Are we Bayesian referring expression generators?

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

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

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

Abstract

A recent paper by Frank and Goodman (2012) proposes a Bayesian model of simple referential games. One of the claims embodied in the model is that choosing which word or property to use to refer to an object depends on the utility of the property. In this paper, we compare this model to other computational models of reference production, in particular the recent PRO (Probabilistic Referential Overspecification) model. We argue that the assumption of utility that guides property choice in the Frank and Goodman (2012) model is inadequate, insofar as it ignores the possibility of overspecification and the role of preference rankings among properties, as a result of which they may be used irrespective of their utility. We show that models that do take this into account, such as PRO, have a better fit to experimental data in which participants have the possibility of overspecifying.

Background: A Bayesian model of property choice

In a recent paper, Frank and Goodman (2012) proposed a Bayesian model of simple referential games (hereafter, the FG model). The model computes the likelihood with which a listener will identify the intended referent of a description in a given context, as a function of (a) the prior probability that the referent itself would be referred to; and (b) the probability that the speaker would use a particular property for the referent.[1] It is with the second of these that this paper is concerned.

Frank and Goodman adopt a rational actor model of the speaker, whereby the probability that a speaker chooses a property for a given referent depends on its utility or informativeness. Letting p be some property of the referent r, P be the set of available properties, and $|p|$ stand for the number of objects of which p is true, the likelihood of using p, given a referent r and a context C, is given in (1).

$$P(w|r,C) = \frac{|p|^{-1}}{\sum_{q \in P} |q|^{-1}} \quad (1)$$

This makes property utility a function of surprisal. If, as suggested by the FG model, speakers compute such quantities, their property choices will be maximally informative for a listener whose task is to identify the intended referent, since a property of the referent is more likely to be selected if it is true of few (or no) other objects.

Consider, for example, a context in which there are two objects, one of which is the intended referent r. Suppose r is a blue circle, while the other object is a red circle. This makes the likelihood of using *blue* higher (at 0.67) than the likelihood of using circle (0.33), because r is the only object with the property *blue*. Thus, this calculation assumes that speakers will prioritise those properties that are most discriminatory for the referent in context.

Frank and Goodman tested the part of the model in (1) against data collected in language games where participants had to select ('bet on') which of two properties a speaker would choose to describe one of the objects in a given context. The task had the following characteristics:

1. Speakers could only choose one property for a given referent;
2. The options to choose from were verbalised in advance, for example: *Which word would you use,* blue *or* circle*?*.

The predictions of the model were found to correlate very highly with actual choices made in the experiment. Note, however, that participants did not have the possibility of choosing more than one property.

Referring Expression Generation

The aims of the FG model are similar in certain respects to those of a number of models proposed in the computational literature on Referring Expression Generation

[1] Frank and Goodman (2012) refer to 'words' rather than 'properties'. For the sake of generality, and in line with the terminology used in the Referring Expressions Generation literature, we refer to properties throughout. Note that nothing in the present discussion of the model hinges on this distinction.

algorithms (REG; see Krahmer & van Deemter, 2012, for an exhaustive review). One of the tasks of these algorithms is to choose properties for an intended referent which jointly distinguish it from its distractors (the other objects in context). In the REG literature, two trends stand out in particular. Some algorithms prioritise brevity, seeking to produce descriptions which contain as little extra information as possible beyond the identification requirement, by taking into account their discriminatory power. More recent work, however, has shifted the focus from discriminatory power to other factors that affect speakers' property choices, chief among these being a property's psychological salience or degree of 'preference'.

Brevity and discriminatory power

In early work on REG, algorithms sought to find the shortest possible referring expression (the smallest possible set of properties) to distinguish the referent. For example, the Full Brevity algorithm (Dale, 1989) searches through possible descriptions in order of length, until a distinguishing one is found.

Such heuristics were often motivated by an appeal to the Gricean Maxim of Quantity (Grice, 1975), under the assumption that more information than strictly required to identify the referent would give rise to unintended implicatures. However, finding the shortest possible description turns out to be intractable in the worst case, because it amounts to an exhaustive search through all available subsets of properties of the referent (Appelt, 1985; Reiter, 1990). In order to avoid this complexity, an approximation was proposed by Dale (1989) in the form of a Greedy Algorithm. Rather than searching through the space of possible descriptions exhaustively, this algorithm proceeds incrementally, adding a single property at a time to the description. At each stage, it considers the one that has the highest discriminatory value, that is, excludes the greatest number of the remaining distractors. The procedure terminates when a distinguishing description has been found, or all properties have been considered.

To continue with our earlier example, if r is a blue circle and there is one other object, a red circle, the Greedy Heuristic will first consider the property *blue* (because it is more discriminatory than *circle*) and add it to the description. Since this fully distinguishes r from the other object, the procedure terminates.

Property preference

More recent algorithms, starting with Dale and Reiter (1995), have maintained the incremental approach to property choice, but have de-emphasised the role of discriminatory power in favour of other factors in property selection. Dale and Reiter's Incremental Algorithm (IA) considers properties for inclusion primarily as a function of the degree to which they are psychologically 'preferred' or 'salient'. It achieves this by assuming that properties are arranged in a fixed linear order, determined by their degree of preference, and then traversing this list, checking each property in turn and including it in the description if it excludes at least one distractor. As with the Greedy heuristic, the procedure terminates when a description is fully distinguishing or there are no properties left to consider.

Suppose, for example, that shape properties are assumed to be more salient or preferred than colour properties. Then, in our running example, the IA considers *circle* first; however, this property does not exclude any distractors (both objects are circles) and is therefore not included. The algorithm next considers *blue* and selects it. In this case, the outcome is similar to that of the Greedy Heuristic. Note, however, that the algorithm can overspecify (i.e. include more information than strictly required for identification) in some cases. Suppose that in addition to the two circles in the example, there is also a red square. Then, on first considering the shape property *circle*, the IA would include it because it now excludes one distractor; the eventual outcome is the description *blue circle*, even though in this context too, *blue* would have sufficed.

The preference-based heuristic was inspired by a large body of psycholinguistic evidence showing that speakers overspecify by including highly preferred properties, especially an object's colour (Pechmann, 1989; Eikmeyer & Ahlsèn, 1996; Belke & Meyer, 2002; Arts, 2004; Engelhardt, Bailey, & Ferreira, 2006, among many others). Such properties tend to be used by speakers even when they appear to violate the Gricean Maxim of Quantity; indeed, these findings go against the grain of early psychological theorising about reference which, like the early REG models, emphasised discriminatory value (e.g. Olson, 1970). Recent comparative evaluations of REG algorithms have also found that those incorporating preference-based heuristics match speaker behaviour better, compared to those that emphasise discriminatory power (Gatt & Belz, 2010; van Deemter, Gatt, Sluis, & Power, 2012).

The Frank and Goodman model and REG

The classic REG algorithms reviewed above are entirely deterministic, in that, given a particular context and a referent, they always return the same description. This makes them poor models of human speaker behaviour (van Deemter, Gatt, van Gompel, & Krahmer, 2012). Frank and Goodman's approach makes property choice probabilistic and is therefore arguably more compatible with what is observed in experimental settings.

However, there is one crucial respect in which the underlying assumptions of the FG model seem to go against the psycholinguistic evidence on reference production, in that it ignores preference-based heuristics. Indeed, modulo its non-determinism, Frank and Goodman's model

(a) Size condition

(b) Colour condition

(c) Colour or size condition

Figure 1: Experimental domains

of the speaker is related to the first approach to REG described above, since utility is dependent on discriminatory power.

By contrast, evidence for overspecification would not only argue against a strict interpretation of the Gricean communicative maxims in reference production, but also suggests that speakers may not be 'rational' agents whose choices are solely guided by utility. Rather, speakers may be relying on simple, preference-based heuristics that override slower cognitive processes (e.g. Tversky & Kahneman, 1982). Interestingly, this seems to be the case even though overspecification may actually be detrimental to listeners, insofar as properties which aren't required for identification may slow down a listener's search for the intended referent based on the incoming description (Engelhardt, Baris Demiral, & Ferreira, 2011).

This discussion raises the question whether there could be a better approach to the speaker's choice problem in reference production, one that combines the virtues of non-determinism with more psycholinguistically plausible heuristics that do not rely exclusively on utility or discriminatory value. In the next section, we describe a model that was developed to take both of these issues into account.

The PRO Model

The PRO (Probabilistic Referential Overspecification) model (van Gompel, Gatt, Krahmer, & van Deemter, 2012) was designed on the basis of experiments with Dutch and English speakers, and sought to give a more precise account of referential choices, combining preference-based heuristics and non-determinism (see Gatt, van Gompel, Krahmer, & van Deemter, 2011, for full details of the experiment).

In the experiments, participants were engaged in a simple director-matcher task in which one participant, the speaker, described a target referent from a group of three objects, to a listener who could see exactly the same objects. Speakers were instructed to describe objects in a way that would allow the listener to identify the intended referent.

Participants were exposed to trials in three different conditions, as shown in Figure 1. Each trial consisted of pictures of three objects, one of which was marked as the intended referent which a speaker had to identify for her listener. In the trials, objects could be described on the basis of two properties, size and colour, in addition to their type. These two properties were chosen because a clear distinction has been found between them in the literature (e.g. Eikmeyer & Ahlsèn, 1996; Belke & Meyer, 2002): colour is highly preferred and tends to be used even when not required; size, in contrast, only tends to be used if absolutely required, presumably because it is a relative property.[2]

Trials were constructed so that either colour alone (C; Figure 1(a)), or size alone (S; Figure 1(b)) sufficed to identify the target referent. In each condition, the other property was also discriminating, but did not identify the referent completely, because it only excluded one distractor. In a third, baseline condition (C/S; Figure 1(c)) either property was fully discriminatory. The experiment was replicated with both Dutch and English speakers.

The proportions of choices of referring expressions are shown in Table 1. As expected, they show a clear preference for colour. For instance, in the S condition, where colour is not required, most speakers (78% in Dutch, 80% in English) opted for a description containing both colour and size. By contrast, in the C condition there are relatively few overspecified descriptions containing both properties (10% in Dutch, 8% in English). In the C/S condition, the majority of speakers opted for a colour-

[2]Type was not considered in this experiment, under the assumption that speakers will always select an object's type in constructing a description; see for example Pechmann (1989) and Dale and Reiter (1995). Note, however, that a second experiment was also conducted, similar to the one reported here, in which the discriminatory value of type and colour were manipulated and size played no role. In conditions where type alone sufficed to identify the referent, most speakers (70%) produced descriptions containing only this property. In conditions where colour alone sufficed, over 90% of speakers produced descriptions containing both type and colour, in line with our assumption in the experiment summarised here. The PRO model described below was also fitted to data from these experiments, with predictions within 3% of observed proportions. See van Gompel et al. (2012) for details.

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

		Description		
		Colour only	**Size only**	**Colour and 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) (baseline)	Dutch	70.8 (266)	3.7 (14)	25.5 (96)
	English	79.1 (280)	3.7 (13)	17.2 (61)

only description.

Modelling property choice

The PRO model, shown schematically in Figure , assumes that speakers always first select the property that is fully discriminating (size in Figure 1(a), colour in 1(b)). If there is more than one such property (as in Figure 1(c)), then one property is initially selected probabilistically, according to preference. After they add the first property, speakers may add a second property. Once again, the probability of doing this depends on the property's degree of preference.

Thus, PRO combines both discriminatory power and preference, and does so non-deterministically. The model has two parameters: x is a maximum-likelihood estimate of the probability of using colour (its 'preference'), while y is an additional parameter that represents the likelihood or 'eagerness' to overspecify on the part of a speaker.

The predictions of PRO were compared to the data obtained in our experiments. The model predicted proportions for all expressions in the three conditions within 2% of the observed frequencies (it also accounts for speech repairs where speakers initially underspecify, but subsequently add a second property; see van Gompel et al., 2012). As an example, Table 2 displays predictions for the English data.

Consider now the predictions of the FG model in relation to the data presented above. Recall that the model does not take into account the possibility of overspecification: the experimental task against which the predictions of (1) were compared only allowed participants to choose one property for an object. However, it is instructive to look at the predicted probabilities for the use of each property in each of the experimental conditions in Figure 1. These are shown in Table 3.

The main observation here is that the FG model systematically overestimates the probability of using a dispreferred property (size) and underestimates the probability of using a preferred one (colour). For example, in the S condition depicted in Figure 1(a), *green* refers to two objects, while *large* refers to only one. The model predicts a probability of 0.33 of using *green*, with a probability of 0.67 of using *small*. In the C/S condition, where either size or colour suffices to distinguish the object (Figure 1(c)), the model predicts that either property is equally probable (0.5). These predictions are clearly incompatible with the data, where 80% of the English descriptions in the S condition included both colour and size, while the use of colour in the C/S condition is well above chance.

Table 3: Predicted probability by the Frank and Goodman model of using colour or size in each experimental condition

	Colour	**Size**
Size Sufficient (S)	0.33	0.67
Colour sufficient (C)	0.67	0.33
Colour or size (C/S)	0.5	0.5

Discussion

Frank and Goodman's Bayesian model assumes a model of the speaker as a rational agent whose choices in a reference task are based on utility. As we have argued, this assumption is compatible with that made by certain Referring Expression Generation algorithms. By contrast, models such as the Incremental Algorithm (Dale & Reiter, 1995) or PRO emphasise the role of preferences rather than utility or discriminatory power. When their predictions are compared on the same experimental data, it turns out that, while PRO predicts speaker choices quite accurately, the Bayesian model does not, because it misrepresents the extent to which a property is preferred over others.

One reason for the mismatch may be that Frank and Goodman tested their model against data from communicative situations which are different from the referential contexts in our experiment. Specifically, participants in their experiments were asked to select exactly one property out of two and were told which words to use ('blue or circle?'). This may have caused participants to explicitly evaluate the properties and choose the one that rules out most distractors. This was not the case in the experiment summarised above, where speakers were simply instructed to identify the target referent for their

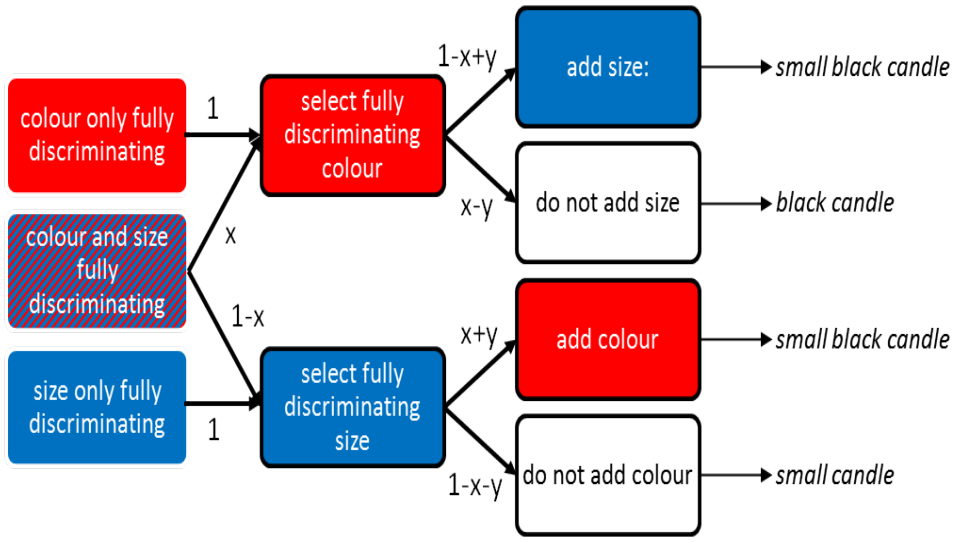

Figure 2: The PRO Model; x = probability of selecting colour; y = ovserspecification likelihood. Using $x = .8673$ and $y = -.0531$, PRO fits the English data almost perfectly. The same is true of the model fitted to the Dutch data.

Table 2: PRO predictions for each description type, by condition. Parameters: $x = .8673$; $y = -.0531$

	Description		
	Colour only	**Size only**	**Colour and size**
Size sufficient (S)	0	0.19	0.81
Colour sufficient (C)	0.92	0	0.08
Colour or size (C/S)	0.80	0.02	0.18

listeners and had no *a priori* limit on how much information to include.

However, a more crucial difference between the two models is that, while PRO explicitly seeks to model preference-based heuristics, Frank and Goodman do not consider this, excluding the possibility that properties selected by speakers may be redundant for the purposes of identification. We have argued that this is a weakness of the model because the robust psycholinguistic and computational findings in this regard suggest that a completely rational, utility-based account cannot be an accurate model of the speaker.

If speakers systematically include properties that are not 'useful' for identification – that is, do not contribute to the discriminatory value of an identifying description – then clearly they are not making choices purely on the basis of utility. Rather, as we have suggested above, these 'preferences' suggest that (at least in visual domains of the sort that we and Frank and Goodman have explored), choices are also made on the basis of simpler heuristics, relying on perceptual or cognitive salience. Thus our conclusion is that, while any psychologically realistic model of reference production has to be non-deterministic (as Frank and Goodman suggest), the assumption that speaker behaviour is guided by a utility-based heuristic is probably false.

References

Appelt, D. (1985). Planning english referring expressions. *Artificial Intelligence*, *26*(1), 1–33.

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

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.

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the 27th annual meeting of the association for computational linguistics (acl'89)* (p. 68-75).

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

Eikmeyer, H. J., & Ahlsèn, E. (1996). The cognitive process of referring to an object: A comparative study of German and Swedish. In *Proceedings of the 16th scandinavian conference on linguistics.*

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.

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

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Gatt, A., & Belz, A. (2010). Introducing shared task evaluation to nlg: The TUNA shared task evaluation challenges. In E. Krahmer & M. Theune (Eds.), *Empirical methods in natural language generation.* Springer.

Gatt, A., van Gompel, R., Krahmer, E., & van Deemter, K. (2011). Non-deterministic attribute selection in reference production. In *Proceedings of the workshop on production of referring expressions: Bridging the gap between empirical, computational and psycholinguistic approaches to reference (PreCogSci'11).*

Grice, H. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics: Speech acts.* (Vol. III). New York: Academic Press.

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.

Reiter, E. (1990). The computational complexity of avoiding conversational implicatures. In *Proc. 28th annual meeting of the association for computational linguistics.*

Tversky, A., & Kahneman, D. (1982). Judgement under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

van Deemter, K., Gatt, A., Sluis, I. van der, & Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, *36*(5), 799-836.

van Deemter, K., Gatt, A., van Gompel, R., & Krahmer, E. (2012). Towards a computational psycholinguistics of reference production. *Topics in Cognitive Science*, *4*(2), 166-184.

van Gompel, R., Gatt, A., Krahmer, E., & van Deemter, K. (2012). Pro: A computational model of referential overspecification. In *Proceedings of the conference on architectures and mechanisms for language processing (*AMLAP *'12).* Trento, Italy.