# Preferences versus Adaptation during Referring Expression Generation

**Martijn Goudbeek**
University of Tilburg
Tilburg, The Netherlands
m.b.goudbeek@uvt.nl

**Emiel Krahmer**
University of Tilburg
Tilburg, The Netherlands
e.j.krahmer@uvt.nl

## Abstract

Current Referring Expression Generation algorithms rely on domain dependent preferences for both content selection and linguistic realization. We present two experiments showing that human speakers may opt for dispreferred properties and dispreferred modifier orderings when these were salient in a preceding interaction (without speakers being consciously aware of this). We discuss the impact of these findings for current generation algorithms.

## 1 Introduction

The generation of referring expressions is a core ingredient of most Natural Language Generation (NLG) systems (Reiter and Dale, 2000; Mellish et al., 2006). These systems usually approach Referring Expression Generation (REG) as a two-step procedure, where first it is decided which properties to include (content selection), after which the selected properties are turned into a natural language referring expression (linguistic realization). The basic problem in both stages is one of choice; there are many ways in which one could refer to a target object and there are multiple ways in which these could be realized in natural language. Typically, these choice problems are tackled by giving preference to some solutions over others. For example, the Incremental Algorithm (Dale and Reiter, 1995), one of the most widely used REG algorithms, assumes that certain 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 result in a unique characterization, other, less preferred attributes such as orientation are tried. The Incremental Algorithm is arguably unique in assuming a complete preference order of attributes, but other REG algorithms rely on similar distinctions. The Graph-based algorithm (Krahmer et al., 2003), for example, searches for the cheapest description for a target, and distinguishes cheap attributes (such as color) from more expensive ones (orientation). Realization of referring expressions has received less attention, yet recent studies on the ordering of modifiers (Shaw and Hatzivassiloglou, 1999; Malouf, 2000; Mitchell, 2009) also work from the assumption that some orderings (*large red*) are preferred over others (*red large*).

We argue that such preferences are less stable when referring expressions are generated in interactive settings, as would be required for applications such as spoken dialogue systems or interactive virtual characters. In these cases, we hypothesize that, besides domain preferences, also the referring expressions that were produced earlier in the interaction are important. It has been shown that if one dialogue participant refers to a couch as a *sofa*, the next speaker is more likely to use the word *sofa* as well (Branigan et al., in press). This kind of micro-planning or “lexical entrainment” (Brennan and Clark, 1996) can be seen as a specific form of “alignment” (Pickering and Garrod, 2004) between speaker and addressee. Pickering and Garrod argue that alignment may take place on all levels of interaction, and indeed it has been shown that participants also align their intonation patterns and syntactic structures. However, as far as we know, experimental evidence for alignment on the level of content planning has never been given, and neither have alignment effects in modifier orderings during realization been shown. With a few notable exceptions, such as Buschmeier et al. (2009) who study alignment in micro-planning, and Janarthanam and Lemon (2009) who study alignment in expertise levels, alignment has received little attention in NLG so far.

This paper is organized as follows. Experiment I studies the trade-off between adaptation

*Proceedings of the ACL 2010 Conference Short Papers*, pages 55–59,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

and preferences during content selection while Experiment II looks at this trade-off for modifier orderings during realization. Both studies use a novel interactive reference production paradigm, applied to two domains – the Furniture and People domains of the TUNA data-set (Gatt et al., 2007; Koolen et al., 2009) – to see whether adaptation may be domain dependent. Finally, we contrast our findings with the performance of state-of-the-art REG algorithms, discussing how they could be adapted so as to account for the new data, effectively adding plasticity to the generation process.

## 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 way (*the left-facing fan*), when in the prior context either the first or the second variant was made salient.

### Method

***Participants*** 26 students (2 male, mean age = 20 years, 11 months), all native speakers of Dutch without hearing or speech problems, participated for course credits.

***Materials*** Target pictures were taken from the TUNA corpus (Gatt et al., 2007) that has been extensively used for REG evaluation. This corpus consists of two domains: one containing pictures of people (famous mathematicians), the other containing furniture items in different colors depicted from different orientations. From previous studies (Gatt et al., 2007; Koolen et al., 2009) it is known that participants show a preference for certain attributes: color in the Furniture domain and glasses in the People domain, and disprefer other attributes (orientation of a furniture piece and wearing a tie, respectively).

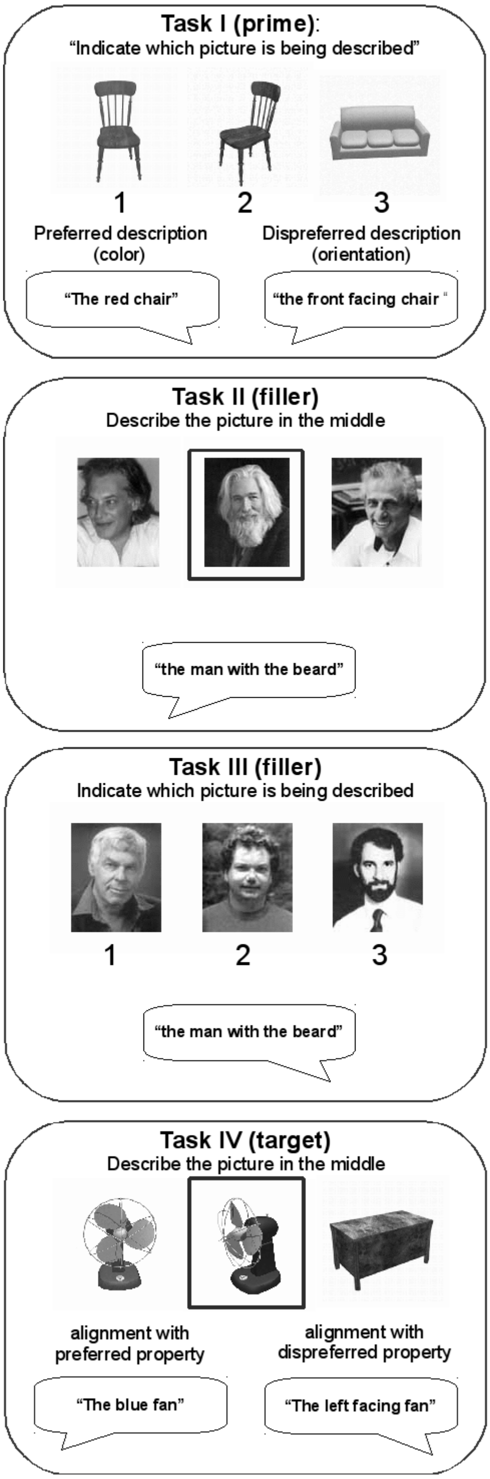
Task I (prime):
"Indicate which picture is being described"
1
2
3
Preferred description (color)
Dispreferred description (orientation)
"The red chair"
"the front facing chair"
Task II (filler)
Describe the picture in the middle
"the man with the beard"
Task III (filler)
Indicate which picture is being described
1
2
3
"the man with the beard"
Task IV (target)
Describe the picture in the middle
alignment with preferred property
alignment with dispreferred property
"The blue fan"
"The left facing fan"

Figure 1: The 4 tasks per trial. A furniture trial is shown; people trials have an identical structure.

***Procedure*** Trials consisted of four turns in an interactive reference understanding and production experiment: a prime, two fillers and the experimental description (see Figure 1). First, participants listened to a pre-recorded female voice referring to one of three objects and had to indicate which one was being referenced. 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 described. The two filler turns always concerned stimuli from the alternative domain and were intended to prevent a 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 *at-*

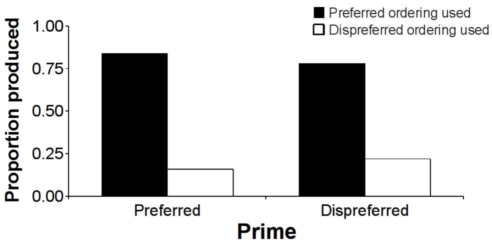

Figure 2: Proportions of preferred and dispreferred attributes in the Furniture domain.

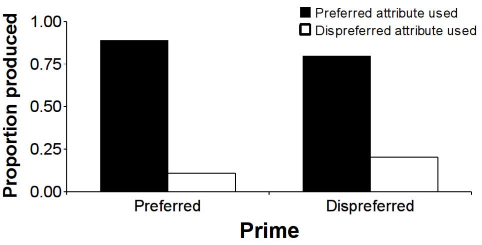

Figure 3: Proportions of preferred and dispreferred attributes in the People domain.

*tributes* are primed, not values; a participant may have heard *front facing* in the prime turn, while the target has a different value for this attribute (cf. Fig. 1).

For the two domains, there were 20 preferred and 20 dispreferred trials, giving rise to 2 x (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 Figure 1). During debriefing, none of the participants indicated they had been aware of the experiment's purpose.

### Results

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. This includes overspecified descriptions (Engelhardt et al., 2006; Arnold, 2008), where both the preferred and dispreferred attributes were mentioned by participants. Overspecification occurred in 13% of the critical trials (and these were evenly distributed over the experimental conditions).

The use of the preferred and dispreferred attribute as a function of prime and domain is shown in Figure 2 and Figure 3. In both domains, the preferred attribute is used much more frequently than the dispreferred attribute with the preferred primes, which serves as a manipulation check. As a test of our hypothesis that adaptation processes play an important role in attribute selection for referring expressions, we need to look at participants' expressions with the *dispreferred* primes (with the preferred primes, effects of adaptation and of preferences cannot be teased apart). Current REG algorithms such as the Incremental Algorithm and the Graph-based algorithm predict that participants will always opt for the preferred attribute, and hence will not use the dispreferred attribute. This is not what we observe: our participants used the dispreferred attribute at a rate significantly larger than zero when they had been exposed to it three turns earlier ($t_{furniture}$ [25] = 6.64, $p < 0.01$; $t_{people}$ [25] = 4.78 $p < 0.01$). Additionally, they used the dispreferred attribute significantly *more* when they had previously heard the dispreferred attribute rather than the preferred attribute. This difference is especially marked and significant in the Furniture domain ($t_{furniture}$ [25] = 2.63, $p < 0.01$, $t_{people}$ [25] = 0.98, $p < 0.34$), where participants opt for the dispreferred attribute in 54% of the trials, more frequently than they do for the preferred attribute (Fig. 2).

## 3 Experiment II

Experiment II uses the same paradigm used for Experiment I to study whether speaker's preferences for modifier orderings can be changed by exposing them to dispreferred orderings.

### Method

***Participants*** 28 Students (ten males, mean age = 23 years and two months) participated for course credits. All were native speakers of Dutch, without hearing and speech problems. None participated in Experiment I.

***Materials*** The materials were identical to those used in Experiment I, except for their arrangement in the critical trials. In these trials, the participants could only identify the target picture using two attributes. In the Furniture domain these were color and size, in the People domain these were having a beard and wearing glasses. In the prime turn (Task I, Fig. 1), these 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 bespectacled and bearded*

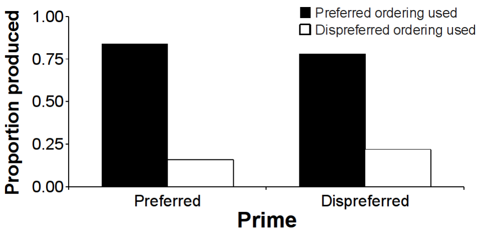

Figure 4: Proportions of preferred and dispreferred modifier orderings in the Furniture domain.

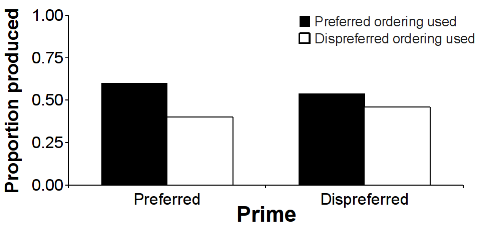

Figure 5: Proportions of preferred and dispreferred modifier orderings in the People domain.

*man*). Google counts for the original Dutch modifier orderings reveal that the ratio of preferred to dispreferred is in the order of 40:1 in the Furniture domain and 3:1 in the People domain.
***Procedure*** As above.

### Results

We use the proportion of modifier ordering alignments as our dependent measure, where alignment occurs when the participant's ordering coincides with the primed ordering. Figure 4 and 5 show the use of the preferred and dispreferred modifier ordering per prime and domain. It can be seen that in the preferred prime conditions, participants produce the expected orderings, more or less in accordance with the Google counts.

State-of-the-art realizers would always opt for the most frequent ordering of a given pair of modifiers and hence would never predict the dispreferred orderings to occur. Still, the use of the dispreferred modifier ordering occurred significantly more often than one would expect given this prediction, $t_{furniture}$ [27] = 6.56, $p < 0.01$ and $t_{people}$ [27] = 9.55, $p < 0.01$. To test our hypotheses concerning adaptation, we looked at the dispreferred realizations when speakers were exposed to dispreferred primes (compared to preferred primes). In both domains this resulted in an increase of the anount of dispreferred realizations, which was significant in the People domain ($t_{people}$ [27] = 1.99, $p < 0.05$, $t_{furniture}$ [25] = 2.63, $p < 0.01$).

## 4 Discussion

Current state-of-the-art REG algorithms often rest upon the assumption that some attributes and some realizations are preferred over others. The two experiments described in this paper show that this assumption is incorrect, when references are produced in an interactive setting. In both experiments, speakers were more likely to select a dispreferred attribute or produce a dispreferred modifier ordering when they had previously been exposed to these attributes or orderings, without being aware of this. These findings fit in well with the adaptation and alignment models proposed by psycholinguists, but ours, as far as we know, is the first experimental evidence of alignment in attribute selection and in modifier ordering. Interestingly, we found that effect sizes differ for the different domains, indicating that the trade-off between preferences and adaptions 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).

To account for these findings, GRE algorithms that function in an interactive setting should be made sensitive to the production of dialogue partners. For the Incremental Algorithm (Dale and Reiter, 1995), 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, and can be estimated in a data-driven way. Alternatively, in the Graph-based algorithm (Krahmer et al., 2003), costs of properties 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 towards interactive REG.

## Acknowledgments

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

## References

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

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. in press. Linguistic alignment between people and computers. *Journal of Pragmatics*, 23:1–2.

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

Hendrik Buschmeier, Kirsten Bergmann, and Stefan Kopp. 2009. An alignment-capable microplanner for Natural Language Generation. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 82–89, Athens, Greece, March. Association for Computational Linguistics.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Paul E. Engelhardt, Karl G. Bailey, and Fernanda Ferreira. 2006. Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, 54(4):554–573.

Albert Gatt, Ielka van der Sluis, and Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the 11th European Workshop on Natural Language Generation*.

Srinivasan Janarthanam and Oliver Lemon. 2009. Learning lexical alignment policies for generating referring expressions for spoken dialogue systems. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 74–81, Athens, Greece, March. Association for Computational Linguistics.

Ruud Koolen, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2009. Need I say more? on factors causing referential overspecification. In *Proceedings of the PRE-CogSci 2009 Workshop on the Production of Referring Expressions: Bridging the Gap Between Computational and Empirical Approaches to Reference*.

Emiel Krahmer, Sebastiaan van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

Robert Malouf. 2000. The order of prenominal adjectives in natural language generation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 85–92.

Chris Mellish, Donia Scott, Lynn Cahill, Daniel Paiva, Roger Evans, and Mike Reape. 2006. A reference architecture for natural language generation systems. *Natural Language Engineering*, 12:1–34.

Margaret Mitchell. 2009. Class-based ordering of prenominal modifiers. In *ENLG '09: Proceedings of the 12th European Workshop on Natural Language Generation*, pages 50–57, Morristown, NJ, USA. Association for Computational Linguistics.

Thomas Pechmann. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27:89–110.

Martin Pickering and Simon Garrod. 2004. Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, 27:169–226.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

James Shaw and Vasileios Hatzivassiloglou. 1999. Ordering among premodifiers. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 135–143.