TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 4 (2012) 166–183

ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2012.01187.x

Toward a Computational Psycholinguistics of Reference Production

Kees van Deemter,[a] Albert Gatt,[b] Roger P.G. van Gompel,[c] Emiel Krahmer[d]

[a]*Department of Computing Science, University of Aberdeen*
[b]*Institute of Linguistics, University of Malta, Tilburg Centre for Cognition and Communication (TiCC), Tilburg University*
[c]*School of Psychology, University of Dundee*
[d]*Tilburg Centre for Cognition and Communication (TiCC), Tilburg University*

Received 2 April 2010; received in revised form 20 January 2011; accepted 20 February 2011

Abstract

This article introduces the topic ''Production of Referring Expressions: Bridging the Gap between Computational and Empirical Approaches to Reference'' of the journal *Topics in Cognitive Science*. We argue that computational and psycholinguistic approaches to reference production can benefit from closer interaction, and that this is likely to result in the construction of algorithms that differ markedly from the ones currently known in the computational literature. We focus particularly on determinism, the feature of existing algorithms that is perhaps most clearly at odds with psycholinguistic results, discussing how future algorithms might include non-determinism, and how new psycholinguistic experiments could inform the development of such algorithms.

Keywords: Referring expressions; Computational models; Psycholinguistic experiments; Non-determinism in language production; Overspecification

1. Introduction

Throughout the last century, reference has been studied in all academic disciplines that study language and communication, varying from theoretical linguistics and philosophy to language acquisition, psycholinguistics and computational linguistics. This is not surprising, given the centrality of reference in human communication: people can only exchange information about an object if they agree about the identity of the object first. Thus, when

Correspondence should be sent to Kees van Deemter, University of Aberdeen, Aberdeen, UK. E-mail: k.vdeemter@abdn.ac.uk

children acquire language, one of the first things they learn is how to refer to objects (Bruner, 1983; Matthews, Lieven & Tomasello, 2007), and when computers produce language, the generation of referring expressions is, invariably, one of the key tasks they perform (Reiter & Dale, 2000). The philosopher John Searle defined reference as follows:

> Any expression which serves to identify any thing, process, event, action, or any other kind of individual or particular I shall call a referring expression. Referring expressions point to particular things; they answer the questions Who?, What?, Which? (Searle, 1969, pp. 26–7)

Searle acknowledged that this definition is imprecise (e.g., if you sign your name, do you refer to yourself?), but he suggests that the proper approach is ''to examine those cases which constitute the center of variation of the concept of referring and then examine the borderline cases in light of their similarities and differences from the paradigms.'' The current issue of the journal *Topics in Cognitive Science*, to which the present article serves as an introduction, is consistent with Searle's methodological remarks. More specifically, it uses definite descriptions (i.e., noun phrases of the form ''the …'') as its paradigm of reference, while including other expressions—such as pronouns and proper names—to the extent that they function in the same way.

In this article, we shall examine some of the main results and methods of two of the most prolific research traditions to study reference, namely *psycholinguistics* and *computational linguistics*, focusing on the *production* of referring expressions. Research in these two traditions has so far proceeded with little mutual influence. *Psycholinguistic* work on reference production seldom references computational work; although it has given rise to interesting ideas (e.g., concerning common ground and alignment), the resulting models are in many ways still sketchy and imprecise, as we shall see. *Computational* work on reference production occasionally mentions psycholinguistic studies; moreover, algorithms in this area have recently come to be tested using methods that owe much to the psycholinguistic tradition. However, while the resulting algorithms are useful for practical applications, they are unsatisfactory as models of linguistic behavior, as we shall argue. This article aims to *bridge the gap* between computational and psycholinguistic approaches to reference production, arguing that each discipline can benefit substantially by taking some of the insights and methods of the other tradition on board. The substance of our argument will focus on the related issues of referential overspecification and non-determinism. Before we get there, we shall sketch briefly how reference production has been studied in each of the two research traditions.

2. Computational approaches to reference production

The computational production of referring expressions is a key component of many Natural Language Generation (NLG) systems (Mellish et al., 2006; Reiter & Dale, 2000). Work on algorithms for Referring Expression Generation (REG) started when computer programs such as SHRDLU (Winograd, 1972) needed an ability to identify objects to human users, for

example, to answer their questions (e.g., User: ''Which block supports the table?'' System: ''The red pyramid.'') Research on REG started in earnest in the 1980s, focusing on the underlying intentions that an agent had in producing referring expressions (Appelt, 1985; Kronfeld, 1990). Inspired by the work of Searle, Austin, and others, Appelt and Kronfeld argued that a referring expression should be seen as part of a larger speech act. In their approach, the generation of a referring expression was part of planning and generating an entire utterance. More recent work has focused on the generation of definite, identifying noun phrases, as a separate NLG task (Dale, 1989; Dale & Reiter, 1995). Since the work of Dale and Reiter, a significant consensus has arisen on the nature of the problem that these REG algorithms are designed to solve, with many approaches subscribing to some version of the following task definition, which emphasizes *content determination* and unambiguous identification:

> Given a domain of discourse, consisting of entities and their properties, and a target referent, find a set of properties (the description) which uniquely distinguishes the target referent from its distractors.

The above definition does not specify how the current domain of discourse is to be determined. It also does not tackle other communicative intentions that could underlie the production of a definite reference, beyond identification (*pace* Jordan, 2002; Jordan & Walker, 2005). Reference, in this tradition, is most often understood as a ''one-shot'' affair, where the referring expression in question cannot rely on information in previous utterances. Exceptions exist, where REG work has started to address the generation of anaphoric referential NPs (see Callaway & Lester, 2002; Krahmer &Theune, 2002; McCoy & Strube, 1999; Passonneau, 1996; Stoia, Shockley, Byron, & Fosler-Lussier, 2006, among others). Heeman and Hirst (1995) went further, by proposing a computational model which addresses the types of collaboration observed in dialog by Clark and Wilkes-Gibbs (1986). Heeman and Hirst's work is rooted in a long tradition in Artificial Intelligence, which seeks to understand dialog using models of rational agency based on communicative intentions and mutually held beliefs (e.g., Allen & Perrault, 1980; Cohen & Levesque, 1991).

An important theoretical influence on many REG algorithms has been the work of Grice (1975), whose Maxim of Quantity was originally interpreted in the context of referring expressions as a constraint to include no more information than is required for a distinguishing description (e.g., Appelt, 1985; Dale, 1989). This *Full Brevity* interpretation was compatible with early psycholinguistic theorizing (e.g., Olson, 1970). Given a domain such as Table 1, a Full Brevity algorithm would yield the minimally distinguishing description

Table 1
An example domain

Entity	Type	Color	Size
e_1	*dog*	*black*	*large*
e_2	*dog*	*white*	*small*
e_3	*dog*	*black*	*small*

{*large*} for e_1 (this description could be realised as *the large one*). Dale and Reiter's (1995) later Incremental Algorithm (IA), which had roots in the work of Winograd and Appelt, relaxed this constraint. The IA, summarized in Box 1, is based on the finding that some properties of a referent (e.g., its COLOR) are more likely to be included in a description than others, even when this leads to *overspecified* referring expressions, which are longer than necessary for identification (e.g., Pechmann, 1989). In the IA, the likelihood of selection is determined by a fixed *preference order* of properties, with more preferred properties being more likely to be selected. Overspecification occurs because the algorithm never withdraws properties once they have been selected. The IA is still an influential REG model, which has informed many subsequent developments (see Krahmer & van Deemter, 2011, for a review). Various papers in the present journal issue build on this algorithm directly, for example, by investigating how the fixed preference order of the IA interacts with other factors, such as alignment with a property used in the previous utterance (Goudbeek & Krahmer, 2012), or in what way normally preferred properties, such as color, can lose their preferential status in settings where they prove to have low utility or when, because of a craftily constructed experiment, the hearer cannot be relied upon to see the same colors (Guhe, 2012).

Many REG algorithms rely on the implicit assumption that generated references should be humanlike. This goal of emulating human behavior is particularly evident in the way such algorithms are typically evaluated. As in many other areas of NLP, a typical REG evaluation scenario involves the comparison of automatically generated referring expressions to a set of human-produced ''reference'' outputs in a corpus (van Deemter, Gatt, van der Sluis & Power, in press; Gupta & Stent, 2005; Jordan & Walker, 2005; Viethen & Dale, 2007). Such evaluations usually incorporate a metric which yields a global score reflecting the degree of match between an algorithm and one or more individual participants. Averaged over an entire corpus, such metrics typically ignore the variation within the corpus and the extent to which an algorithm ''agrees'' with a subset of the participants represented in the corpus.

Humanlikeness, of course, is not the only conceivable goal for a REG algorithm. An alternative is to aim for *effectiveness*, by generating descriptions that are easy for a comprehender to understand and/or resolve (e.g., Paraboni, van Deemter, & Masthoff, 2007), and such a comprehension-oriented perspective is of great importance in practical applications. The two goals of humanlikeness and effectiveness are not necessarily compatible, since human speakers do not always produce optimal expressions (Dale & Viethen, 2010; Gatt & Belz, 2010; Oberlander, 1998). We shall return to this issue at the end of the next section.

3. Psycholinguistic approaches to reference production

Production has been studied extensively by psycholinguists as well. In this section, we focus on two areas which are of particular relevance not only to psycholinguistic but also to computational models, namely *overspecification and ambiguity avoidance* and *interaction and audience design*.

Computational algorithms have sometimes been inspired by psycholinguistic work. An example is Dale and Reiter's reliance on the work of Pechmann (1989) (see also Levelt, 1989) in connection with the observation that human speakers sometimes overspecify their references, and hence so should REG algorithms. The theme of overspecification has again arisen in more recent debates concerning the question whether and how speakers ensure that their referring expressions are maximally interpretable to the addressee. It has become clear, for example, that in some situations, speakers tend to produce ambiguous or underspecified referring expressions. For example, Ferreira, Slevc, and Rogers (2005) found that speakers will use the expression *the bat* (for a flying mammal) even when a baseball bat is also present, making the description ambiguous. Khan and colleagues observed that syntactic ambiguities (as in ''the old men and women,'' whose syntax leaves it unclear whether it denotes all women or just the old ones) are frequent as well, and proceeded to investigate the conditions under which such surface-ambiguities are resolvable by the hearer (Khan, van Deemter, & Ritchie, 2012).

*Over*specification appears to occur more frequently, though. Recent studies have suggested that redundant information is frequent in the referring expressions produced by people (e.g., Arts, 2004; Engelhardt, Bailey, & Ferreira, 2006). This is consistent with the Incremental Algorithm, which predicts overspecification in specific situations. In fact, the IA makes some precise predictions concerning overspecification: assuming that color is preferred over size (Box 1), then a generated description can include color as an overspecified property, but not size (after all, if color were sufficient to distinguish a target, the IA would select color and then terminate, so it would not consider less preferred properties). This is an interesting prediction that has never been tested experimentally, as far as we know.

Many factors are known to affect speakers' tendency to overspecify. It has been shown, for example, that speakers overspecifiy more frequently in fault-critical situations, that is, where confusion would tend to cause problems (Arts, 2004; Arts, Maes, Noordman, & Jansen, 2011). Paraboni et al. (2007) showed how speakers use overspecification to identify objects in large domains, and how readers benefit: If there is only one photocopier in a building, located on a different floor, then simply directing a new member of staff to *the photocopier* would be unhelpful; *the photocopier, on the 2nd floor, opposite the elevator* is much better. Koolen, Gatt, Goudbeek, and Krahmer (2011) found that speakers overspecify more often when referring to ''complex'' targets, for example, when referring to persons rather than furniture items and when referring to plural rather than singular targets. All these results suggest that there may not be a single reason why speakers overspecify their references. Be that as it may, overspecification has to be used with caution: Engelhardt et al. (2006), for example, found that although listeners do not rate overspecified references as worse than minimal ones, eye-tracking data suggest that overspecified descriptions (*Put the apple on the towel in the box* where *the apple* would suffice) can be confusing for listeners when they cause syntactic ambiguity. More recently, Engelhardt and colleagues presented evidence that even overspecification with properties such as color, which is realized prenominally and does not give rise to syntactic ambiguity in the NP, result in hearers taking longer to resolve references in visual domains (Engelhardt, Demiral, & Ferreira, 2011).

The avoidance of misunderstandings is, of course, highly relevant for determining whether a target should be referred to using a full description or a reduced form. A pronoun provides less information about its antecedent than a full description does, so pronouns are predicted to be used when their antecedents are highly salient. Various proposals have been put forward, linking the form of a referring expression to the saliency of the referent in the discourse (see e.g., Ariel, 2001; Arnold, 2001; Givon, 1983; Stevenson, Crawley, & Kleinman, 1994). Recently, Arnold and Griffin (2007) showed that speakers produced fewer pronouns (and more names) when an additional character was present in the discourse than when it was not, suggesting that the presence of the additional character competes with, and therefore, reduces the saliency of the target referent. Fukumura, van Gompel, and Pickering (2010) showed that the presence or absence of an additional character in the visual context has a similar effect, supporting the idea that not only saliency in the discourse but also saliency in the visual context affects the choice of referring expression.

A large body of work now shows the limitations of viewing reference as a ''one-shot'' affair. Clark and colleagues (Brennan & Clark, 1996; Clark & Murphy, 1983; Clark & Wilkes-Gibbs, 1986), for example, have shown that references change during the course of an interaction, becoming more reduced and eventually converging on a single description for a referent. They interpret these findings in terms of a collaborative process, whereby speaker and addressee both converge on a descriptive form for a referent (referred to as a ''conceptual pact'' by Brennan & Clark, 1996). A more general account of adaptation in dialog is the Interactive Alignment Model of Pickering and Garrod (2004), which claims that dialog participants may align their linguistic representations at all levels of interaction, ranging from alignment of phonological and phonetic categories (Bard & Aylett, 2004) up to lexical and syntactic choice. Many authors see alignment as the result of a mechanistic, largely automatic process where speakers produce expressions that are easy to comprehend for their addressee because they rely on representations that were used (''primed'') earlier on in the interaction. We do not know of any attempts to model the priming mechanisms outlined by Pickering and Garrod in computational REG, though they may be discerned in other subtasks of Natural Language Generation (e.g., Buschmeier, Bergmann, & Kopp, 2010; Purver & Kempson, 2004).

While the importance of adaptation and audience design is not disputed, the *extent* to which speakers are capable of taking the addressee into account is a matter of intensive debate, with some researchers arguing that speakers in fact find it hard to do this, for example, when some subset of a referential domain is in a speaker's privileged ground, rather than in common ground (Horton & Keysar, 1996; Keysar, Lin, & Barr, 2003; Wardlow Lane, Groisman, & Ferreira, 2006). These authors would argue that speakers frequently plan their utterances ''egocentrically.'' This conclusion has sometimes been questioned (Brennan & Hanna, 2009; Brown-Schmidt, 2009). Heller and colleagues add to these questions by re-examining the data of an earlier study by Wu and Keysar (2007), Heller, Skovbroten, and Tanenhaus (2012), and Wu and Keysar (2007). When subjects in a new experiment were given the utterances from the earlier study to listen to, they were generally able to determine whether a given item of information was privileged. In other words, where Wu and Keysar had found expressions (e.g., names) that they regarded as egocentric, these expressions may

actually have been produced for good reasons (e.g., to ''teach'' the hearer the name). We assume that the last words on this debate have not yet been said.

4. Toward a computational psycholinguistic theory of reference

The previous two sections have illustrated how computational and psycholinguistic research on the production of referring expressions are often seen as separate areas, with only limited interaction between them. Yet each of the two areas suffers from limitations which the other would be well placed to rectify. Clearly, as we have seen, there are various aspects of referential behavior which current REG algorithms ignore, but which have received a lot of attention among psycholinguists. Conversely, psycholinguistic theories often rely on intuitive notions such as ''common ground,'' ''adaptation,'' ''alignment,'' or ''salience,'' without defining these precisely. A study by Poesio, Stevenson, Eugenio, and Hitzeman (2004) shows how a computational approach can be useful in such cases. In discussing the Centering model of discourse anaphora (Grosz, Joshi, & Weinstein, 1995), these authors demonstrated the extent to which the underlying assumptions of psycholinguistic models need to be explicated. Psycholinguistic experimentation (e.g., Brennan, 1995; Gordon & Hendrick, 1999; Gordon, Kendrick, Ledoux, & Yang, 1999) has suggested that the preference for a pronoun over a name in both production and comprehension is affected by factors such as the salience of the antecedent and in which utterance it occurred. However, the notions of ''salience'' and ''utterance'' have remained vague. This has given rise to several parameters in the Centering model. Poesio and colleagues argued that while ''the best way to test such preferences is through behavioral experiments,'' this is in practice difficult because of ''the enormous number of possible ways of setting the theory's parameters'' (Poesio et al., 2004, p. 310). They, therefore, set about testing several of these alternative parameter settings computationally, using an annotated corpus to test different versions of the theory, and explicitly formalizing several hitherto underspecified parameters in the process.

Even though there is undoubted scope for increased collaboration between practitioners in the psycholinguistic and computational camps, we shall see that this requires a re-evaluation of the goals that REG algorithms are designed to achieve, as well as a different focus in psycholinguistic studies.

4.1. Goals of computational algorithms and their relevance for psycholinguistics

As we have seen, the exact goal of REG algorithms, as these are presented in the literature, is often unclear. Dale and Reiter's (1995) Incremental Algorithm is a good example. On the one hand, the authors argued that one way of creating computational models is to ''determine how speakers generate texts and build an algorithm based on these observations (the Incremental Algorithm Interpretation)'' (Dale & Reiter, 1995, p. 252) and, consequently, their Incremental Algorithm is often understood as aiming to produce referring expressions that resemble those that speakers produce. Yet they state: ''The argument can be made that psychological realism is not the most important consideration for developing algorithms for

embodiment in computational systems; in the current context, the goal of such algorithms should be to produce referring expressions that human hearers will understand, rather than referring expressions that human speakers would utter'' (Dale & Reiter, 1995, p. 253). The ambiguity of their goal also shines through when they write: ''The fact (for example) that human speakers include redundant modifiers in referring expressions does not mean that natural language generation systems are also required to include such modifiers; there is nothing in principle wrong with building generation systems that perform more optimizations of their output than human speakers. On the other hand, if such beyond human-speaker optimizations are computationally expensive and require complex algorithms, they may not be worth performing; they are clearly unnecessary in some sense, after all, since human speakers do not perform them'' (Dale & Reiter, 1995, p. 253).

The ambiguities surrounding the aim of REG models raise significant problems for evaluating such models. The goal of humanlikeness would call for comparison against corpora or against the results of language production experiments (e.g., van Deemter et al., in press; Gatt & Belz, 2010; Jordan & Walker, 2005; van der Sluis & Krahmer, 2007; Viethen & Dale, 2007). By contrast, the goal of producing expressions that are easiest to understand (Paraboni et al., 2007) would tend to make reading (e.g., self-paced reading or eye movements during reading; Almor, 2000; Garrod, Freudenthal, & Boyle, 1994; Gordon, Grosz, & Gilliom, 1993) or auditory language comprehension paradigms (e.g., recording of eye movements while people identify objects in a visual scene; Brown-Schmidt, 2009; Sedivy, Tanenhaus, Chambers, & Carlson, 1999) the evaluation methods of choice.

4.2. The psychological reality of computational algorithms

Even granted that REG algorithms do not seek to model the actual language production *process*, but only its output, there are aspects to these algorithms that make them psychologically implausible. Perhaps the most striking property of most computational algorithms that is problematic from a psycholinguistic point of view is their *determinism*: They always generate the same referring expression in a particular situation or condition. For example, in a situation where there is no other object of the same category as the target object (say, a single car), most algorithmic models either always generate minimally specified expressions (*the car*) or always generate overspecified expressions (e.g., *the red car*). But given this specific situation, they would not generate a minimally specified expression in some cases and an overspecified expression in others. This contrasts with the results from experiments with human speakers, which show that they produce various types of referring expressions in a specific condition. For example, Pechmann (1989) showed that across different speakers, both minimally specified referring expressions (on 21% of experimental trials) and overspecified expressions (on 75% of trials) were produced when there was no other object of the same category (e.g., only a single car), while underspecified expressions were also chosen on a small percentage (4%) of trials. Very similar non-deterministic results were observed by Engelhardt et al. (2006), while Dale and Viethen (2010) showed that even when referring to simple objects in simple scenes, different speakers used a large variety of referring expressions to refer to the same object, while the same speaker was likely to vary

his or her choice of referring expression considerably in very similar (or even isomorphic) scenarios. Inter-person variability is not a feature of (spoken or written) language alone: De Ruiter and colleagues, for example, report a large amount of variation between subjects in terms of the type and role of their *gestures* (de Ruiter, Bangerter, & Dings, 2012). Explanations of inter-person variation are not difficult to think of. Variability may be partly explained by children's exposure to different stimuli—compare Matthews, Butcher, Lieven, and Tomasello (2012), in this journal issue, for relevant experiments. Yet theoretical models struggle to give variation a natural place; at best, they offer a many-to-many relationship between contents and forms, allowing, in particular, that a given content can be expressed through different forms. A good example is Gundel's reformulated givenness hierarchy, which specifically allows different types of referring expressions to be associated with each level in the hierarchy (Gundel, Hedberg, & Zacharski, 2012).

It is important to note that even probabilistic REG algorithms (such as the systems described and evaluated in Gatt & Belz, 2010) are usually deterministic. These models typically use a probability distribution learned from training data to return the most probable referring expression given a particular situation. However, their output in isomorphic situations will always be the same. An exception is the model proposed by Fabbrizio, Stent, and Bangalore (2008), who proposed a probabilistic model which incorporates individual preferences for particular referring expressions, thereby altering the type of expression it generates depending on the preferences of individual speakers. Another non-deterministic model was proposed by Dale and Viethen (2010), who used different algorithms to mimic different speakers and found that this increased the correlation between the model and human responses as compared to a deterministic model. For a specific speaker, however, the output of both these models remains deterministic; that is, it is assumed that a single speaker always produces the same referring expression in a particular situation. The results of experimental studies are normally reported averaged across participants, so they do not report whether individual human speakers are deterministic. However, closer examination of the data of individual participants of almost any study reveals that their responses vary substantially, even within a single experimental condition. For example, we examined the data of Fukumura and van Gompel (2010), who conducted experiments that investigated the choice between a pronoun and a name for referring to a previously mentioned discourse entity. The clear majority (79%) of participants in their two main experiments behaved non-deterministically, that is, they produced more than one type of referring expression (i.e., both a pronoun and a name) in at least one of the conditions.

Indeed, variability in many aspects of individual behavior seems to be the rule rather than the exception. A classic example comes from ballistic research around 1900, which observed that the bullets of a skilled target shooter do not always hit the target, but pile up close to the bull's eye, with fewer and fewer strikes further away from it, giving rise to a bell-shaped probability distribution (Holden & Van Orden, 2009). Linguists have long known that language use is variable as well. Sociolinguists, for example, believe that language change and social register (e.g., idiolects associated with different social strata) cause a phenomenon known as *diglossia*, where different grammars are represented in the head of a single individual at the same time (Kroch, 2000). For each utterance, the individual is thought to

''choose'' between different grammars, where the probability of choosing a given grammar is affected by the recent history of the individual. The link with reference was made in Gibbs and Van Orden (2012), who discuss variability in speakers' pragmatic choices, including the choice how to refer (e.g., whether to express privileged information, cf. Horton & Keysar, 1996), proposing to explain these by assuming that ''the bases of any particular utterance (...) are contingencies, which are to an underappreciated extent the products of indiosyncracy in history, disposition, and situation'' (Holden & Van Orden, 2009).

The idea that human responses may best be viewed as non-deterministic, even within a single speaker, suggests that non-determinism should be an important property of a psychologically realistic algorithm. One approach to model non-determinism is exemplified by so-called roulette-wheel generation models (Belz, 2007). Rather than always generating the same, most probable output given a specific input, these models sample alternatives from a non-uniform distribution, returning outputs in proportion to their likelihood. To our knowledge, models of this kind have not yet been exploited for generating referring expressions; however, this may be a promising way to incorporate non-determinism. Another possibility would be to turn current deterministic algorithms for the generation of referring expressions into non-deterministic algorithms. Regardless of which approach is chosen, the goal should be to make quantitative, and testable, predictions.

Consider the Incremental Algorithm (IA) once again (Dale & Reiter, 1995). The original, deterministic version always generates *the cup* to refer to a black cup in the presence of a blue ashtray and yellow candle. The reason is that it assumes a fixed preference order, causing it to check the category of the object (cup) before its color (red), and since *cup* rules out both distractors, color is not tried. As we have seen, research by Pechmann (1989) suggests that speakers do produce overspecified expressions such as *the red cup* in this situation. To account for this, the IA could be revised slightly, so that color is selected first when it is a discriminating feature. But this would still not fully account for Pechmann's (1989) results, because he showed that although overspecified expressions (e.g., *the black cup*) are produced most frequently, minimally specified expressions are produced on one quarter of the trials. To account for this, the IA would need to incorporate some form of non-determinism. One possibility would be to include a random process by which the algorithm checks color before type three-quarters of the time and type before color in the remaining quarter (both across speakers and within a single speaker).

If we assume that the decision about which property is checked first is a probabilistic, non-deterministic process, then the algorithm makes interesting predictions that are relevant to psycholinguists. For example, a non-deterministic version of the Incremental Algorithm makes exact, quantitative predictions about when overspecification occurs. Although several psycholinguistic studies have shown that overspecification is common, it remains unclear under exactly what conditions it occurs and psycholinguistic models do not make clear predictions concerning this issue. We therefore, believe that the algorithm provides an important step toward a better understanding of the possible psychological mechanisms involved in overspecification.

To let us make this more concrete, assume that when referring to a small black cup in the context of a large white cup and a large red cup (so color or size can be used to uniquely

characterize the target), speakers produce *the black cup* four times more often than *the small cup*. In that case, there is a 80%-20% color-size preference (ignoring, for the sake of argument, possible overspecified expressions like *the small black cup*). According to the non-deterministic version of the Incremental Algorithm, this pattern arises because speakers first check color in 80% of cases, whereas they first check size in 20% of cases (and the category *cup* is obligatorily added, because *the black one* or *the small one* sounds awkward). Once we have determined the color-over-size preference, we can predict how often overspecification occurs in other situations. When referring to a small black cup in the context of a large white cup and a large black cup (i.e., only size is required to produce a distinguishing description), the algorithm initially chooses color over size in 80% of cases, but because this does not uniquely identify the target, it subsequently adds size, resulting in an overspecified expression (*the small black cup*) in 80% of cases. In the other 20%, it first checks size, and because this uniquely identifies the target, color will not be added. An 80%-20% split is also predicted to occur when the same target (a small black cup) occurs in a context with a small white cup and a large white cup (so color is required). In 80% of cases, color is checked first, and because this uniquely identifies the target, the algorithm produces *the black cup*. In the other 20%, size is selected first, but because it does not uniquely identify the target, color is added, resulting in *the small black cup*. Thus, the algorithm makes clear quantitative predictions that arise from the fact that color is usually checked before size. These predictions can be tested in psycholinguistic studies (Gatt, Gompel, Krahmer, & Deemter, 2011).

Other algorithms can be made non-deterministic in similar ways. For example, the Greedy Algorithm (Dale, 1989) iteratively selects the property which rules out most distractors from among those that have not yet been ruled out by the properties selected so far. But ties can occur, where two or more properties rule out the same number of distractors, and in this case the choice could be made probabilistic. When referring to a small black cup in the context of a large white cup and large red cup, for example, the revised greedy algorithm might produce *the black cup* 80% of the time and *the small cup* 20% of the time (assuming, as before, that the category—cup—is always added given that omission of the category sounds awkward). One interesting prediction is that the same 80%-20% color-over-size preference should occur when referring to e_1 in Table 2 even though a priori the color of the target rules out more distractors (e_2 and e_3) than its size (e_2). The reason is that the Greedy Algorithm first selects the property *with spoon*, because this rules out most distractors (e_3, e_4, and e_5). Next, only the distractor e_2 remains, which can be removed by either color or size, resulting in a tie. Assuming an 80%-20% color-over-size preference, people should

Table 2
Another referential domain

Entity	Type	Color	Size	Spoon
e_1	*cup*	*black*	*small*	*with*
e_2	*cup*	*white*	*large*	*with*
e_3	*cup*	*white*	*small*	*without*
e_4	*cup*	*black*	*small*	*without*
e_5	*cup*	*black*	*small*	*without*

```
1: Pref ← the preference order
2: Desc ← the description
3: r ← the intended referent
4: Dist ← the distractors
5: while Pref is not exhausted do
6:     a ← the next attribute
7:     v ← the value of r for a
8:     if v excludes some distractors then
9:         Add v to Desc
10:        Remove the distractors from Dist
11:    end if
12:    if there are no more distractors then
13:        Return Desc
14:    end if
15: end while
```

(a) The algorithm

The IA uses a preference order to model attribute salience, for example TYPE > COLOUR > SIZE in Table 1. To identify an intended referent r (say e_1), the algorithm traverses the preference order (lines 5–12), checking at each stage checks whether r's value on a given attribute excludes some distractors. Thus, for e_1, TYPE is tied first, but is not selected (all entities are dogs). Subsequently, COLOUR is found to exclude e_2 and SIZE excludes e_3.

The result is $\{black, large\}$, which is overspecified. Dale and Reiter (1995) also proposed a function to add TYPE in case it is omitted by the search. Thus, this description could be realised as *the large black dog*.

(b) An example

Fig. 1. The Incremental algorithm.

produce *the black cup with the spoon* in 80% of cases and *the small cup with the spoon* in 20% of cases. We believe this is a striking new prediction, especially if color *is* always chosen over size when it rules out more distractors. If this prediction were to be confirmed, this would provide initial support for the idea that human speakers first select the property that rules out most distractors.

Existing metrics of ''humanlikeness'' (van Deemter et al., in press; Gupta & Stent, 2005; Jordan & Walker, 2005; Viethen & Dale, 2007) were not designed to measure the extent to which an algorithm reflects the variation in a corpus. This is most easily seen in connection with variation between speakers. Consider a simple example involving just one reference

task, to which just two speakers are exposed: Speaker *a* utters NP_1 and speaker *b* utters NP_2. Suppose these two NPs are as different as they can be, so it is impossible to match (i.e., resemble) both of them. Now consider two algorithms, each of which is run twice. One algorithm generates the two human-produced NPs, while the other behaves deterministically, generating NP_1 on both runs:

Speaker *a*: NP_1. Speaker *b*: NP_2.

Algorithm 1: NP_1 (first run); NP_2 (second run).

Algorithm 2: NP_1 (first run); NP_1 (second run).

Intuitively speaking, Algorithm 1 captures the variation among the two speakers much better than Algorithm 2. Existing evaluation metrics, however, attribute the same score to both algorithms, because these metrics compute the extent to which the descriptions generated by a given algorithm match the speaker-generated descriptions *on average*, comparing each generated description with each human-produced description. Since both of the descriptions, NP_1 and NP_2, match one human-generated description fully (leading to a score of 1) while failing to match the other one entirely (scoring 0), both algorithms end up with the same averaged score of 0.5. The development and deployment of metrics that are able to measure the variation among speakers (or, indeed, within a speaker) is an example of the way in which computational research will need to change if a more psycholinguistic angle on language production is adopted.

Although it is the ultimate aim of REG to produce referring noun phrases, most current REG algorithms are limited to content determination: They determine which properties are expressed, but they seldom determine how and in which order they are realized, leaving the decision between, for example, *the big red car* and *the red big car* to a generic and independent *realisation* algorithm (*pace* Khan et al., 2012; Siddharthan & Copestake, 2004). But if algorithms are to offer full models of reference, they will need to address linguistic realization in its full generality. Psycholinguists have long suggested that incrementality plays an important role here as well. Pechmann (1989), for example, observed that participants often realized color before they realized size, even though this is not the preferred word order, and argued that this is because color is recognized faster than size, because color is not a relative property. Recently, this idea has received support from a study by Brown-Schmidt and Tanenhaus (2006), who showed that the timing of speakers' fixations to a distractor (e.g., a large triangle) predicted whether they produced a prenominal adjective (*the small triangle*) or a repair following the noun (*the triangle … the small one*). Furthermore, fixations to the distractor were made earlier when speakers produced a prenominal adjective than a postnominal modifying phrase (*the triangle with the small squares* in the context of a triangle with large squares). This supports the idea that information that is first fixated and accessed is encoded first in the referring expression. Future algorithms could incorporate these findings, for example, by making sure that highly salient properties that are rapidly encoded are realized earlier in referring expressions than less salient properties.

Finally, as was already pointed out by Dale and Reiter (1995), some algorithms are unlikely to be psychologically realistic, because they are computationally very costly to run when there is a large number of distractors. Under such conditions, which are quite typical of real-life situations, such algorithms would thus be very slow. Given the limited processing capacity of humans and their reliance on ''fast and frugal'' heuristics rather than exact calculations (Gigerenzer & Goldstein, 1996; Simon, 1956; Tversky & Kahneman, 1982), it seems likely that human processing mechanisms use clever shortcuts that will need to be incorporated into future algorithms.

5. Concluding remarks

Psycholinguists and computational linguists who study reference production approach their work with different questions, mentalities, and dispositions. Psycholinguists are driven by a wish to understand how human language production works, whereas computational linguists are at least partially motivated by a wish to contribute to practical applications. Consequently, psycholinguists are interested in the human production *process*, while computational linguists tend to focus on the *product*—the process is irrelevant as long as it can be implemented to run fast enough. Conversely, the algorithms constructed by computational linguists require a level of detail that seems foreign to psycholinguistic models of reference production, which focus on high-level issues such as audience design and alignment.

It appears to us, however, that the differences between the two disciplines are becoming increasingly irrelevant. For, on the one hand, computational linguists are increasingly making use of experimental-statistical research methods. On the other hand, there is a growing awareness that psycholinguistic models would become more interesting and more useful if they gained in algorithmic detail. The papers in the present issue of *Topics in Cognitive Science* illustrate both these trends.

Acknowledgments

The authors have contributed equally to this paper, and their names appear in alphabetical order. Thanks are due to Wayne Gray and the anonymous reviewers for their constructive comments. Emiel Krahmer and Albert Gatt thank The Netherlands Organisation for Scientific Research (NWO) for VICI grant ''Bridging the Gap between Computational Linguistics and Psycholinguistics: The Case of Referring Expressions'' (277-70-007). Kees van Deemter thanks the EPSRC Platform Grant ''Affecting People with Natural Language.''

References

Allen, J., & Perrault, C. R. (1980). Analyzing intention in utterances. *Artificial Intelligence*, *15*, 143–178.

Almor, A. (2000). Constraints and mechanisms in theories of anaphor processing. In M. Pickering, C. Clifton, & M. Crocker (Eds.), *Architectures and mechanisms for language processing* (pp. 341–354).Cambridge, UK: Cambridge University Press.

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

Ariel, M. (2001). Accessibility theory: An overview. In T. Sanders, J. Schilperoord & W. Spoorten (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 29–87). Amsterdam, The Netherlands: John Benjamins.

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, *31*(2), 137–162.

Arnold, J. E., & Griffin, Z. M. (2007). The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, *56*(4), 521–536.

Arts, A. (2004). *Overspecification in instructive texts*. Unpublished PhD thesis, Tilburg University.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification in written instruction. *Linguistics*, *49*(3), 555–574.

Bard, E., & Aylett, M. (2004). Referential form, word duration, and modeling the listener in spoken dialogue. In J. Trueswell & M. Tanenhaus(Eds.), *Approaches to studying worldsituated language use: Bridging the language-as-product and language-as-action traditions* (pp. 173–191). Cambridge, MA.: MIT Press.

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

Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, *10*, 137–167.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology*, *22*(6), 1482–1493.

Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, *1*(2), 274–291.

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*, 171–190.

Brown-Schmidt, S., & Tanenhaus, M. (2006). Watching the eyes talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*(4), 592–609.

Bruner, J. (1983). *Child's talk: Learning to use language*. New York: Norton.

Buschmeier, H., Bergmann, K., & Kopp, S. (2010). Modelling and evaluation of lexical and syntactic alignment with a priming-based microplanner. In E. Krahmer & M. Theune (Eds.), *Empiricial methods in natural language generation* (pp. 85–104). Berlin, Germany: Springer.

Callaway, C., & Lester, J. (2002). Pronominalization in generated discourse and dialogue. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)* (pp. 88–95). Philadelphia.

Clark, H. H., & Murphy, G. (1983). Audience design in meaning and reference. In J. F. LeNy & W. Kintsch (Eds.), *Language and comprehension* (pp. 287–299). Amsterdam, The Netherlands: North Holland.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

Cohen, P., & Levesque, H. (1991). Confirmations and joint actions. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-91)* (pp. 951–975). Sydney, Australia.

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'89)* (pp. 68–75). Vancouver, BC, Canada.

Dale, R., & Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, *18*, 233–263.

Dale, R., & Viethen, J. (2010). Attribute-centric referring expression generation. In E. Krahmer & M. Theune (Eds.), *Empirical methods in natural language generation* (pp. 163–179). Berlin: Springer Verlag.

de Ruiter, J., Bangerter, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2012.01183.x.

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

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

Fabbrizio, G. di, Stent, A., & Bangalore, S. (2008). Trainable speaker-based referring expression generation. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CONLL'08)*. Manchester, England.

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

Fukumura, K., van Gompel, R., & Pickering, M. J. (2010). The use of visual context during the production of referring expressions. *Quarterly Journal of Experimental Psychology*, *63*, 1700–1715.

Fukumura, K., & van Gompel, R. P. G. (2010). Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language*, *62*, 52–66.

Garrod, S., Freudenthal, D., & Boyle, E. (1994). The role of different types of anaphor in the on-line resolution of sentences in a discourse. *Journal of Memory and Language*, *33*, 39–68.

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* (pp. 264–293). Berlin: Springer Verlag.

Gatt, A., Gompel, R. van, Krahmer, E., & Deemter, K. van. (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 (pre-cogsci'11)*. Retrieved January 2012, from http://pre2011.uvt.nl/workshop-program.html.

Gibbs, R., & Van Orden, G. (2012). Pragmatic choice in conversation. *Topics in Cognitive Science*, *4*, 1.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Givon, T. (1983). *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam: Benjamins.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, *17*(3), 311–347.

Gordon, P. C., & Hendrick, R. (1999). The representation and processing of coreference in discourse. *Cognitive Science*, *22*, 389–424.

Gordon, P. C., Kendrick, R., Ledoux, K., & Yang, C. L. (1999). Processing of reference and the structure of language: An analysis of complex noun phrases. *Language and Cognitive Processes*, *14*, 353–379.

Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering and referential overspecification. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2012.01186.x.

Grice, H. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics: Speech acts (Vol. III)* (pp. 41–58). New York: Academic Press.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203–225.

Guhe, M. (2012). Utility-based generation of referring expressions. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2012.01185.x.

Gundel, J., Hedberg, N., & Zacharski, R. (2012). Underspecification of cognitive status in reference production: Some empirical predictions. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2012.01184.x.

Gupta, S., & Stent, A. (2005). Automatic evaluation of referring expression generation using corpora. In *Proceedings of the 1st Workshop on Using Copora in Natural Language Generation (UCNLG'05)* (pp. 1–6). Brighton, England.

Heeman, P. A., & Hirst, G. (1995). Collaborating on referring expressions. *Computational Linguistics*, *21*(3), 351–382.

Heller, D., Skovbroten, K., & Tanenhaus, M. (2012). To name or to describe: shared knowledge affects referential form. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2012.01182.x.

Holden, J., & Van Orden, G. (2009). Dispersion of response times reveals cognitive dynamics. *Psychological Review*, *2*, 318–342.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59,* 91–117.

Jordan, P. (2002). Contextual influences on attribute selection for repeated descriptions. In K. van Deemter & R. Kibble (Eds.), *Information sharing: Reference and presupposition in language generation and interpretation* (pp. 295–328). Stanford, CA: CSLI Publications.

Jordan, P., & Walker, M. (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, *24*, 157–194.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*, 25–41.

Khan, I., van Deemter, K., & Ritchie, G. (2012). Managing ambiguity in reference generation: The role of surface structure. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2011.01167.x.

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

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of descriptions in context. In K. van Deemter & R. Kibble (Eds.), *Information sharing: Givenness and newness in language processing* (pp. 223–264). Stanford, CA: CSLI Publications.

Krahmer, E., & van Deemter, K. (2011). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218.

Kroch, A. (2000). Syntactic change. In M. Baltin & C. Collins (Eds.), *Handbook of contemporary syntactic theory* (pp. 699–729). Oxford, England: Blackwell.

Kronfeld, A. (1990). *Reference and computation: An essay in applied philosophy of language*. Cambridge, UK: Cambridge University Press.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.

Matthews, D., Butcher, J., Lieven, E., & Tomasello, M. (2012). Two- and four-year-olds learn to adapt referring expressions to context: Effects of distracters and feedback on referential communication. *Topics in Cognitive Science*, DOI: 10.1111/j.1756-8765.2012.01181.x.

Matthews, D. E., Lieven, E., & Tomasello, M. (2007). How toddlers and preschoolers learn to uniquely identify referents for others: A training study. *Child Development*, *78*, 1744–1759.

McCoy, K., & Strube, M. (1999). Generating anaphoric expressions: Pronoun or definite description? In *Proceedings of the ACL'99 Workshop on Discourse and Reference Structure* (pp. 63–71). College Park: University of Maryland.

Mellish, C., Scott, D., Cahill, L., Paiva, D., Evans, R., & Reape, M. (2006). A reference architecture for natural language generation systems. *Natural Language Engineering*, *12*, 1–34.

Oberlander, J. (1998). Do the right thing . . . but expect the unexpected. *Computational Linguistics*, *24*(3), 501–507.

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

Paraboni, I., van Deemter, K., & Masthoff, J. (2007). Generating referring expressions: Making referents easy to identity. *Computational Linguistics*, *33*, 229–254.

Passonneau, R. (1996). Using centering to relax Gricean informational constraints on discourse anaphoric noun phrases. *Language and Speech*, *39*, 229–264.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*, 98–110.

Pickering, M., & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*, *27*, 169–226.

Poesio, M., Stevenson, R., Eugenio, B. D., & Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational Linguistics*, *30*, 309–363.

Purver, M., & Kempson, R. (2004). Context-based incremental generation for dialogue. In *Proceedings of the 3rd International Conference on Natural Language Generation (INLG'04)* (pp. 151–160). Brockenhurst, England.

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

Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge, UK: Cambridge University Press.

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

Siddharthan, A., & Copestake, A. (2004). Generating referring expressions in open domains. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL-2004)* (pp. 407–414). Berlin: Springer-Verlag.

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138.

Stevenson, R. J., Crawley, R. A., & Kleinman, D. (1994). Thematic roles, focus and the representation of actions. *Language and Cognitive Processes*, *9*, 519–548.

Stoia, L., Shockley, D. M., Byron, D. K., & Fosler-Lussier, E. (2006). Noun phrase generation for situated dialogs. In *Proceedings of the 4th International Conference on Natural Language Generation (INLG'06)* (pp. 81–88). Morristown, NJ: 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* (pp. 3–22). Cambridge, UK: Cambridge University Press.

van Deemter, K., Gatt, A., van der Sluis, I., & Power, R. (in press). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*.

van der Sluis, I., & Krahmer, E. (2007). Generating multimodal referring expressions. *Discourse Processes*, *44*(3), 145–174.

Viethen, J., & Dale, R. (2007). Evaluation in natural language generation: Lessons from referring expression generation. *Traitement Automatique des Langues*, *48*(1), 141–160.

Wardlow Lane, L., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants! Speakers' control over leaking private information during language production. *Psychological Science*, *17*, 273–277.

Winograd, T. (1972). *Understanding natural language*. New York: Academic Press.

Wu, S., & Keysar, B. (2007). The effect of information overlap on communication effectiveness. *Cognitive Science*, *31*, 1–13.