# Conceptual alignment in reference with artificial and human dialogue partners

**Koen van Lierop (koenvanlierop@gmail.com)**

**Martijn Goudbeek (m.b.goudbeek@uvt.nl)**

**Emiel Krahmer (e.j.krahmer@uvt.nl)**

Tilburg center for Cognition and Communication (TiCC), Faculty of Humanities, Tilburg University,
PO Box 90153, 5000 LE Tilburg, The Netherlands

## Abstract

Previous work on reference in dialogue has shown that speakers adapt to the concepts that were used in earlier references during an interaction (such as orientation when a dialogue partner describes a chair as "the chair seen from the front"), even if these concepts are generally dispreferred. Here, we investigate to what extent it matters whether speakers interact with an artificial or a human dialogue partner (Study 1) and whether this adaptation indeed takes place at the conceptual level (Study 2). For Study 1 participants interacted either with a computer or with a human confederate and it was found that participants adapt in similar ways and just as much to a human dialogue partner as to a computer. Study 2 used a cross-language interaction paradigm, in which bilingual participants listened to English descriptions after which they had to refer in Dutch (thereby reducing the possibilities for lexical and syntactic alignment). The results showed that even with crosslinguistic prime-target pairings, participants aligned with the attributes used by their dialogue partner, providing further evidence for alignment at the conceptual level.

**Keywords:** referring expressions, alignment, human-computer interaction, conceptual alignment

## Introduction

During conversations, people continuously refer to other people, objects or events, for example using descriptions such as *the man with the beard* or *the blue chair.* Since such descriptions are so common (Poesio & Vierra, 1998), their underlying production process has drawn many researchers' attention, both from a computational and from a psycholinguistic perspective. Much of this research focusses on the question of what makes people choose one possible way of referring to an object over another. Why do speakers include certain properties in their descriptions and others not?

Computational studies of reference often frame the production of referring expressions as a problem of choice where a (usually fixed) preference order determines the order in which attributes (such as color or orientation) are considered for inclusion in a referring expression. The Incremental Algorithm (Dale & Reiter, 1995), for instance, which is probably the most widely used algorithm in this field, operates according to this principle, assuming the existence of a domain-dependent preference order on the relevant attributes, and first tries out preferred attributes before less preferred ones are considered, thereby modeling the intuition that speakers prefer certain attributes over others, partly based on findings of Pechmann (1989). For example, when referring to a chair, speakers are more likely to refer to its color (*the blue chair*) than to its orientation (*the chair facing left*) even though both may be successful in singling out a particular chair.

However, one could argue that the Incremental Algorithm is "addressee-blind" (Clark & Bangerter, 2004) in that it pays no attention to references that were produced earlier in an interaction (the same holds, incidentally, for the various alternatives that have been proposed to the Incremental Algorithm and which are surveyed in Krahmer and van Deemter, 2012). Indeed, it has been shown that speakers do take prior references into account during reference production. One study, for example, found 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, Pickering, Pearson, & McLean, 2010). This can be seen as a lexical form of "alignment" (Garrod & Pickering, 2004; Pickering & 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, for example, their intonation patterns and syntactic structures.

Goudbeek and Krahmer (2010, 2012) have shown that a similar kind of adaptation can take place at the conceptual level of attributes. While the Incremental Algorithm and its ilk only predict the use of dispreferred attributes when preferred attributes alone are unable to uniquely identify a given target object, Goudbeek and Krahmer showed that participants do use dispreferred attributes when these were used earlier in an interaction. In particular, they found that when one dialogue partner used a dispreferred attribute to refer to an object, the other dialogue partner was more likely to use values of that attribute in subsequent references as well. For example, in the furniture domain used in their experiments, participants could always uniquely identify an object by using its color (e.g., *the green chair*[1]) or its orientation (*the chair seen from the front*). Of these two attributes, color is the preferred one, in the sense that

[1] Here and elsewhere we provide English translations of Dutch originals.

without prior context speakers are more likely to use color than orientation, as was firmly established independently for Dutch (Koolen, Gatt, Goudbeek, & Krahmer, 2011) and English (van Deemter, Gatt, van der Sluis, & Power, 2012). Yet, Goudbeek and Krahmer found that when the participants had been exposed to descriptions such as *the chair seen from the front,* they were themselves more likely to use the dispreferred orientation attribute in their own references, despite the fact that using color would have been perfectly sufficient. Goudbeek and Krahmer (2010, 2012) established these effects in two different referential domains (people and furniture).

Crucially, Goudbeek and Krahmer suggest that their findings cannot be explained in terms of lexical alignment: participants were primed, for instance, with *seen from the front,* while the target was *facing left*. They argue that instead they found evidence for what they call "conceptual alignment", where participants align at the level of attributes (such as orientation) and not values (such as seen from the side). However, two potential criticisms could be levelled against this claim: first, in their experiments, participants interacted with a computer rather than with another participant, and it is conceivable that this influenced the conclusions; and second, given that primes and targets were always referred to using Dutch descriptions, the possibility that some, possibly indirect, form of lexical alignment influences the results cannot be discarded. In this paper, we address these two criticisms.

The first question we therefore ask is to what extent the results obtained with participants interacting with a computer, using a procedure in which participants had to repeatedly listen to a prerecorded description and respond with a description in front of a computer screen, are representative for human-human interaction. While using a computer-based paradigm offers several advantages in an experimental context (especially concerning controllability), Branigan et al. (2010) point out some dangers in drawing conclusions about alignment in human-human interactions from human-computer interactions (HCI). They indicate that alignment in the latter kind of interactions is mainly based on considerations of communicative success, the speaker's model of the computer and what they think the computer might or might not be able to know. Such considerations can override strong linguistic preferences, which Branigan and colleagues interpret as signs of HCI not always being a reliable predictor of real interactions between humans. Branigan et al. (2010) also argue that alignment in HCI is potentially stronger than alignment in interactions between humans; humans, they reason, may have doubts about the communicative capabilities of computers, and hence might be more inclined to adapt to the computer. This suggests that the previous findings of Goudbeek and Krahmer could have overstated the amount of alignment in interactive referring expression generation.

On the other hand, one could also argue that the participants in these studies did not strictly engage in human-computer interaction, but rather could be argued to be interacting with an "imaginary audience". Various recent studies, including Ferreira, Slevc and Rogers (2005) and Van der Wege (2009), have shown that there generally are only small differences between referring for a real and an imagined audience. References produced by participants are not more precise when they are interacting with a real instead of an imaginary addressee (Van Der Wege, 2009) nor do participants avoid potential ambiguities in their references more when they are speaking to a real addressee (Ferreira et al., 2005).

Ultimately, however, this an empirical question, which we address in Study 1. In this study, we attempt to replicate the previous findings from Goudbeek and Krahmer with human dialogue partners. Here, participants took part in an otherwise identical interactive alignment paradigm, the only difference being that instead of with a computer, participants communicated with another person which (unbeknownst to them) was a confederate of the experimenter. The participants and the confederate took turns in referring to objects and identifying objects based on the descriptions produced by their dialogue partner. As in the original paradigm of Goudbeek and Krahmer (2010), participants could always use either a preferred or a dispreferred attribute to refer to an object, and depending on the condition, the confederate either included a preferred or a dispreferred attribute when referring to an object. We compute the amount of alignment, and compare it with the amount of alignment observed in the earlier, computer-based study.

The second question we address in this paper is to what extent the alignment observed in the referential tasks indeed occurs at the *conceptual* level. In Levelt's model of speech production (Levelt, 1989; 1999), the conceptual level is the level at which the speaker decides which information to put into an utterance. In contrast to lexical alignment, where the use of *sofa* by one dialogue partner may trigger the switch from *couch* to *sofa* in the other partner, conceptual alignment refers to alignment with respect to the attribute and not necessarily with the value used by the speaker. In the previous experiments, there certainly was no direct relation between the attribute value of the prime and that of the target (and the prime and target were always separated by a pair of unrelated fillers). Nevertheless, it could be that attribute values occurring elsewhere in the experiment as primes somehow lexically primed values over longer distances, or that some other form of lexical or syntactic alignment played a role in one way or another.

To more directly test the claim of conceptual alignment, a crosslinguistic version of the interactive alignment paradigm was devised, inspired by earlier crosslinguistic priming experiments (e.g., Loebell, & Bock, 2003; Schoonbaert, Hartsuiker, & Pickering, 2007). In this experiment, described as Study 2, bilingual participants are exposed to primes in English (*the chair seen from the side*), and subsequently have to describe (after two fillers items) a target in Dutch (e.g., *de stoel van voren;* English (literally) "the chair from the front"). If Dutch participants align equally frequent with English as with Dutch primes, this would be further, and arguably more compelling evidence for conceptual alignment, since lexical and syntactic priming are relatively less likely between two different languages, even when they are relatively similar as Dutch and English (Schoonbaert et al., 2007). Again, we compare the amount of alignment when primes are in Dutch (as in the original studies) with the amount of alignment when primes are in English.

## Study 1: Artificial vs. Human Partner

This first experiment tests whether the finding of Goudbeek and Krahmer (2010), that people align with their interaction partner by using a dispreferred target attribute in their referring expressions, also holds when the interaction partner is a real person rather than a computer.

### Method

**Participants** 29 Students (23 female) from Tilburg University participated in this experiment, either for partial course credit or a small payment. All participants were fluent speakers of the Dutch language, and had normal hearing and normal (or corrected to normal) vision. None had participated in one of the earlier studies of Goudbeek and Krahmer (2010; 2012).

**Materials** The stimulus material for this study consisted of a set of furniture images and images of people, which have frequently been used in previous research on the production of referring expressions (e.g., van Deemter et al., 2012) and which where also used in Goudbeek and Krahmer (2010; 2012)[2]. The target images were always one of five different furniture items, which varied in both color and orientation. An overview of the possible combinations is provided in Table 1.

The pictures from the people domain (all black and white photographs of famous mathematicians) served as filler images, and were included to distract participants' attention from the goal of this study. Target images were always presented together with two distractors from the same domain, and were shown to participants on two synchronized monitors, to ensure that the set up was identical for the confederate and participant, in order to not raise any suspicions about the confederate with the participants. Targets could always be distinguished both in terms of color (preferred) and in terms of orientation (dispreferred). That color information is preferred and orientation information is not was determined independently from corpus data for Dutch and English (van Deemter et al., 2012; Koolen et al., 2011), showing that speakers frequently use color spontaneously and that they rarely use orientation when referring to the furniture items under study here.

Table 1: The attributes and their possible values.

| Attribute | Possible values |
|---|---|
| type | chair, desk, fan, sofa, television |
| color | red, green, blue, grey, black |
| orientation | front, back, side |

**Procedure** Participants took part in an interactive understanding and referring task, closely modelled on the paradigm presented in Goudbeek and Krahmer (2010). In this study, however, participants interacted with a female confederate (a student of the same age as the participants) instead of with a computer. The experiment consisted of two blocks, each featuring 30 trials[3]. During block one, the confederate systematically used preferred attributes when referring to furniture items (e.g., *the green chair*), during block two, the confederate systematically used dispreferred attributes (e.g., *the chair seen from the front*). The order of blocks was counterbalanced across participants. In every trial, participants first had to listen to the confederate describing a critical target (to which we refer as the prime), who referred to one of three pieces of furniture, which participants subsequently had to identify on their screen. After doing so, the next slide of images (three persons) appeared for both participant and confederate, and the participant had to describe a filler target from the person domain to the confederate, who identified it on her screen. Third, the confederate would describe a filler target (again from the people domain). In the fourth turn, the participant described a critical target that could be described with a preferred or dispreferred attribute (or both) to the confederate. Figure 1 shows an example of a critical trial.

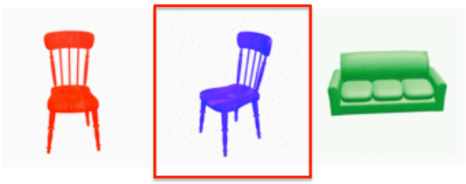

Figure 1. Example of a critical trial. The target is indicated by a red border and can be distinguished both in terms of its color (blue) or its orientation (facing left).

The use of a confederate was motivated from the fact that participants were unlikely to systematically use dispreferred properties, which would hinder a direct comparison with the results of Goudbeek and Krahmer (2010; 2012). The confederate was instructed to engage in each interaction and ask questions when participants provided insufficient information to identify a target. After

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

[3] The order of the trials within a block was the same for every participant

the experiment participants were debriefed. None suspected the other person to be a confederate.

## Results

The number of times participants aligned with the attribute they were primed with was used as the dependent variable. This includes cases in which participants used an overspecified referring expression, where both preferred and dispreferred attributes of the target where used by the participant, even though only one of them would suffice for the purpose of identification.

The analysis focuses on the proportional use of dispreferred attributes (orientation) when participants are exposed to dispreferred primes (note that when participants use a preferred prime, i.e., color, we cannot tell whether they used it because it is preferred or because it was primed). The results show that with dispreferred primes, the proportion of dispreferred attributes used by participants is considerable (*M* = 0.41, *SD* = 0.46). Contrary to the predictions of the Incremental Algorithm, the proportion of dispreferred attributes is significantly larger than zero; *t*(28) = 4.82, *p* < .001.

To investigate whether it matters if participants were interacting with a computer or a real person, the data of Goudbeek and Krahmer (2010) was compared with the data from the current experiment. Figure 2 displays the proportion of alignment of the participants who had been interacting with a computer (*M* = 0.53, *SD* = 0.43) and those of who interacted with a confederate (*M* = 0.41, *SD* = 0.46). A one-way analysis of variance with interaction partner (computer versus confederate) as the independent variable and proportion of alignment as the dependent variable showed no significant difference between interacting with a computer and interaction with a human *F*(1,48) = 2.20, *p* = .14).

Study 1 revealed that participants have a strong tendency to align with their conversation partner; when their partner uses a dispreferred attribute (referring to a piece of furniture in terms of its orientation) they are more likely to do so themselves later on. Whether their conversation partner is a computer of person has no significant influence.

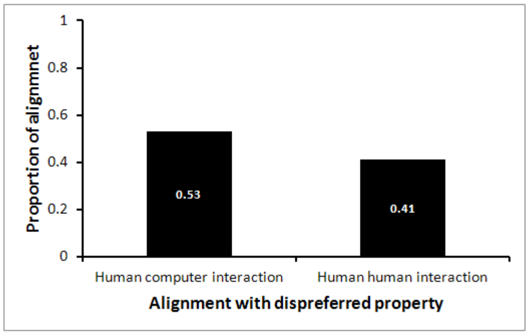

Figure 2. The proportion of alignment in human-computer and in human-human interaction

# Study 2: Crosslingual priming

The second study aims to find more conclusive evidence for the claim that the kind of alignment under discussion here takes place at the conceptual level, rather than the realization one. This is addressed using a cross-language priming experiment, where bilingual participants are not primed in their native Dutch, but in English. They do have to refer in Dutch, however, as in the previous experiment.

## Method

**Participants** 40 Students (31 female) participated in this experiment in exchange for partial course credit. All were unbalanced bilinguals, namely native Dutch speakers with formal instruction in English for 7 years or more. All had normal hearing and (corrected to normal) vision. None had participated in Study 1 or in any of the other studies described in Goudbeek and Krahmer (2010; 2012).

**Materials** The stimuli were identical to those described in Study 1, except that primes were now pre-recorded descriptions produced by a native English speaker (of roughly the same age as the participants), that referred to the objects (e.g., *the chair seen from the left*) with the same preferred and dispreferred attributes (colour and orientation, respectively) as before.

**Procedure** Before starting the experiment, participants were told they had to identify pre-recorded descriptions provided by an English speaker but had to describe the objects themselves in Dutch. In this experiment, following Goudbeek and Krahmer (2010), descriptions were once again produced by a computer, which was warranted by the findings of Study 1. The remainder of the procedure was identical to that of the previously described study.

## Results

As in Study 1, the number of times participants aligned with the dispreferred attribute when this attribute was used earlier in the interaction was used as the dependent measure. If this alignment indeed takes place on the conceptual level, it should not matter whether primes were in English or in Dutch. If, however, the alignment that participants' showed in Krahmer and Goudbeek (2010) was of a lexical or syntactic nature, we predict that people in this study will align not or to a lesser extent with the dispreferred primes, since the linguistic realization of the prime and the target differ considerably.

Figure 3 shows the use of color (the preferred attribute) and orientation (the dispreferred attribute) for each prime type. Clearly, when participants were primed with the dispreferred attribute orientation, they start using that attribute themselves more often (and the color-attribute less).

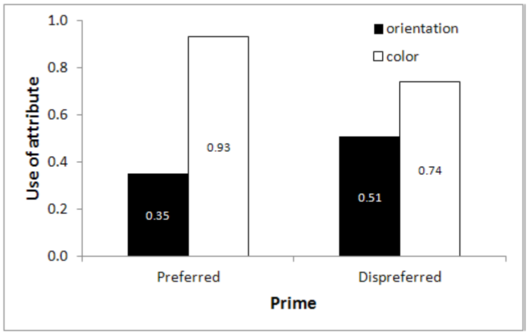

Figure 3. Participants' proportional use of the preferred and dispreferred attribute following preferred and dispreferred primes.

Statistical analysis of the data shows that despite the primes being in English, people still use the dispreferred attribute orientation significantly more than zero with $t(39) = 7.29$, $p < .001$, contrary to the prediction of the Incremental Algorithm. Moreover, participants used the dispreferred target attribute more often when they had been primed with a dispreferred attribute than when they had been primed with a preferred attribute, $F(39) = 10.92$, $p < .005$.

Furthermore, a comparison was made between the current data and the data that was collected through the experiment conducted by Goudbeek and Krahmer (2010) to test whether the language of the primes influenced the level of alignment. The results of this comparison are depicted in Figure 4. A statistical analysis showed no significant effect of the language of the primes on the amount of alignment with the dispreferred target attribute, with $t(57) = 0.83$, $p = 0.41$. Given that the experimental set-up was exactly the same, apart from the manipulation of the primes (English here, Dutch in the earlier study), this shows that the language of the prime has no impact on the amount of alignment with the dispreferred attribute, which we take as further evidence for the claim that the kind of alignment observed here is conceptual rather than lexical.

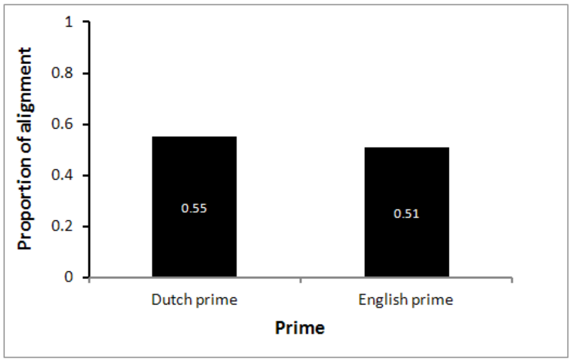

Figure 4. A comparison of the amount of alignment with the dispreferred attribute as a function of the language of the prime.

## Discussion

In this paper, we studied how speakers refer to objects in an interactive setting, and how this influences speakers' decision of which attributes to include in a description. In both studies, we found that speakers have a strong tendency to use dispreferred attributes in their descriptions when these were used earlier in the interaction, even though there was always the possibility to rely only on a preferred attribute. These results replicate the earlier findings of Goudbeek and Krahmer (2010; 2012), and extend them in two important directions.

The results of the first study show that alignment between humans is statistically indistinguishable from alignment between humans and computers. This indicates that the use of a pre-recorded conversation partner in the earlier studies did not influence the results, and is in line with the claims from Ferreira et al. (2005) and van der Wege (2009) that referring for an "imaginary audience" is similar to referring for a real audience. The results of the second study show that participants which are primed with dispreferred attributes in English, use these attributes to the same extent in their descriptions in Dutch as participants that were primed with dispreferred attributes in Dutch. This strongly suggests that the kind of alignment under study here is conceptual, rather than lexical or syntactic. What seems to be primed is a way to conceptualize or look at an object (in terms of orientation rather than color, for instance), rather than a specific property (such as facing left or being blue).

It is interesting to contrast the current findings with the predictions of state-of-the-art computational models for Referring Expression Generation (REG), including the Incremental Algorithm (Dale & Reiter, 1995) as well as more recent extensions and variations of these models. Generally, these models fail to account for the alignment results presented here; the Incremental Algorithm, for instance, predicts that a dispreferred attribute would never be used if a preferred attribute is sufficient to uniquely characterize a target object. The basic problem is that these algorithms treat the decision of which attributes to include in a description as a decision that can be made independent of context, and hence does not need to take into account the reference history, something which our data clearly contradict.

To make these algorithms suitable for the generation of referring expressions in interactive settings (as is required for many applications), they should become more sensitive to the preceding interaction and the references that were produced in it. For the Incremental Algorithm, this could be achieved, for example, by combining the (fixed, domain dependent) list of preferred attributes with a (flexible) list of "previously mentioned" attributes. The relative weighting of these two lists can be estimated based on data such as those presented here.

Gatt, Goudbeek and Krahmer (2011) go one step further, proposing a new model for alignment in reference production that integrates alignment and preference order based attribute selection. Their 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. Gatt et al. (2011) show that their model can account for the alignment findings well.

## Conclusion

When producing referring expressions in interactive settings, speakers have a strong tendency to re-use attributes that were used before, even if these attributes are otherwise dispreferred. The frequency with which dispreferred attributes are re-used does not depend on whether the interaction is with a computer or with another person, nor on whether the preceding primes where produced in a different language or not. Taken together, these results suggest that the alignment is automatic, and of a conceptual nature. Current state-of-the-art computational models of reference production fail to account for this, since they tend to be "addressee-blind" and mostly rely on domain-dependent preferences only.

## 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). We thank Manon Yassa for serving as the confederate in Study 1 and Elsa Jonker for assistance in running Study 2.

## References

Branigan, H. P., Pickering, M. J., Pearson, J., & McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, *42*(9), 2355-2368.

Clark, Herbert H., & Bangerter, A. (2004). Changing ideas about reference. In Ira A. Noveck and Dan Sperber, editors, *Experimental Pragmatics*. Palgrave Macmillan, Basingstoke, pages 25–49.

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

van Deemter, K., Gatt, A., van der Sluis, I., & Power, R. (2012). Generation of Referring Expressions: Assessing the Incremental Algorithm. *Cognitive Science*, in press.

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

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

Gatt, A., M. Goudbeek and E. Krahmer (2011). Attribute preference and priming in reference production: Experimental evidence and computational modeling. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society (CogSci 2011)*, July 20-23, Boston, Massachusetts, 2627-2632.

Goudbeek, M., & Krahmer, E. (2010). Preferences versus adaptation during referring expression generation. *Proceedings of the 48th annual meeting of the Association for Computational Linguistics (ACL)*. July 2010, Uppsala, Sweden, 55-59.

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

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

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*. Cambridge, MA: MIT Press.

Levelt, W. (1999). Producing spoken language: A blueprint of the speaker. In C. Brown & P. Hagoort (red.), *The Neurocognition of Language* (p. 83-122). Oxford: Oxford University Press.

Loebell, H., & Bock, K. (2003). Structural priming across languages. *Linguistics, 41,* 791-824

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 description use. *Computational Linguistics, 24*, 183–216.

Schoonbaert, S., Hartsuiker, R., & Pickering, M.J. (2007). The representation of lexical and syntactic information in bilinguals: Evidence from syntactic priming. *Journal of Memory and Language, 56*, 153-171.

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