Does domain size impact speech onset time during 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 P.G. 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 referring to a target referent, speakers need to choose a set of properties that jointly distinguish it from its distractors. Current computational models view this as a search process in which the decision to include a property requires checking how many distractors it excludes. Thus, these models predict that identifying descriptions should take longer to produce the larger the distractor set is, independent of how many properties are required to identify a target. Since every property that is selected is checked, they also predict that distinguishing a target should take longer the more properties are required to distinguish it. This paper tests this prediction empirically, contrasting it with two alternative predictions based on models of visual search. Our results provide support for the predictions of computational models, suggesting a crucial difference between the mechanisms underlying reference production and object identification.

Keywords: Referring expressions, language production, visual search, computational modeling

Introduction

When a speaker refers to a target referent in a visual domain, she identifies it for an addressee by using properties which distinguish it from its distractors. For example, in order to identify the object surrounded by a red border in Figure 1, a speaker needs to refer to it using both its colour and its size (*the large blue aeroplane*); leaving out either of these properties would result in an underspecified description.

Most psycholinguistic accounts of reference in such domains assume that the discriminatory value of properties plays an important role, since the objective is to identify an object for the addressee (Olson, 1970). On the other hand, it is also well-established that certain properties are 'preferred' in that speakers often include them when they are not required to distinguish the target, thus producing overspecified descriptions (Pechmann, 1989 ; Belke & Meyer, 2002 ; Arts, 2004).

The present paper is concerned with the mechanisms underlying the selection of properties. Specifically, we ask whether this process is best viewed as a *search*, along the lines suggested by current computational models of Referring Expression Generation (REG; see Krahmer & van Deemter, 2012, for a survey). In these models (described more fully in the next section), the decision to include a property in a description requires checking it against the distractor set to determine whether it excludes at least some of them. If speakers do perform such a procedure, then larger domain sizes should result in more effort (and this should be indicated, for example, by increased speech onset times). This is because more objects have to be checked every time a property is considered for inclusion.

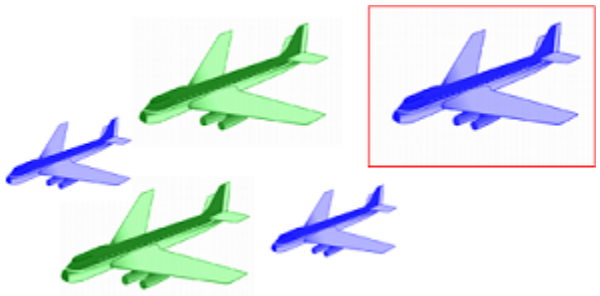

Figure 1: An example domain

This prediction is compatible with a classic finding in the visual search and attention literature, where search time has been shown to increase linearly with domain size (Treisman & Gelade, 1980). However, whereas REG models predict an impact of domain size irrespective of the number of properties required to distinguish a target referent, the task used by Treisman and Gelade only evinces a linear increase with targets distinguished by a *conjunction* of properties (e.g. *blue and large*). When targets are distinguished by a single property, a 'pop-out' effect is observed and domain size has no impact.

Yet a third possibility is suggested by more recent visual search models (e.g. Itti & Koch, 2001), which give a more central role to parallel processing. In these models,

an initial, first-pass overview of a visual domain results in the parallel activation of salient features, forming a saliency map. From a production perspective, such models would predict that, irrespective of how many properties are needed for identification, domain size should have no impact on search time because the salient, contrastive features of a referent could be 'read off' the saliency map without exhaustive checking against the distractors.

In short, there are at least three alternative models that could account for how a speaker selects properties, each with different predictions concerning the impact on search time of (i) the size of the visual domain in which the referent must be distinguished, (ii) the number of properties that are required to achieve this. In the remainder of this paper we first discuss these models in more detail and then describe an experiment that sought to investigate these relationships. In our experiment, we focus on speech onset time as an indicator of the amount of search effort required to produce a distinguishing description.

Three alternative models

Computational models of Referring Expression Generation (REG) are core components of systems which automatically generate text or speech from non-linguistic data (Reiter & Dale, 2000). Their aim is to determine the content of a distinguishing description of a referent in a given domain. REG algorithms usually view this process of content selection as a search (cf. Bohnet & Dale, 2005) and this is often modelled as an incremental procedure (e.g. Dale, 1989 ; Dale & Reiter, 1995, as well as several models based on these). The models that this paper is concerned with focus on 'first-pass' references, that is, the generation of initial descriptions in domains which are assumed to be mutually known between speaker and hearer.

The input to a REG algorithm is a domain of objects (such as Figure 1), one of which is the target referent. These algorithms iterate through the available properties (for example, the size and colour) of the objects. Each property is considered as a possible candidate for inclusion in the description.[1]

For each property, the algorithm checks whether there is at least one distractor that it excludes. If this is the case, then the property is included in a description and the distractor set is updated. For example, suppose the first property to be considered in Figure 1 is ⟨COLOUR:*blue*⟩. A check of each of the four distractors shows that there are two non-blue objects. Since it has some discriminatory value, this property is added to the description and the distractor set updated to leave only the two remaining distractors. Since, at this stage, the target referent is not yet distinguished (it is not the only blue object, as shown by the presence of the two remaining distractors), the process does not terminate, but considers the next available property, ⟨SIZE:*large*⟩.[2] Upon checking, the algorithm discovers that the two remaining distractors are both excluded by this property. Since, at this stage, there are no remaining distractors, the procedure terminates with a description that contains both size and colour. From the perspective of the present paper, this content selection procedure makes the following important predictions:

1. Search time should increase linearly with the number of distractors, since each candidate property has to be checked against the distractor set to determine its discriminatory value;
2. Since search is incremental and every candidate property is checked, the effect of domain size should be observed irrespective of whether a distinguishing description contains a single property or a conjunction;
3. Independently of (2) above, the more properties are required to distinguish the target referent, the longer the search time should be, because each property represents a cycle in an iterative search procedure.

The first of these predictions is compatible with classic findings in the visual search literature. Treisman et Gelade (1980) reported a steep linear increase in search time with increasing domain size in tasks in which participants have to determine whether some object is present in a visual domain in response to a question presented beforehand (e.g. *Is there a red vertical?*). Various replications of this effect have shown that reaction time increases by as much as 31ms with every new object (e.g. Spivey, Tyler, Eberhard, & Tanenhaus, 2001). However, the effect only holds when participants search for a target defined by a conjunction of properties. Single property search (e.g. *Is there a vertical?*) evinces a 'pop-up' phenomenon, attributed to parallel activation of salient features, which obviates the need for serial search and intergration of multiple features. Interestingly, in the case of conjunction search, the linear increase in search time with domain size is altered if participants hear the description of a target *concurrently* with the presentation of a visual scene (Spivey et al., 2001). In this case, the slope is significantly shallower, perhaps because concurrent presentation of description and domain allows listeners to incrementally circumscribe the search domain

[1] An important factor that distinguishes algorithms from each other is how they prioritise properties during search. For example, Dale et Reiter (1995) propose to prioritise properties based on the preferences evinced by humans in psycholinguistic experiments (e.g. Pechmann, 1989). By contrast, Dale (1989) proposes a model which prioritises properties based on their discriminatory value. Here, we abstract away from these differences, focusing on the basic search mechanisms common to all of them.

[2] We are assuming that the *large* value of the SIZE attribute is determined as part of this procedure, perhaps by an algorithm along the lines described by Deemter (2006).

as each property is processed (cf. Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995 ; Sedivy, Tanenhaus, Chambers, & Carlson, 1999, for related observations).

The third model we alluded to in the Introduction makes neither of the predictions of REG models. It is possible that speakers default to a 'fast and frugal' heuristic whereby, instead of searching for a distinguishing property or combination thereof, they rely on a first-pass overview of the coarse visual features of the domain to select all the properties that seem to have some contrastive value. Such a procedure would be compatible with more recent models of visual search and attention, where salient contrasts are activated through an initial, parallel process that results in a saliency map (Itti & Koch, 2001). Under this model, we would expect the number of properties required to distinguish a referent to have no impact on search time. We would also not expect the function modeling the impact of domain size on search time to increase linearly.

As the foregoing discussion suggests, REG models view the mechanism underlying reference production as fundamentally similar to that of object identification, namely, as a search process. However, there is a crucial difference which accounts for the different predictions made by the two classes of models.

In reference production, search is *object-driven*. A speaker knows which target is intended. Her task, as it is modelled in REG, is to identify a set of properties which are individually discriminatory (each contributes to the overall goal of identification) and jointly distinguishing. Thus, even determining whether a single property is discriminatory requires a check against the distractor set. By contrast, in the standard identification task, search is *property-driven*: a description of the object serves as an instruction to pick out a particular entity. If the description contains only a single feature, a pop-out search is sufficient, for the listener need not verify that the feature in question has discriminatory value – that assumption should follow from the fact that the speaker is being cooperative and is not including redundant information. Indeed, recent work has suggested that the inclusion of properties with no discriminatory value – a strong tendency among speakers, as we discussed above – results in increased processing effort for listeners, suggesting that they do in fact make this assumption (Engelhardt, Baris Demiral, & Ferreira, 2011).

Experiment

In the experiment, participants were shown visual domains of the kind displayed in Figure 1 and asked to produce a distinguishing description of the target referent. We measured the time it took participants to initiate a description, as a function of the size of the domain and the number of properties (one or two) required to distinguish the target.

Participants

The experiment was conducted at the Tilburg center for Cognition and Communication. Forty native speakers of Dutch participated in return for course credit.

Materials and design

The experimental stimuli consisted of 64 items selected from a version of the Snodgrass and Vanderwart set of line drawings with colour and texture (Rossion & Pourtois, 2004). The items were selected on the basis of a pretest in which seven native speakers of Dutch 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 agreed on the name of the object. These were subsequently manipulated to create versions in different sizes and colours. For each item, 8 versions of a visual domain were constructed, each consisting of a target referent indicated by a red border, and a number of distractors. The 8 versions represented combinations of the following two factors:

- *Properties* (2 levels): Either size only (S) or both colour and size (CS) were required to distinguish the target. Figure 1 is an example of the CS condition.
- *Distractors* (4 levels): There were 2, 4, 8 or 16 distractors in addition to the target, representing increasing domain size.

In each domain, all objects (target and distractors) were of the same type (e.g. all were aeroplanes). In S trials, distractors were identical to the target except for size. Distractors were also identical to each other (e.g. the target was a small blue aeroplane and all distractors were large blue aeroplanes). In the CS trials, half the distractors were identical to the target except for their size and the other half were identical to the target except for their colour (e.g. the target was a large blue aeroplane, half the distractors were small green aeroplanes and the other half were large blue aeroplanes).

In addition to the experimental items there were 108 fillers. In 64 of these, the target could be distinguished using size only or both size and colour, as in the trials. However, there was variation in the types of distractors (not all distractors were of the same same type as the target). In the remaining 64 fillers, the target could be distinguished by using its type only. There were equal numbers of fillers containing 2, 4, 8 or 16 distractors.

In each trial, objects were presented in a sparse grid. For each item, the position of the target was fixed in advance and was the same in all conditions. The position of the distractors was also fixed in the 2-, 4-, 8- and 16-distractor conditions. Both items and participants were randomly divided into 8 groups. Each participant saw exactly 8 items in each condition; item and participant groups were rotated through a latin square so that each

item was seen in a condition by an equal number of participants.

Procedure

Participants did the experiment individually in a soundproof booth, wearing a headset through which their descriptions were recorded. The experiment was run using the DMDX package for stimulus presentation (Forster & Forster, 2003). They were asked to imagine that they were describing objects for a listener who could see the exact same objects but did not know which one was the target referent. In order to avoid the use of descriptions containing locative expressions, participants were also told that their putative listener would see the objects in different positions (none of the participants used locatives).

A trial was initiated with a warning bell and a fixation cross appearing for 500ms in the middle of the screen. Subsequently, the visual domain appeared. After they had described the target, participants pressed the Enter key on their keyboard to move to the next trial.

Trials were presented in two blocks to allow participants to take a break. Speech onset time was measured using the DMDX voice trigger from the point when the visual domain was presented to the point when a participant began to speak.

Data pre-processing

Descriptions were transcribed and annotated for whether they contained size, colour or both. Descriptions in the S condition which contained both size and colour were classified as overspecified. Descriptions in the S condition which contained only colour, or those in the CS condition which contained only one of the two properties, were classified as underspecified. All other descriptions were classifed as well-specified. Data from two participants was excluded because they produced utterances which compromised the calculation of speech onset time (for example, starting all of their descriptions with *I see a...*). The remaining 38 participants produced well-specified descriptions 71% of the time, with 27% overspecified descriptions and 2% underspecifications. The relative frequency of over- compared to underspecifications is to be expected, given previous work (see the Introduction).

Speech onset times were manually tuned using CheckVocal (Protopapas, 2007), a program for the detection and correction of voice key mistriggers (due to lip smacks, coughs, background noise etc) in DMDX result files. For each sound file, we ensured that the speech onset time was taken at the precise point where the participant's description began. In case the description included a determiner, this meant the onset of the determiner. In case a description began with a hesitation (e.g. *uhhhh het kleine rode bed*), the onset time was still the onset of the description, that is, following the initial hesitation.

Following tuning, an onset time was defined as an outlier if it exceeded the mean ±2SD in its condition. 106 data points (4.4%) were considered outliers by this criterion and were treated as missing.

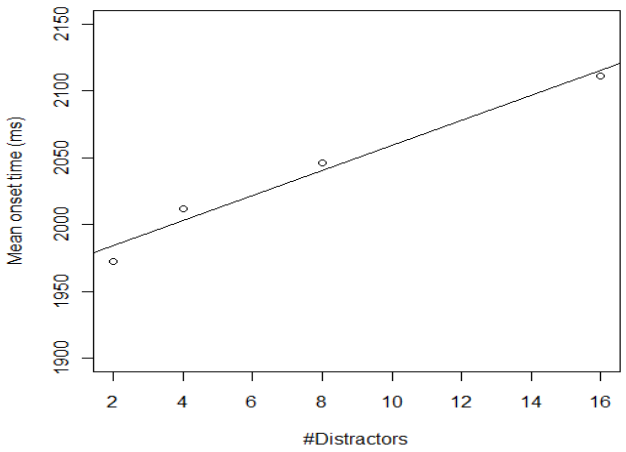

Figure 2: Mean onset times by number of distractors.

Results

In what follows, we report results based on descriptions which were well-specified, excluding over- and underspecified cases.[3] This is because in the case of underspecified descriptions, participants presumably did not check against the distractor set to see whether a selected property combination was distinguishing; where participants overspecified, the inclusion of a redundant property may not have involved such a check because it was extra information.

Table 1 displays mean speech onset times in each Distractor condition for different levels of Property as well as overall, while Figure 2 displays the relationship between onset times and Distractors. The means show an increase in speech onset time in the CS compared to the S condition. As the Figure shows, the relationship between speech onset time and domain size appears linear.

We report a linear mixed effects analysis with Properties and Distractors as fixed effects, and random intercepts for participants and items.[4] The Properties factor was scaled and centred; Distractors was treated as a continuous variable, since our aim is to model the change in

[3]The statistical tests reported here were also conducted on the full dataset; the general trends are identical to the ones reported here.

[4]A comparison of the model we report here to one with a random intercept and slope for each participant showed that the latter did not provide a better fit to the data (model $\chi^2 = 8.05, p > .5$); neither did adding a random intercept and slope for each item ($\chi^2 = 3.34, p > .9$).

	2	4	8	16	overall
CS	2022.91 (492)	2022.44 (507)	2105.93 (525)	2139.67 (809)	2073 (538)
S	1872.30 (458)	1990.95 (566)	1921.73 (209)	2046.73 (473)	1955 (506)
overall	1972.35 (486)	2011.97 (527)	2046.45 (527)	2111.16 (572)	–

Table 1: Mean speech onset times and standard deviations in each condition

speech onset times as a function of continuous increases in domain size.

There were strong main effects of both Properties ($t = -3.40, p < .001$) and Distractors ($t = 3.72, p < .001$), but no interaction ($t = .22, p > .8$). Thus, both the increase in speech onset time with two properties compared to one, and the increase with domain size are reliable.[5] To further investigate the nature of the effect of Distractors, we carried out planned comparisons by re-running a linear mixed effects analysis with Distractors as the only fixed effect and random intercepts for participants and items. For this model, Distractors was recoded as a factor using forward difference coding to perform contrasts between adjacent levels. None of these contrasts proved significant, although the difference between domains with 8 and 16 distractors approached significance ($p = .06$). Post-hoc pairwise comparisons using t-tests with Bonferroni adjustment showed a significant difference between domains with 2 distractors and those with 8 ($p = .03$) and those with 16 ($p = .004$) distractors, but no other differences. This suggests that the primary contrast is between relatively small domains and those with many more objects in the visual display.

Discussion

Our experimental results show that the time speakers take to produce references is influenced both by the number of distractors from which they have to distinguish the referent and by the number of properties that they need to include in their description in order for it to be identifying.

The effect of domain size was very strong and crucially did not interact with the number of properties required to distinguish the target referent. This is compatible with the predictions of computational REG models, in which every property needs to be checked against the distractor set in order to determine whether or not it contributes to the goal of identifying a target referent. By contrast, the literature on visual search and object identification only reports robust effects of domain size with conjunctions. In our initial discussion, we suggested that this is in part due to the difference between the task of reference production and that of object identification or online reference resolution. A speaker needs to ensure that every selected property contributes to the overall referential goal.[6] By contrast, there is no need to check contrastive value in a task involving search and identification for an object based on an instruction (e.g. Treisman & Gelade, 1980) or a definite description (as in the reference resolution experiments of Tanenhaus et al., 1995). Here, reliance on pop-up search is a good strategy for targets identified by a single feature (Treisman & Gelade, 1980) whereas, when a description is presented concurrently with a visual domain, each property in a description can be used to circumscribe the set of relevant objects (Tanenhaus et al., 1995 ; Spivey et al., 2001). Thus, the search mechanisms of speakers and listeners are qualitatively different.

A second prediction in relation to domain size was that its effect on speech onset time would be linear. Figure 2 does suggest a linear effect, although the increase becomes less steep as domain size grows larger. We also do not find differences between adjacent levels of the Distractors factor, with post-hoc analyses showing that the primary differences are between the smallest and the largest domains. Further research is required to confirm these findings. Nevertheless, the main effect, which was obtained by modelling Distractors as a continuous variable, does support the predictions of REG algorithms. It also runs counter to the predictions of models of visual search based on parallel activation of salient, contrastive properties (the third class of models discussed at the beginning of this paper), which would predict no main effect.

Finally, the finding that speakers take longer to initiate descriptions containing two properties, compared to only one, also supports the incremental search procedure in REG models, where each candidate property is checked against the distractor set. Although our experimental design does not allow us to determine whether the effect of properties is linear or not, REG models do make an interesting prediction in this regard.

Consider the effect of an incremental procedure of the

[5]The p-values for the LME models were estimated using Baayen's (2008) `pvals.fnc` function, included in the `languageR` package.

[6]The fact that speakers overspecify (that is, add properties that do not contribute to identification) may be due to the incremental nature of reference production, whereby speakers include properties one by one, starting from properties (such as colour) which are highly 'preferred' (Pechmann, 1989). Such properties may have discriminatory value individually, but may turn out to be redundant once the description has been fully formulated. Note, however, that even under this account – which is essentially the account incorporated by the incremental REG models we reviewed above (Dale & Reiter, 1995) – speakers would still check a property for its individual contrastive value.

sort we have described (Dale, 1989 ; Dale & Reiter, 1995) when it selects two or more properties. Since discriminatory value is a prerequisite for selection, every property that is selected results in a decrease in the size of the distractor set, an effect akin to the incremental domain circumscription that Spivey et al. (2001) suggest as an interpretation of their results. Thus every property that is selected leaves fewer objects against which to check the next candidate property, predicting that the effect of properties should be non-linear. This is a possibility that we intend to explore in future work, by including conditions where more than two properties are required to identify the referent.

Conclusions and future work

This paper focused on the predictions of computational models of reference production, comparing them to some well-known findings in the visual search and attention literature. We reported an experiment that showed that speakers take longer to refer to an object the more properties they require to distinguish it, and the more objects there are in the domain. Our findings lend some support to current computational models, and also highlight some important differences between the search mechanisms involved in reference production and those involved in object identification or reference resolution.

We have identified a number of avenues for future work. In the medium term, we plan to investigate the effect of domain size further in order to determine more precisely the nature of the relationship between domain size and search time. We also plan to investigate the nature of the effect of properties on search time, testing the predictions made by incremental computational models on the effect of adding properties to a description.

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 are supported by a grant from the Netherlands Organization for Scientific Research (NWO). Kees van Deemter is supported by the EPSRC Platform Grant *Affecting people with Natural Language*.

Références

Arts, A. (2004). *Overspecification in instructive texts*. Thèse de doctorat non publiée, Tilburg University.

Baayen, R. (2008). *Analyzing linguistic data*. Cambridge : Cambridge University Press.

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

Bohnet, B., & Dale, R. (2005). Viewing referring expression generation as search. In *Proc.* IJCAI'05.

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.

Deemter, K. van. (2006). Generating referring expressions that involve gradable properties. *Computational Linguistics*, *32*(2), 195–222.

Engelhardt, P. E., Baris Demiral, S., & Ferreira. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, *77*(2), 304-314.

Forster, K. I., & Forster, J. C. (2003). Dmdx: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*, 116-124.

Itti, L., & Koch, C. (2001, March). Computational modeling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194-203.

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

Olson, D. R. (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.

Protopapas, A. (2007). Check vocal: A program to facilitate checking the accuracy and response time of vocal responses from dmdx. *Behavior Research Methods*, *39*(4), 859-862.

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

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.

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

Spivey, M., Tyler, M., Eberhard, K., & Tanenhaus, M. (2001). Linguistically mediated visual search. *Psychological Science*, *12*, 282-286.

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

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.