of the participants indicated they had been aware of the experiment's true purpose.

2.2. Results and discussion

The proportions of preferred and dispreferred attributes as a function of prime and domain are shown in Figs. 2 and 3. The black bars indicate use of the preferred attribute and the white bars indicate use of the dispreferred attribute. In both domains, the preferred attribute is used more frequently than the dispreferred attribute with the preferred primes, which serves as a manipulation check (our participants indeed overall preferred the preferred attributes to the dispreferred ones).

Fig. 2 shows a clear effect of prime for the Furniture domain: Participants use the preferred attribute (color, as in *the red fan*) more when they were primed with this attribute. Conversely, when they were primed with the dispreferred attribute orientation (as in *the fan seen from the front*), they used this attribute more often in the critical trials. Fig. 3 reveals a

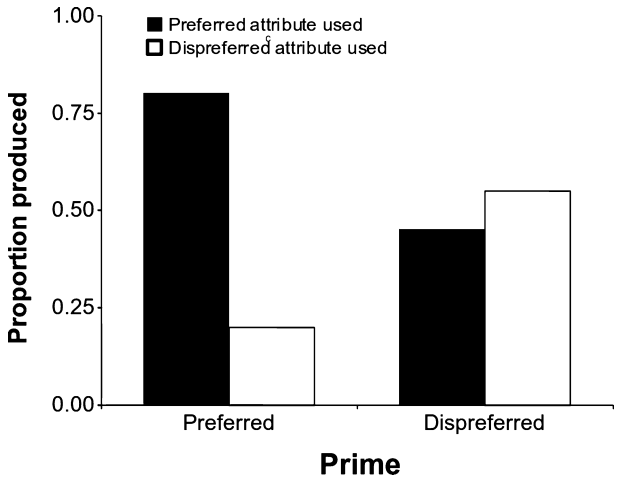

Fig. 2. Proportions of preferred and dispreferred attributes in the Furniture domain.

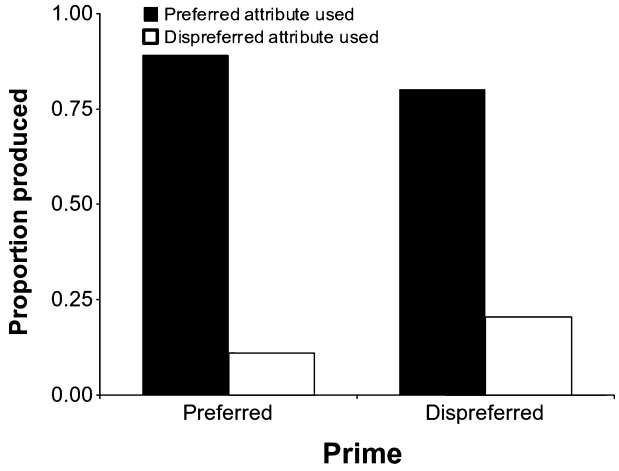

Fig. 3. Proportions of preferred and dispreferred attributes in the People domain.

similar picture for the People domain (when speakers were primed with the dispreferred attribute, they use it themselves more often), but clearly much less pronounced.

For our statistical analysis, we use the proportion of attribute alignment as our dependent measure. Alignment occurs when a participant uses the same attribute in the target as occurred in the prime. Our definition of alignment consequently includes overspecified descriptions (Arnold, 2008; Engelhardt et al., 2006), where both the preferred and dispreferred attributes were mentioned by participants (e.g., *the red front-facing chair* aligns with *the red chair* and with *the front-facing chair*). On average, overspecification occurred in 13% of the critical trials (and these were evenly distributed over the experimental conditions, see also Table 5). Table 1 displays the mean alignment and standard deviations per prime (preferred vs. dispreferred) for the Furniture and People domain.

We conducted a 2×2 repeated measures analysis of variance with Alignment as the dependent variable and Domain (Furniture vs. People) and Prime (Preferred vs. Dispreferred) as independent variables. The results of this analysis are shown in Table 2. A significant main effect was found of Prime, showing that the prime influenced the selection of the attributes in the critical trial: when primed with dispreferred attributes, our participants used the dispreferred attributes more often than when they were primed with

Table 1
Mean alignment (and standard deviations) as a function of Domain (Furniture and People) and Prime (Preferred and Dispreferred)

Domain	Prime	Mean Alignment (*SD*)	
		Experiment I	Experiment II
Furniture	Preferred	0.89 (0.32)	0.84 (0.13)
	Dispreferred	0.60 (0.49)	0.22 (0.18)
People	Preferred	0.97 (0.16)	0.60 (0.33)
	Disperred	0.25 (0.43)	0.54 (0.33)

Table 2
Summary of within subjects analysis of variance for Experiment I (attribute selection)

	F	*df*	*p* Value	η^2
Domain	10.88	1, 25	.01	0.30**
Prime	6.43	1, 25	.02	0.21*
Domain × Prime	5.74	1, 25	.02	0.19*

Note. Significant at *$p < .05$ and **$p < .01$.

preferred attributes. A significant main effect was found of Domain, confirming that there is significantly more alignment in the Furniture domain. Finally, a significant interaction was found, confirming our observation that the effect of the Prime was less pronounced in the People domain.

As a final test of our hypothesis that adaptation processes play an important role in attribute selection for referring expressions, we looked at participants' expressions with the *dispreferred* primes (with the preferred primes, effects of adaptation and of preferences cannot be teased apart). The Incremental Algorithm, and many REG algorithms, predict that the dispreferred attribute will never be used, because the preferred attribute is sufficient to uniquely characterize the target. However, in both the Furniture and the People domain there was significantly more alignment than would be expected based on this prediction, $t_{\text{furniture}}(25) = 6.86, p < .01$; $t_{\text{people}}(25) = 4.81, p < .01$.

3. Experiment II

Experiment II investigates the realization of referring expressions. It uses the same interactive elicitation paradigm as used for Experiment I to study whether speaker's preferences for modifier orderings can be changed by exposing them to dispreferred orderings.

3.1. Method

3.1.1. Participants

The participants ($n = 28$) were all students (10 males, mean age = 23 years 2 months) from Tilburg University who participated in exchange for course credits. All participants were native speakers of Dutch and did not report any hearing or speech problems. None participated in Experiment I.

3.1.2. Materials

The materials were identical to those used in Experiment I, except for their arrangement in the critical trials. In these trials, the target picture could only be distinguished from its competitors using two attributes (besides their type). In the Furniture domain these were the color and size of the furniture items; in the People domain these were having a beard and wearing glasses. Different orderings of these sets of attributes are possible; participants

Table 3
Google counts (in Dutch) of the preferred and dispreferred modifier orderings

	People	Furniture
Preferred realization	177	44,863
Dispreferred realization	67	893

Note. The counts for the People domain are based on the hits for ''bespectacled and bearded'' and ''bearded and bespectacled.'' The counts for the Furniture domain are based on the mean of all relevant combinations of color and size. Counts were made on July 1, 2009.

could describe furniture targets, for instance, as *the big red chair* or *the red big chair*, and people targets as *the bearded and bespectacled man* or *the bespectacled and bearded man*. Note that all these variants are legal in Dutch, but some are more preferred than others. We assessed the preference of the various modifier orderings with Google counts (performed on July 1, 2009). Using the web is an increasingly popular way to compute statistics about word combinations (e.g., Keller & Lapata, 2003, but see Kilgarriff, 2007 for a critical review). We calculated the mean number of hits (on Dutch webpages) returned by Google on bigrams *big red* and *red big* for all colors and sizes, as well as the number of hits for *bespectacled and bearded* versus *bearded and bespectacled*. As can be seen in Table 3, the count ratio of preferred to dispreferred is in the order of 50:1 in the Furniture domain and 3:1 in the People domain. Given that the preferences are less pronounced in the People domain, we predict stronger effects of alignment for this domain.

In the prime turn (Task I, see Fig. 1), attributes were realized in a preferred way (''size first'': e.g., *the big red sofa*, or ''glasses first'': *the bespectacled and bearded man*) or in a dispreferred way (''color first'': *the red big sofa* or ''beard first'': *the bearded and bespectacled man*). The prime descriptions were spoken by the same speaker as in Experiment I.

3.1.3. Procedure

The procedure is otherwise identical to that of Experiment I. Again, none of the participants reported any awareness of the experiment's goal at debriefing.

3.2. Results and discussion

Figs. 4 and 5 show the use of the preferred and dispreferred modifier ordering per prime and domain. It can be seen that the Google counts offer a good indication of the participants' preferences in the Furniture domain, but less so in the People domain. Given that the difference in preference between the two modifier orderings was much smaller in the People domain than in the Furniture domain and that Google counts need not be directly related to speaker preferences, this finding is not entirely unexpected.

The data in Figs. 4 and 5 show that participants' modifier orderings are clearly influenced by the descriptions they were previously exposed to. To statistically assess the observed effects, we again conducted a 2×2 repeated measures analysis of variance with Domain (Furniture vs. People) and Prime (Preferred vs. Dispreferred) as independent variables.

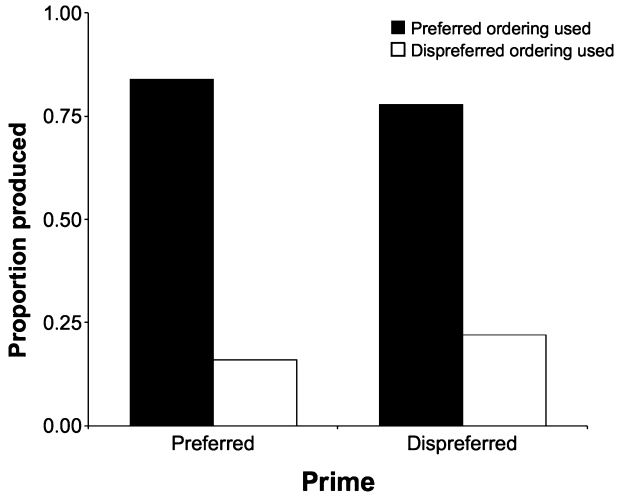

Fig. 4. Proportions of preferred and dispreferred modifier orderings in the Furniture domain.

Table 4
Summary of the within subjects analysis of variance for Experiment II (modifier orderings)

	F	*df*	*p* Value	η^2
Domain	41.03	1, 27	.01	0.60**
Prime	4.51	1, 27	.04	0.14*
Domain × Prime	1.33	1, 27	.26	0.05

Note. Significant at *$p < .05$ and **$p < .01$.

Similarly as in Experiment I, we use the proportion of aligned modifier orderings as our dependent measure, where alignment occurs when the participant's ordering coincides with the primed ordering, irrespective of the exact realization. Note that overpecification with a preferred or dispreferred attribute is not an issue here, since both attributes have to be used to produce a correct, distinguishing description. Table 4 shows the results of the analysis of variance. A significant main effect of Domain was found, reflecting the initial difference in preference between the modifier orderings in the People and Furniture domains. The significant main effect of Prime shows that our participants were once again sensitive to the difference in primes, and more often used the dispreferred modifier orderings when they were primed with these. The interaction between Prime and Domain failed to reach significance, but a separate analysis per domain revealed a significant effect of prime in the People, $F(1, 27) = 3.95$, $p < .05$, $\eta^2 = .13$, but not in the Furniture domain, $F(1, 25) = 1.474$, *ns*. Typically, natural language realization algorithms will opt for the most frequently encountered modifier ordering; hence, these algorithms would not produce any dispreferred orderings. However, *t* tests showed that both in the Furniture, $t(27) = 6.54$, $p < .01$, and in the People, $t(27) = 7.44$, $p < .01$, domain the amount of alignment to dispreferred primes was significantly higher than zero.

4. Experiment III

This experiment looks at overspecification, where participants are once more asked to refer to a target, which could be distinguished using a minimal referring expression only including a preferred attribute, while in the prior context overspecified references were primed, including both preferred and dispreferred attributes.

4.1. Method

4.1.1. Participants

Participants ($n = 28$) were all students (8 males, mean age = 20 years 6 months) from Tilburg University. They participated in exchange for course credits. None had participated in Experiments I or II and none reported any history of hearing or speech problems.

4.1.2. Materials and procedure

The materials and procedure were identical to those in Experiment I, with the exception of the referring expressions in the prime turn (see Fig. 2). In Experiment III, these prime descriptions were always overspecified ones. Thus, in the Furniture domain participants heard descriptions such as *the red chair seen from the front* and in the People domain they heard descriptions such as *the man with the glasses and the tie*. All these descriptions were overspecified in that they use two attributes (in addition to the type attribute), both a preferred and a dispreferred one, where either attribute would be sufficient. All expressions were once again produced by the same speaker as in Experiment I and II.

4.2. Results and discussion

Fig. 6 and Table 5 display the amount of overspecified references in Experiment I (single prime) and III (dual prime) for both domains. A reference was considered overspecified when both the preferred and the dispreferred attributes were used in the description. References that consisted of one of the primed attributes and another arbitrary attribute (0.18% of all produced expressions) were not included in the analysis.

The results show that when participants were primed with *both* the preferred and the dispreferred attribute, 52% of the Furniture trials and 57% of the People trials were produced with both attributes, although the preferred attribute would be sufficient to distinguish the target. Clearly, overspecification occurs much more often in Experiment III than in Experiment I, where overspecification occurred in 11%–15% of the cases. To statistically analyze these results, we combined the data for Experiment I (single prime) and Experiment III (dual prime) and conducted a mixed effects anova with amount of overspecification as the dependent variable and Domain (Furniture vs. People) as within subjects variable and Prime (single prime vs. dual prime) as between subjects variable. Table 6 summarizes the outcome of this analysis. The results show a substantial effect of Prime, $F(1, 52) = 32.50$, $p < .001$, $\eta^2 = .36$; the dual primes result in more overspecified descriptions (and thus a more frequent use of the dispreferred property) than the single primes. The effect of Domain was not

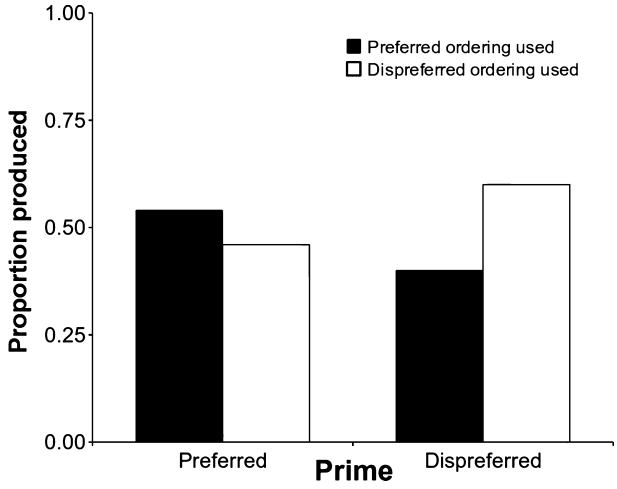

Fig. 5. Proportions of preferred and dispreferred modifier orderings in the People domain.

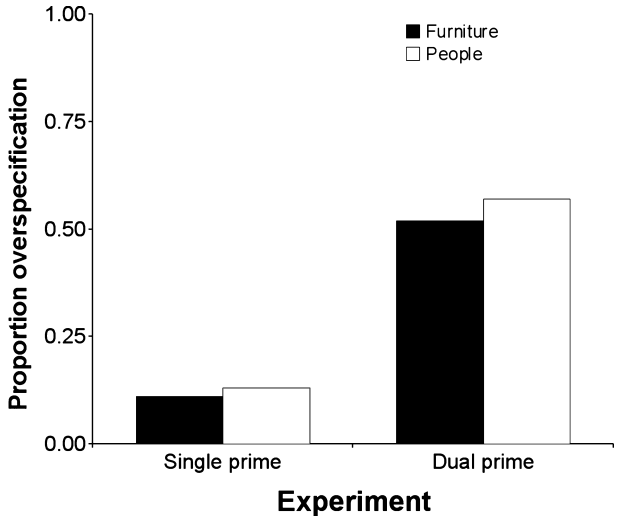

Fig. 6. Proportion of overspecification with single (*the chair seen from the front/the man with the tie*) and dual primes (*the blue chair seen from the front/the man with the glasses and the tie*) in the People and Furniture domain.

significant, and neither was the interaction between Domain and Prime. This indicates that the effect of Prime was the same for both domains.

As noted earlier, the Incremental Algorithm and other REG algorithm predict that dispreferred attributes are never used when a more preferred attribute is sufficient to uniquely characterize a target object. In other words, these algorithms predict that none of the speakers in Experiment III would use the dispreferred attribute. This is clearly not supported by the data; in both domains, the dispreferred attribute was used significantly more

Table 5
Mean overspecification (and standard deviations) for Experiments I (selection) and III (overspecification) per Domain (Furniture and People) and Prime (Preferred and Dispreferred)

	Experiment I		Experiment III
	Preferred	Dispreferred	
Furniture	0.13 (0.34)	0.11 (0.31)	0.52 (0.37)
People	0.15 (0.36)	0.13 (0.33)	0.57 (0.34)

Table 6
Summary of the mixed effects analysis of variance for Experiments I and III (referential overspecification)

	F	*df*	*p* Value	η^2
Domain	0.67	1, 52	.42	0.01
Prime	32.50	1, 52	.01	0.36**
Domain × Prime	0.13	1, 52	.72	0.00

Note. Significant at $^{**}p < .01$.

often than predicted by these algorithms, $t_{\text{furniture}}(27) = 8.46$, $p < .001$; $t_{\text{people}}(27) = 9.02$, $p < .001$.

5. General conclusion and discussion

In this article, we presented three studies, looking at attribute selection, modifier ordering, and referential overspecification using an interactive reference production paradigm. Experiment I (content selection) showed that speakers are more likely to include a (preferred or dispreferred) attribute in a referring expression when this attribute was primed a number of turns before. In Experiment II (modifier ordering), we showed that a similar influence of previously encountered material holds for the realization of referring expressions. Our participants used a certain modifier ordering significantly more often when they had previously encountered it. Finally, Experiment III (referential overspecification) showed that alignment and overspecification are closely related. While some participants were still reluctant to select a dispreferred attribute for their referring expressions in Experiment I, participants aligned frequently with an overspecified referring expression that contained *both* a preferred and dispreferred attribute, although only including the preferred one would have been sufficient to distinguish the target object from its distractors. In none of the experiments did participants realize that they were primed with specific descriptions, suggesting that the observed alignment effects were indeed largely automatic, as predicted by the Interactive Alignment Model (Garrod & Pickering, 2004; Pickering & Garrod, 2004).

In all experiments, stimuli came from two different domains: a set of furniture pictures and a set of male portraits (all famous mathematicians, none recognized by our participants). Although alignment effects were observed in both domains, we found in Experiments I and

II that the strength of the alignment effect differed per domain. Arguably, these differences are not due to differences between the domains per se, but rather to differences in strength between the preferred and dispreferred primes. Based on earlier collected referential datasets using these two domains (Gatt et al., 2007; Koolen et al., 2009), we know that the dispreferred attribute in the People domain is more dispreferred (i.e., is used less frequently) than the dispreferred attribute in the Furniture domain. Hence, we would expect alignment effects to be stronger in the Furniture domain, which is indeed what we found. Similarly, web occurrence statistics revealed that in the Furniture domain the dispreferred modifier ordering (e.g., *red large chair*) is much more infrequent than the preferred ordering (*large red chair*), while in the People domain the difference between the two primed variants is much smaller. Hence, we would expect the effects of alignment to be stronger in People domain, which is once again exactly what we found. These Domain effects indicate that the trade-off between preferences and alignment is a gradual one, also influenced by the a priori differences in preference. It is more difficult to make people say something truly dispreferred than something more marginally dispreferred.

It could be argued that the interactive nature of the experimental paradigm is somewhat limited, in that participants do not truly interact with the speaker of the referring expressions they have to comprehend. Rather, our speakers interact with an imaginary audience, which allowed us to guarantee that all participants were primed in exactly the same way. Using an imaginary audience is a standard experimental procedure to study interactive communication, and recent studies of Van Der Wege (2009) and Ferreira, Slevc, and Rogers (2005) have shown that the differences between a real audience and an imagined audience (as is the case in our studies) are small. Participants' references do not get more precise when they are interacting with a real addressee (Van Der Wege, 2009) nor do they avoid potential linguistic ambiguity in their referential expressions more when they are speaking to a real addressee compared to an imagined one (Ferreira et al., 2005). If anything, we would expect that the effects we reported in this article would be even stronger in truly interactive settings. Another potential limitation of the current studies is that we only focused on the *production* of referring expressions in interaction; an interesting question for future research is how these aligned references are comprehended by listeners; for instance, to see what effects alignment failures may have on comprehension, along the lines of Metzing and Brennan (2003).

It is worth emphasizing that our results cannot readily be explained in terms of well-understood phenomena such as lexical or syntactic alignment. In Experiment I, what is primed are not lexical items, but attributes. A prime in the Furniture domain may be *the front-facing chair*, where *front-facing* is the relevant value of the orientation attribute, while in the critical trial participants should produce a referent for, say, a fan whose orientation is to the left. Arguably, what is being primed is a way to look at an object, thereby making certain attributes of the object more salient. Similarly, the effects in Experiment II are not merely syntactic alignment effects, since both modifier orderings have the same underlying syntactic structure. Rather, the alignment effects in this experiment seem to arise at the level of surface realization in the Formulator of Levelt's model of speaking (Levelt, 1989, 1999), after the syntactic structure has been decided upon.

Current state-of-the-art REG algorithms, including the Incremental Algorithm (Dale & Reiter, 1995) and the Graph-based algorithm (Krahmer et al., 2003), fail to capture the alignment effects we found in Experiments I and III. These algorithms predict that a dispreferred attribute would never be used if a preferred attribute would be sufficient to uniquely characterize a target. And while both algorithms account for some amount of overspecification, they would never redundantly use a dispreferred (expensive) attribute, although our participants did this in over half of the cases in Experiment III. In a similar vein, current linguistic realization algorithms would always opt for the most frequent modifier ordering, which is not what our participants did.

Many REG researchers assume (sometimes implicitly) that REG algorithms should be evaluated in terms of their human-likeness (Gatt & Belz, 2010). On this view, an REG algorithm performs well if it generates referring expressions that closely resemble those produced by speakers. Of course, other goals are conceivable as well, for example, that automatically generated referring expressions should be easy to understand for listeners. A detailed discussion of what possible goals REG algorithms can have, and how this might influence their evaluation, is beyond the scope of this article (but see van Deemter, Gatt, van Gompel, & Krahmer, 2011). In any case, human-produced referring expressions are likely to remain an important source of inspiration for REG algorithms (if only, because human produced references are presumably easy to understand for listeners), and interactive REG algorithms would do good to take the current findings into account. The amount of alignment that participants engaged in was found to vary with the domain and the experiment. Sometimes the effect sizes were relatively small, but in other cases the alignment is so substantial (e.g., in Experiment III) that this would be difficult to ignore when developing interactive REG algorithms. In addition, our data show a large degree of variation between participants, with some aligning all the time and others hardly ever. Capturing such interspeaker variation is an important goal of future REG algorithms (Dale & Viethen, 2010; van Deemter, Gatt, van Gompel, & Krahmer, 2011).

As a first step, REG algorithms for interactive settings should become more sensitive to the prior interaction and the references that were produced therein. For the Incremental Algorithm, this could be achieved by augmenting the list of preferred attributes with a list of ''previously mentioned'' attributes. The relative weighting of these two lists will be corpus dependent, as we have seen, and can be estimated in a data-driven way. Alternatively, in the Graph-based algorithm (Krahmer et al., 2003), costs of attributes could be based on two components: a relatively fixed domain component (preferred is cheaper) and a flexible interactive component (recently used is cheaper). Which approach would work best is an open, empirical question, but either way this would constitute an important step toward interactive REG. Gatt, Goudbeek, and Krahmer (2010) go one step further, proposing a new model for alignment in reference production that integrates alignment and preference order based attribute selection. This model consists of two parallel processes: a preference-based search process based on the Incremental Algorithm, and an alignment-based process. These two processes run concurrently and compete to contribute attributes to a limited capacity working memory buffer that will produce the referring expression. This model was tested against the data of Experiment III and showed a similar amount of overspecification as the

human participants (Gatt et al., 2010). Note that applying REG algorithms in an interactive setting is not only relevant for applications which engage in a dialog with their user but also has a number of other theoretical and practical advantages. First of all, allowing REG algorithms to align with the addressee may drastically reduce their search space; not all alternatives need to be explored, because the search process can largely be driven by the attributes that were used previously. In addition, as we have seen, preference orders need to be empirically determined for each new domain. But what do you do when your REG algorithm is applied in a domain of which the preference order is unknown? Our experiments suggest that a good strategy might be to simply follow the addressees, preferring the same attributes that they seem to prefer and ordering modifiers in the same way as they do.

Various recent studies have shown a benefit for alignment in human–computer interaction (see Branigan, Pickering, Pearson, & McLean, 2010, for a recent survey). It has been shown that people certainly align with computers, even more so when they perceive them as dated. Based on such results, Branigan et al. (2010) argue that since alignment is an important ingredient of effective communication, this implies that human–computer interaction where alignment does not occur is a less effective way of communicating than human–human interaction. Conversely, a lot is to be gained for computers that align: ''Speakers should also feel more positive affect when interacting with a computer that aligns with them than with one that does not'' (Branigan et al., 2010). In this article, we presented new evidence on how a computer could achieve this kind of alignment for a specific subtask: the production of referring expressions in an interactive setting.

Notes

1. The attribute type (e.g., the kind of furniture item; chair, table, etc.) has a special status in the Incremental Algorithm and is always included, even if it does not rule out any distractors (i.e., if all objects are of the same type). This is because the type of an object is typically realized as the head noun (*chair*) and omitting it might result in an awkward referring expression.
2. Here and elsewhere we give English translations of Dutch originals.
3. The picture of furniture items was taken from the Object Databank, developed by Michael Tarr at Carnegie Mellon University and freely distributed at http://www.tarrlab.org.

Acknowledgments

The research reported in this article forms part of the VICI project ''Bridging the gap between psycholinguistics and computational linguistics: the case of referring expressions,'' funded by the Netherlands Organization for Scientific Research (NWO grant 277-70-007). We thank Dr. Mirjam Broersma of the Radboud University Nijmegen for assistance in recruiting the participants of Experiment I. In addition, we thank

Dr. Wayne Gray and three anonymous referees for their careful reading and insightful comments on the manuscript.

References

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

Arts, A. (2004). *Overspecification in instructive texts*. Unpublished PhD thesis, Tilburg University.

Belke, E., & Meyer, A. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during same-different decisions. *The European Journal of Cognitive Psychology*, *14*(2), 237–266.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387.

Bock, K. & Griffin, Z. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, *129*, 177–192.

Bock, K., Dell, G. S., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition*, *104*, 437–458.

Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*, 2355–2368.

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

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*, 171–190.

Buschmeier, H., Bergmann, K., & Kopp, S., (2010). Modelling and evaluation of lexical and syntactic alignment with a priming-based microplanner. In E. Krahmer & M. Theune (Eds.), *Empiricial methods in natural language generation* (pp. 85–104). Berlin: Springer.

Chartrand, T., & Bargh, J. (1999). The chameleon effect: The perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*, 893–910.

Clark, H. H. & Bangerter, A. (2004). Changing ideas about reference. In I. A. Noveck & D. Sperber (Eds.), *Experimental pragmatics* (pp. 25–49). Basingstoke, UK: Palgrave Macmillan.

Clark, H. H. & Murphy, G. (1983). Audience design in meaning and reference. In J. F. LeNy & W. Kintsch (Eds.), *Language and comprehension* (pp. 287–299). Amsterdam, The Netherlands: North Holland.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(2), 233–263.

Dale, R. & Viethen, J. (2010). Attribute–centric referring expression generation. In E. Krahmer & M. Theune (Eds.), *Empirical methods in natural language generation* (pp. 163–179). Berlin: Springer Verlag.

Dijksterhuis, A., & Bargh, J. A. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 33, pp. 1–40). San Diego, CA: Academic Press.

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language*, *54*(4), 554–573.

Ferreira, V. S., Slevc, L. R., & Rogers, E. S. (2005). How do speakers avoid ambiguous linguistic expressions? *Cognition, 96*(3), 263–284.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic coordination. *Cognition*, *27*, 181–218.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*, 8–11.

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* (pp. 264–293). Berlin: Springer Verlag.

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*, July 17-20, Sloss Dagstuhl, Germany.

Gatt, A., Goudbeek, M., & Krahmer, E. (2010). A new computational model of alignment and overspecification in reference (abstract). In J. Hajič, S. Carberry, S. Clark & J. Nivre (Eds.), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. (pp. 55-59). Stroudsburg, PA: ACL.

Gupta, S., & Stent, A. (2005). Automatic evaluation of referring expression generation using corpora. In A. Belz & S. Varges (Eds.), *Proceedings of the Corpus Linguistic 2005 Workshop on Using Copora in Natural Language Generation*. (pp. 1-6). Stroudsburg, PA: ACL.

Heeman, P. A., & Hirst, G. (1995). Collaborating on referring expressions. *Computational Linguistics*, *21*(3), 351–382.

Horacek, H. (1997). An algorithm for generating referential descriptions with flexible interfaces. In P.R. Cohen & W. Wahlster (Eds.), *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997)* (pp. 206–213). Stroudsburg, PA: ACL.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*, 91–117.

Janarthanam, S. & Lemon, O. (2009, March). Learning lexical alignment policies for generating referring expressions for spoken dialogue systems. In E. Krahmer & M. Theune (Eds.), *Proceedings of the 12th european workshop on natural language generation (enlg 2009)* (pp. 74–81). Athens, Greece: Association for Computational Linguistics.

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

Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, *29*, 459–484.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*, 25–41.

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, *33*(1), 147–151.

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). July 29, 2009, Amsterdam, The Netherlands.

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

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

Lane, L. W., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants!: Speakers' control over leaking private information during language production. *Psychological Science*, *17*, 273–277.

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

Levelt, W. (1999). Producing spoken language: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). Oxford, England: Oxford University Press.

Malouf, R. (2000). The order of prenominal adjectives in natural language generation. In H. Ida (Ed.), *Proceedings of the 38th annual meeting of the association for computational linguistics* (pp. 85–92). Stroudsburg, PA: ACL.

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.

Metzing, C. A. & Brennan, S. E. (2003). When conceptual pacts are broken: Partner effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*, 201–213.

Mitchell, M. (2010). A statistical approach to class-based ordering of prenominal modifiers. In E. Krahmer & M. Theune (Eds.), *Empirical methods in natural language generation* (pp. 1–2). Berlin: Springer, Lecture Notes in Computer Science 5790.

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.

Reiter, E. & Dale, R. (2000). *Building natural language generation systems*. Cambridge, England: Cambridge University Press.

Reitter, D. (2008). *Context effects in language production: Models of syntactic priming in dialogue corpora*. Ph.D. dissertation, University of Edinburgh.

Reitter, D., Moore, J. D., & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci)* Vancouver, Canada, July 26-29, 2006.

Shaw, J. & Hatzivassiloglou, V. (1999). Ordering among premodifiers. In N. Calzolari, B. Di Eugenio, H. Tou Ng, C. Paris, K-Y. Su, & K. Vijay-Shanker (Eds.), *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. (pp. 135–143). Stroudsburg, PA: ACL.

Stoia, L., Byron, D. K., Shockley, D. M., & Fosler-Lussier, E. (2006). Noun phrase generation for situated dialogs. In Colineau, C. Paris, S. Wan & R. Dale (Eds.), *Proceedings of the Fourth International Natural Language Generation Conference (INLG 2006)* (pp. 81–88). Stroudsburg, PA: ACL.

Stone, M. & Webber, B. (1998). Textual economy through close coupling of syntax and semantics. In *Proceedings of the 9th International Workshop on Natural Language Generation (INLG 1998)*, Niagara-on-the-Lake, Ontario, Canada, August 5-7, 1998.

van Deemter, K., Gatt, A., van Gompel, R., & Krahmer, E. (2011). Towards a computational psycholinguistics of reference production. *Topics in Cognitive Science*.

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 2007)*, Borovets, Bulgaria, September 27-29, 2007.

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