Non-deterministic attribute selection in reference production

Gatt, A.; van Gompel, M.; Krahmer, E.J.; van Deemter, K.

Published in:
Proceedings of the Workshop on the Production of Referring Expressions (PRE-CogSci 2011)

Publication date:
2011

Non-deterministic attribute selection in reference production

Albert Gatt (albert.gatt@um.edu.mt)
Institute of Linguistics, University of Malta
Tilburg center for Cognition and Communication (TiCC), Tilburg University

Roger van Gompel (r.p.g.vangompel@dundee.ac.uk)
School of Psychology, University of Dundee

Emiel Krahmer (e.j.krahmer@uvt.nl)
Tilburg center for Cognition and Communication (TiCC), Tilburg University

Kees van Deemter (k.vdeemter@abdn.ac.uk)
Department of Computing Science, University of Aberdeen

Abstract

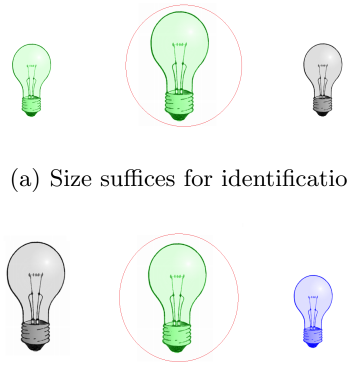
In producing identifying descriptions, speakers often overspecify and manifest preferences for certain attributes. However, current computational models which incorporate this observation tend not to make precise predictions about when and how much speakers do this. The present paper proposes and evaluates two alternative models, based on the results of a new experiment. Unlike current models, the new ones are non-deterministic and seek to make precise quantitative predictions about the extent to which speakers overspecify.

Introduction

When speakers produce definite descriptions to identify objects, they manifest preferences for certain attributes. One source of evidence for this is overspecification. For example, the referent in the circle in Figure 1(a) – the *target* – is likely to be referred to as *the large green lightbulb*, although size alone would suffice to distinguish it from the other objects, its *distractors*. By contrast, speakers would be less likely to overspecify in the case of Figure 1(b), where colour alone suffices to distinguish the target, suggesting that size is less preferred (Pechmann, 1989; Belke & Meyer, 2002). Overspecification has also been attested with other attributes, such as an object's location (Engelhardt, Bailey, & Ferreira, 2006; Arts, 2004).

From a procedural perspective, these observations suggest that in the course of incrementally constructing an identifying description, speakers prioritise preferred attributes and choose values of those attributes first (Pechmann, 1989). Another motivation for selecting a given attribute is its *discriminatory value*. For example, in Figure 1(a), the target is the only large object; thus, it might be quite evident to a speaker that size is relevant to distinguish the target, though colour might also be added due to its preferred status.

Both of these psycholinguistic insights have influenced computational REG models, which are a crucial component of many Natural Language Generation (NLG) systems (see Krahmer & van Deemter, 2011, for an extensive review). However, current REG algorithms are not fully adequate as models of human reference production. In particular, they do not make precise predictions about when speakers overspecify and to what extent they do so. Part of the problem is that these models tend to be completely deterministic, in the sense that they will always output the same description in the same situation (see van Deemter, Gatt, van Gompel, & Krahmer, 2011, for discussion). This runs counter to what we know of human reference production, or indeed, human language behaviour in general, which is widely acknowledged to be stochastic in nature (see Jurafsky, 2003, for discussion).



(a) Size suffices for identification

(b) Colour suffices for identification

Figure 1: Two domains

This paper focuses on attribute selection and overspecification, reporting the results of an experiment and describing two non-deterministic REG models which were designed and evaluated against the experimental data. The primary aim is to design computational models which can be used to make precise predictions about when overspecification occurs.

Computational approaches to REG

REG models typically take as input a a domain such as those in Figure 1, in which the properties of the target referent r and its distractors are represented as attribute-value ($\langle$ A : $v\rangle$) pairs. Their task is to select a combination of properties – the description – that singles out the target from its distractors. In what follows, we shall distinguish between an *attribute*, and a *property* (the value that an entity has for a particular attribute), sometimes using A and P to abbreviate the two. We use $[\![\ p \ ]\!]$ for the extension of a property (the set of entities it is true of). A description D can be represented as a set of properties; it is *distinguishing* if $[\![\ D \ ]\!] = \bigcup_{p \in D} [\![\ p \ ]\!] = \{r\}$, that is, there is only the target referent in the extension of the description.

Early REG models, such as the Full Brevity algorithm (Dale, 1989), were based on a strong interpretation of Grice's Maxim of Quantity, interpreted as an injunction to include no more properties in a description than absolutely necessary for identification, in line with early psycholinguistic proposals (e.g., Olson, 1970). Later models relaxed this assumption, because a strict computational interpretation of the maxim is provably intractable in the worst case. Moreover, the psycholinguistic evidence for attribute preferences and overspecification argues against this strategy.

The Greedy heuristic (Dale, 1989), proposed as a tractable approximation to Full Brevity, selects properties incrementally based on *discriminatory value*, adding to the description that property of the target that excludes most distractors at a given stage. The discriminatory value *disc*(P) of a property P can be estimated as follows:

$$disc(\text{P}) = \frac{1}{|[\![\ \text{P} \ ]\!] - \{r\}|} \qquad (1)$$

that is, the more distractors in the extension of a property, the lower its discriminatory value. As an example, in Figure 1(a), this algorithm will select the size property of the target first. Since this fully distinguishes the target, the algorithm stops here. What this procedure lacks is (i) a way of making choices in case of a tie (that is, when two properties are equally discriminatory); (ii) a way of deciding when to stop not only based on whether a description identifies the target, but also on the attributes that it contains.

An alternative, highly influential, REG model is the Incremental Algorithm (IA) proposed by Dale and Reiter (1995), which partially addresses the psycholinuistic findings about attribute preferences. The IA works using a pre-specified *preference order* of attributes (e.g. colour > size for the example domains in Figure 1). It traverses the preference order, and selects the value of an attribute of the target if it excludes at least one of the remaining distractors. As in the Greedy Algorithm, the procedure terminates when the distractor set is empty, or there are no more attributes to choose from. It is easy to see that this procedure will generate a description containing both colour and size in Figure 1(a) (colour, which is tested first, is included because it removes the grey lightbulb, leaving the small green one, which is excluded once size is considered). By contrast, it will only select colour in Figure 1(b). In this respect, the model has a certain *prima facie* plausibility, as it seems to overspecify when human speakers would.

However, the IA is only adequate as a model of human reference production to the extent that human speakers *always* make the same choices in structurally isomorphic domains. More generally, given a preference for attribute A over attribute B, the IA predicts that in any domain in which there is at least one distractor that has a different value for A compared to the target, descriptions will contain A 100% of the time, which is clearly unrealistic. By hypothesis, humans do not always overspecify in all isomorphic situations, in spite of the robust evidence for attribute preferences.

One consequence of the way both the IA and the Greedy heuristic are formulated is that there are domains in which they predict no overspecification at all. Consider, for instance, a domain in which *either* size *or* colour would suffice for a distinguishing description. Here, the IA will output a colour-only description 100% of the time; the Greedy heuristic's output will depend on how ties are resolved (this is not specified in the original formulation), but there is no principled account in either case of whether overspecification is ever warranted in this situation.

It is worth noting that similar problems also arise with more recent REG models which are stochastic and data-driven (e.g., Fabbrizio, Stent, & Bangalore, 2008; Dale & Viethen, 2010, among others). These models tend to be based on probability distributions learned from data, but will also output the same description in all isomorphic situations, because they choose the *most likely* alternative, rather than varying their output according to the distribution.

There are at least two ways of re-interpreting a model like the IA non-deterministically. A simple way is to vary the preference order according to a probability distribution. In our example domains, this could be done by by sometimes considering size before colour. This is the essence of the first model we test below. A second possibility is to combine elements of both the Greedy and the Incremental heuristics, prioritising attributes based on discriminatory value, but resolving ties and making decisions about overspecificiation non-deterministically based on preference. This is the essence of the second model we test below. First, however, we describe an experiment that was designed to enable us to determine choice probabilities for the models, and to evaluate the outcome of the models against human data.

Experimental evidence

We conducted an experiment on two separate groups of English and Dutch-speaking participants, for which results are reported separately. The experiment was carried out in two different languages to ensure generalisability. Its aims were (i) to compare the predictions of the IA, which deterministically selects attributes based on its preference order, against human data; (ii) more importantly, to enable precise predictions about the nature of the non-determinism observed, identifying the probabilistic parameters that govern atribute choices and the concomitant tendency to overspecify.

For the purposes of this experiment, we used domains such as those in Figure 1, focusing on the size and colour attributes. The reason for this is that, as indicated in the introductory section, the relative preference for colour over size is well-established and has been replicated numerous times. However, our ultimate aim is to design an algorithm which can be generalised to new domains; we return to the issue of generalisability in the concluding section.

Participants The Dutch version of the experiment was conducted on 36 pairs of undergraduates from Tilburg University, who were all native speakers of Dutch and who participated in return for course credit. For the English version, there were 30 pairs of undergraduates at the University of Dundee, who participated voluntarily. All participants had normal or corrected-to-normal vision.

Materials Thirty-six domains such as those shown in Figure 1 were constructed using a version of the Snodgrass and Vanderwart line drawings with colour and texture (Rossion & Pourtois, 2004). Each domain consisted of one target object and two distractors. We used thirty-six different objects for the targets. These were selected from the picture set on the basis of a pretest in which seven native speakers of Dutch and seven of English were asked to name greyscale versions of the pictures. For the items, we selected only those pictures for which at least 5 out of the 7 speakers of either language agreed on the name of the object.

Design The experiment used domains consisting of three objects, one of which was designated as the target, as shown in Figure 1. There were three different conditions: (i) **S**: Size sufficed to distinguish the target (see Figure 1(a)); **C**: Colour sufficed to distinguish the target (see Figure 1(b)); (iii) **C/S**: A baseline condition in which either colour or size sufficed to distinguish the target. The 36 items were distributed across three lists, with each item appearing in each list in a different condition. Participants were divided into three groups so that each group saw a different item list, with twelve items per condition. The experiment also included 72 filler items, consisting of domains in which the target could be identified using its type only (e.g. *the kettle*, in a domain where the distractors were two chairs), or the target consisted of two objects rather than one, or the target was identifiable via attributes that were not being manipulated in the experiment (e.g. stripes, spots, or orientation). None of the 36 experimental targets were used in the fillers. Items were displayed to participants in a pseudo-random order, with each experimental item being preceded by two fillers.

Procedure The experiment used a director-matcher paradigm. Participants were tested in pairs, with one randomly assigned to the role of speaker/director and the other to the role of listener/matcher. Participants did not switch roles. The director and matcher faced each other in the experiment; each had a computer screen that could not be seen by the other. The speaker used a keyboard to request an item, whereupon she identified the target for the listener, who clicked on the target on his own screen. Participants were instructed to keep the interaction to a minimum, with the listener only responding by indicating to the speaker that he had finished identifying the target.

Annotation The speakers' descriptions were transcribed and classified according to whether they contained colour, size or both. Responses that did not contain the attributes of interest (colour or size) were excluded from the analysis. This resulted in the exclusion of 11 (1%) of the English responses and 12 of the Dutch. The data for four pairs of Dutch participants was excluded because of technical problems that compromised the recordning, or because they did not follow instructions.

Results

Table 1 displays the proportion of each type of description in each condition. The data suggests that the proportion of description types differed substantially as a function of condition, in the direction that previous research would lead us to expect, with a majority of overspecified descriptions containing both colour and size in the S condition, and a majority of colour-only descriptions in the C condition. In the baseline C/S condition, there is a larger proportion of colour-only descriptions, as the preference for colour would predict; however, this is smaller than in the C condition, with approximately a quarter of Dutch descriptions and 17% of English ones in this condition including both attributes.

These preliminary impressions were confirmed by separate ANOVAs by participants (F_1) and items (F_2) on arcsin-transformed proportions of each description type. For both languages, there were signif-

Table 1: Percentage of each description type in the experiment for Dutch and English speakers. Frequencies are in parentheses.

		Colour only	**Size only**	**Colour + size**
Size sufficient (S)	Dutch	0.3 (1)	21.1 (80)	78.6 (297)
	English	3.3 (12)	16.5 (59)	80.2 (288)
Colour sufficient (C)	Dutch	89.5 (334)	0.3 (1)	10.2 (38)
	English	91.9 (327)	0 (0)	8.1 (29)
Colour or size (C/S)	Dutch	70.8 (266)	3.7 (14)	25.5 (96)
	English	79.1 (280)	3.7 (13)	17.2 (61)

Table 2: Predicted response proportions for Model 1

		Colour only	**Size only**	**Col+size**
S	Dutch	0	5%	95%
	English	0	4%	96%
C	Dutch	95%	0	5%
	English	96%	0	4%
C/S	Dutch	95%	5%	0
	English	96%	4%	0

icant differences between conditions in the proportion of colour-only descriptions (Dutch: $F_1(2,31) = 294.59; F_2(2,35) = 386.79$; English: $F_1(2,29) = 1159.42; F_2(2,35) = 4840.05$), size-only descriptions (Dutch: $F_1(2,31) = 18.70; F_2(2,35) = 71.41$; English: $F_1(2,29)214.262; F_2(2,35) = 707.11$) and descriptions containing both colour and size (Dutch: $F_1(2,31) = 81.89; F_2(2,35) = 121.29$; English: $F_1(2,29) = 503.43; F_2(2,35) = 1655.25$), with $p < .001$ in all cases.

Although the data clearly replicate previous findings as far as attribute preferences are concerned, proportions also diverge significantly from the IA's predictions. For example, we observe roughly 80% colour+size descriptions in the S condition, compared to the 100% predicted by the IA. In other words, in around 20% of cases, the algorithm would have resulted in a mismatch with what participants actually produced. Similarly, in both English and Dutch data there is a significant number of overspecified descriptions in the C/S condition, where the IA would produce colour-only 100% of the time. In short, the evidence for preferences and the tendency to overspecify are subject to some variation that deterministic models can't handle. In what follows, we first propose an initial, probabilistic version of the IA, and then turn to an alternative which combines both preferences and considerations of discriminatory value.

Model 1

We first consider a non-deterministic version of the IA which probabilistically varies the preference order - that is, the likelihood that colour will be considered before size or vice versa. We obtain probabilities from the baseline C/S conditon, focusing on the proportion of colour-only or size-only descriptions. These are used as indicators of the probability of selecting either attribute first. For example, Dutch participants produced 266 colour-only descriptions and 14 size-only descriptions in this condition, giving us a 95/5% colour/size preference; the English data evinces a 96/4% split. Thus, in the Dutch data, Model 1 predicts that a size-only description should be produced 5% of the time in the S condition, while descriptions should be overspecified 95% of the time in this condition (since the algorithm checks colour first 95% of the time). Similarly, this procedure would produce a colour+size description 96% of the time in the S condition in the English data.

Predicted response proportions were calculated in this way for the English and Dutch data separately. The results are summarised in Table 2. We evaluated the model using ANOVAs by participants and items to compare its predictions to the experimental data, focusing on the cases where the algorithm would select colour first. In the S condition, there was a significant difference in the proportions of overspecified colour+size descriptions, both by participants and items (Dutch: $F_1(1,31) = 12.67, p = .001; F_2(1,35) = 47.02, p < .001$; English: $F_1(1,29) = 13.79, p = .001; F_2(1,35) = 25.6, p < .001$). The model showed a slightly better fit to the data in the C condition, where the difference in proportions of colour-only descriptions was not significant by participants, though it was significant by items (Dutch: $F_1(1,31) = 1.53, ns; F_2(1,35) = 4.24, p < .05$; English: $F_1(1,29) = 1.53, ns; F_2(1,35) = 5.58, p < .05$).

These results suggest that a simple reinterpretation of the IA, which simply varies the preference order non-deterministically, will not achieve an optimal fit to the human data. The model fares particularly badly in the S condition, where a comparison between Tables 1 and 2 shows that the model's predictions are well above the actual proportions of overspecified colour+size descriptions. This suggests that in this condition, speakers were not only being influenced by the nature of the attributes available, but also by their discriminatory value. Note further that the model, like the original IA, predicts no overspecification in the C/S condition, which runs counter to the evidence. In our earlier discussion of current REG models, we suggested an alternative model, one combining attribute preferences with discriminatory value. It is to this model that we turn next.

Model 2

The rationale behind Model 2 is the following: given a choice of attributes to include in a description, the choice

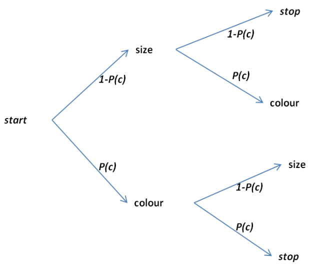


Figure 2: Model 2 choices in Condition 3

of the next attribute is based on (i) the discriminatory power of the value for that attribute (more discriminatory ones are preferred), estimated according to equation (1) and (ii) the relative preference of the available attributes, in case there is more than one attribute value with the same discriminatory power. Preferences also influence termination: Model 2 has a non-deterministic stopping criterion, where termination depends in part on which attributes have already been selected and included in the description, and which choices remain.

To make these ideas more precise, let $\mathcal{A}$ be the set of attributes belonging to a target referent r. We assume an empirically determined function $P : \textsc{a} \rightarrow [0, 1]$, which reflects the relative preference of the attributes in $\mathcal{A}$. We describe how this distribution is determined from our experimental data below. Suppose that at a given stage of processing, the description D is not yet distinguishing and $A_{max} \subseteq \mathcal{A}$ is the set of remaining attributes whose values have maximal discriminatory power with respect to r. In this case, the model selects the value of the attribute $\textsc{a}_{next} \in A_{max}$ with probability $P(\textsc{a}_{next})$. This 'roulette-wheel' behaviour will select the most preferred (most probable) attribute in the majority of cases, but will sometimes select less probable attributes. The final piece of the equation is the stopping criterion. Suppose $[\![D]\!] = \{r\}$, but there are still attributes that can be included. In this case, the algorithm has a choice between choosing $\textsc{a}_{next}$ with probability $P(\textsc{a}_{next})$, or stopping with probability $P(\textsc{stop})$, defined as follows:

$$P(\textsc{stop}) = \prod_{\textsc{a} \in D} P(\textsc{a}) \qquad (2)$$

In other words, the decision to stop once $[\![D]\!] = \{r\}$ is made non-deterministically based on the relative probability of the combination of attributes already selected in D, compared with the probability of the next candidate attribute $\textsc{a}_{next}$.

The tree in Figure 2 shows the decisions taken by Model 2 in the baseline C/S condition of our experiment. The first attribute choice involves a tie in discriminatory value between colour and size which is resolved probabilistically. Since either attribute suffices to distinguish the referent in this condition, the next step is a probabilistic choice between the remaining attribute and termination. In the S and C conditions, where size or colour suffice to distinguish the target, the initial choice is determined based exclusively on discriminatory power. Overspecification occurs if the algorithm non-deterministically adds another attribute after this initial choice. The data from the C/S condition suggests that $P(c) > P(s)$ (see Table 1). Thus, in the S condition, we expect the model to overspecify (choosing colour after the initial choice of size) more than in the C condition, since the decision to stop in the former is based on a comparison between the probability of size (the only attribute selected at the start) and the probability of selecting colour, whereas colour is selected first in the other condition.

Table 3: Predicted response proportions for Model 2

		Colour	**Size**	**Colour+size**
S	Dutch	N/A	16%	84%
	English	N/A	11%	89%
C	Dutch	84%	N/A	16%
	English	89%	N/A	11%
C/S	Dutch	71% (from data)	3%	26%
	English	79% (from data)	1.2%	21%

Determining the probability distribution As in the case of Model 1, the probability distribution $P(\textsc{a})$ is determined empirically from our baseline experimental condition. As shown in Figure 2, since there are only two attributes in the model, only one parameter needs to be determined, namely $P(c)$, the probability of selecting colour; the probability of selecting size is simply $1-P(c)$. Given that a colour+size description can be obtained via two alternative routes in Figure 2, the probability of obtaining such a description is $2 \times [P(c) \times 1 - P(c)]$, whereas the probability of a colour-only description is $P(c)^2$. Thus, $P(c)$ is straightforwardly computed as the square root of the proportion of colour-only descriptions in the C/S condition (.71 in the Dutch data; .79 in the English data; see Table 1). This single parameter suffices to estimate the model's predictions for the other conditions; these are displayed in Table 3.

Model evaluation As before, we evaluated Model 2 by comparing the predicted and observed response proportions of the different types of descriptions in the S and C conditions. Once again, we focus on the proportions of descriptions formed by choosing colour first. In the S condition, there was no difference in the proportion of overspecified colour+size descriptions by participants (Dutch: $F_1(1, 31) = 0, ns$; English: $F_1(1, 29) = 1.82, ns$). The difference reached significance by items only in the case of the Dutch data (Dutch: $F_1(1, 35) = 13.07, p = .001$; English: $F_2(1, 35) = 2.72, ns$). In the C condition, the algorithm diverged significantly from the observations in the proportion of colour-only descriptions for both languages by both participants and items (Dutch: $F_1(1,31) = 10.91, p = .002; F_2(1,35) = 5.98, p = .02$; English: $F_1(1,29) = 9.93, p = .004; F_2(1,35) = 7.04, p - .01$).

In summary, Model 2 has a slightly better fit to the data than Model 1, at least in the case where the initial choice is the relatively dispreferred size attribute. Here, the model correctly predicts that a further choice will be made in addition to the most discriminatory attribute. However, the results are still not optimal. As a comparison between Tables 1 and 2 shows, the model underestimates the likelihood of a non-overspecified colour-only description in the C condition. The difference in goodness of fit between the C and S conditions suggests the configuration of the domain – that is, which attributes are most salient and discriminatory – plays a role in determining the parameters for making choices. Where an attribute is both highly discriminatory and highly preferred, the likelihood of choosing it alone increase.

Conclusions and future work

This paper has focused on attribute preference and overspecification in reference and their implications for computational models. We have argued that current models provide a poor fit to human data and proposed two alternative, non-deterministic models that implement choices in a 'roulette-wheel' fashion (see Belz, 2007, for an example of such models in a different NLG context), using empirically determined parameters. The two models we have tested, though presenting slight improvements on the original deterministic models, are not perfect. In particular, their goodness of fit varies as a function of experimental condition. For Model 2, this suggests that the two decision criteria – discrimination and preference – need to be better combined to deal with different types of domains (for example, domains where a highly preferred attribute is also highly discriminatory or salient, versus those where highly preferred attributes have low discriminatory value).

While the models described here are intended to be general, we have so far tested them on experimental data using relatively simple domains, with only two attributes. However, the current work does suggest that 'roulette-wheel' models, based on empirically determined distributions are a promising way forward for computational REG, and potentially for other areas of NLG as well. In our current work, we are seeking to refine and scale up our model based on a follow-up experiment involving more complex domains with more objects and a greater array of attribute choices. In this way, we are also aiming to systematically trade off discriminatory value and preference in a more fine-grained fashion, to address the main research question raised by our second model.

Acknowledgments

This work forms part of the project *Bridging the gap between psycholinguistics and computational linguistics: The case of Referring Expressions*. Albert Gatt and Emiel Krahmer were supported by the Netherlands Organization for Scientific Research (NWO).

References

Arts, A. (2004). *Overspecification in instructive texts.* Unpublished doctoral dissertation, Univiersity of Tilburg.

Belke, E., & Meyer, A. (2002). Tracking the time course of multidimensional stimulus discrimination. *European Journal of Cognitive Psychology*, *14*(2), 237–266.

Belz, A. (2007). Automatic generation of weather forecast texts using comprehensive probabilistic generation space models. *Natural Language Engineering*, *14*(4), 431–455.

Dale, R. (1989). Cooking up referring expressions. In *Proc. ACL'89.*

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

Dale, R., & Viethen, J. (2010). Attribute-centric referring expression generation. In E. Krahmer & M. Theune (Eds.), *Empirical methods in natural language generation* (Vol. 5790). Berlin and Heidelberg: Springer.

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

Fabbrizio, G. D., Stent, A. J., & Bangalore, S. (2008). Trainable speaker-based referring expression generation. In *Proc. CONLL'08.*

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics.* Cambridge, Ma.: MIT Press.

Krahmer, E., & van Deemter, K. (2011). Computational generation of referring expressions: A survey. *Computational Linguistics*. (To appear)

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

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

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwarts object databank : the role of surface detail in basic level object recognition. *Perception*, *33*, 217–236.

van Deemter, K., Gatt, A., van Gompel, R., & Krahmer,

E. (2011). Towards a computational psycholinguistics of reference production. *Topics in Cognitive Science*. (to appear)