Language, Cognition and Neuroscience, 2014
Vol. 29, No. 8, 899–911, http://dx.doi.org/10.1080/23273798.2014.933242

Routledge
Taylor & Francis Group

Models and empirical data for the production of referring expressions

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

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

(*Received 21 October 2013; accepted 29 May 2014*)

This paper introduces a special issue of *Language, Cognition and Neuroscience* dedicated to *Production of Referring Expressions: Models and Empirical Data*, focusing on models of reference production that make empirically testable predictions, as well as on empirical work that can inform the design of such models. In addition to introducing the volume, this paper also gives an overview of recent experimental and modelling work, focusing on two principal aspects of reference production, namely, choice of anaphoric referential expression and choice of semantic content for referential noun phrases. It also addresses the distinction between dialogue and non-dialogue settings, focussing especially on the impact of a dialogue setting on referential choice and the evidence for audience design in the choices speakers make.

Keywords: language production; computational modelling; referring expressions

Reference is one of the most intensively studied aspects of human language. Since the ability to refer and identify entities in spoken and in written discourse is so central to communication, it is unsurprising that, following the foundational work of philosophers such as Frege (1892) and Russell (1905), the study of reference has been a central concern for theoretical linguists and philosophers of language (e.g. Abbott, 2010), psycholinguists (e.g. Belke & Meyer, 2002; Brennan & Clark, 1996; Olson, 1970; Pechmann, 1989), developmental psychologists (e.g. Deutsch & Pechmann, 1982; Ford & Olson, 1975; Matthews, Butcher, Lieven, & Tomasello, 2012), and computational linguists (e.g. Dale, 1989; Dale & Reiter, 1995; Krahmer & van Deemter, 2012).

This special issue of *Language, Cognition and Neuroscience* focusses on explicit models that account for some aspect of reference production and make precise empirical predictions about it, as well as new experimental data that can inform the development of such models.

The dual emphasis on empirical work and predictive models stems from our observation that, despite a wealth of experimental research on reference in psycholinguistics, as well as computational work on algorithms that generate referring expressions, the two fields frequently proceed in parallel with little cross-fertilisation (van Deemter, Gatt, van Gompel, & Krahmer, 2012). Relatively few experimental results are harnessed in the development of computational models for reference production. Furthermore, the role of such models is somewhat different in the two fields. Yet the concerns of researchers in these fields also evince a striking degree of convergence. Indeed, models of reference production have to address the broad range of choices faced by a speaker in a given context who wishes to refer to a specific entity or set thereof. Several types of choices have been identified and discussed at length in the literature. The papers in this volume concentrate on two of these in particular:

(1) Choice of (anaphoric) referring expression in discourse (e.g. full definite noun phrase, reduced Noun Phrase (NP), pronoun; e.g. Ariel, 2001; Gundel, Hedberg, & Zacharski, 1993; Grosz, Joshi, & Weinstein, 1995).
(2) *Conceptualisation, or choice of properties*, in a referring expression that takes the form of a full definite noun phrase (e.g. Arts, 2004; Belke, 2006; Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Olson, 1970; Pechmann, 1989).

The goal of this introduction is to sketch the broader context within which the papers presented in this special issue are situated, with reference to these two questions. Clearly, the two are interrelated. Consider a language producer who needs to identify the object surrounded by a red border for an interlocutor in either of the two visual domains in Figure 1, consisting of aeroplanes of different sizes and colours.

What kind of referring expression is produced will depend on a variety of constraints that affect the salience of the intended referent. If the referring expression is produced in the context of a text or a dialogue, in the course of which the referent has been mentioned already,

*Corresponding author. Email: albert.gatt@um.edu.mt

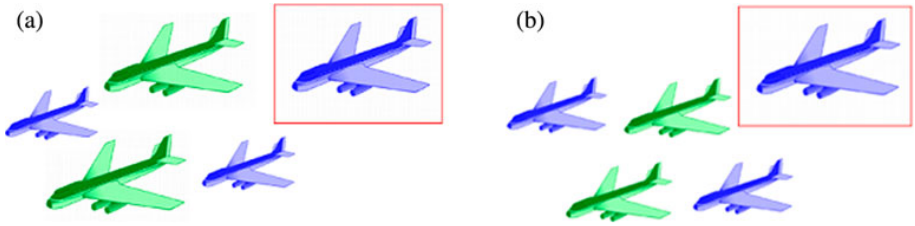

Figure 1. Two example domains.

then the language producer may opt for a reduced expression, in the form of a pronoun (*it*) or a deictic (*that one*). Such anaphoric references make an important contribution to coherent discourse. On the other hand, if the expression selected is a full or a reduced NP – for example, because this is a language game in which 'one-shot' references are being produced for entities in a visually copresent scene or because the entity was introduced much earlier in the discourse and so is not very salient anymore – then the conceptualisation problem arises. Does an expression like *the large aeroplane* suffice to convey the intended referent to a receiver? It clearly does not in Figure 1(a) (there are other large aeroplanes), but it might in Figure 1(b). Still, a language producer may, nevertheless, opt to call the target in Figure 1(b) *the large blue aeroplane*. Which constraints affect a producer's choice in such cases?

For both choices, we are interested not only in reviewing the experimental evidence but also the extent to which computational models have played a role in motivating further empirical work while explaining or characterising the processes involved. Therefore, we will begin with an overview of the role of models in these two disciplines before turning to a review of experimental and computational literature. In so doing, we hope to show that there is substantial convergence among researchers working on reference production, despite some differences of orientation.

Reference production models: convergences and divergences

A glance at the computational linguistics literature and the psycholinguistic literature on modelling[1] reveals a number of differences. Two of these are especially important for the present discussion: (1) differences in the *why* of computational modelling, that is, in what a model is expected to achieve; (2) differences in the way in which models are evaluated. Despite these differences, we argue, there is considerable scope for convergence between the two fields.

Why build models?

Algorithms designed by computational linguists often aim for practical benefits, such as efficiency and coverage. Thus, models of reference production, especially in the field known as Referring Expression Generation (REG), are intended to be incorporated into larger Natural Language Generation (NLG) systems, whose task is to produce text or speech from non-linguistic input (see Reiter & Dale, 2000, for an overview).[2] REG algorithms aim to produce output which is understandable by human readers or listeners but are not typically committed to the cognitive validity of the processes they incorporate, although some of the most influential were partly motivated by psycholinguistic results (a well-known example is Dale & Reiter, 1995, which we shall revisit below).

By contrast, the aim of cognitive models is to characterise and/or explain a human phenomenon based on empirical evidence (see Lewandowsky & Farrell, 2011, for discussion). This is consistent with their historical role as tools for cognitive investigation: as recently noted by McClelland (2009), it was the development of computational methods in the first half of the twentieth century that gave the cognitive sciences the wherewithal to explore processes at a level of detail that behavioural experiments alone do not permit (cf. Newell, 1973, for an early argument to this effect). This development was, of course, facilitated by the fact that mental processes were increasingly being conceptualised in computational (e.g. symbol-processing) terms (see Boden, 2008, for a history of the development of these ideas).

To what extent can these two perspectives – the practical and the explanatory – be brought together? An example from each class of referential choice outlined in the Introduction may serve to illustrate that they can (we return to both cases in the following sections).

In the computational literature, a well-known model of anaphoric reference that has also become influential in the psycholinguistic literature is Centring Theory (Grosz et al., 1995). Centring aims to account for anaphoric reference between consecutive utterances, each of which has a set of forward-looking centres (discourse entities) that are ranked according to their salience, depending on grammatical role (Brennan, Friedman, & Pollard, 1995; Grosz et al., 1995). Every utterance also has a single backward-looking centre, which is the discourse entity in the current utterance that was the highest-ranked forward-looking

centre in the previous utterance. Centring does not consider the whole plethora of forms available to the producer but aims to formalise when pronouns are preferred. Several slightly different versions of a constraint on pronominalisation have been formulated. Grosz, Joshi, and Weinstein (1983) proposed that a pronoun should be used if the backward-looking centre in the current utterance is the same as that in the previous utterance, while Grosz et al. (1995) proposed that if any forward-looking centre is pronominalised, then the backward-looking centre must be; finally, Gordon, Grosz, and Gilliom (1993) argued that the backward-looking centre should always be pronominalised. Although there are obviously differences in the predictions of these different versions, they all assume that reference to John in the final sentence in (1a) should be with a pronoun, whereas a pronoun is not required for reference to Kate in (1b). The reason for this is that John is the backward-looking centre in both (1a) and (1b), whereas Kate is not.

(1) (a) John noticed Kate. He hated her. He/Peter was often offended by her.
(b) John noticed Kate. He hated her. She/Kate often offended him.

In view of the variations among different Centring proposals, Poesio, Stevenson, Di Eugenio, and Hitzeman (2004) harnessed corpus-based computational methods to test the predictions of the theory against naturally occurring data. Most importantly, for the purposes of the present discussion, this study highlighted a significant degree of underspecification in the Centring model. The computational approach enabled a systematic comparison of the various instantiations of Centring and their predictions. For example, Gordon and colleagues (Gordon et al., 1993; Gordon & Chan, 1995) have found evidence that reading times for the backward-looking centre (e.g. John in 1a) were faster when it was realised as a pronoun than a repeated name. However, Poesio et al. (2004) found that backward-looking centres were often (in 45% of cases) not realised as pronouns, suggesting that Gordon et al.'s (1993) version of the constraint on pronominalisation was too strong. Grosz et al.'s (1995) version of the constraint (if any forward-looking centre is pronominalised, then the backward-looking centre must be) appeared to account best for the data: this rule was violated in only 3% of cases, while Grosz et al.'s (1983) version also fared fairly well (19% violations). The study by Poesio et al. (2004) also highlights a methodological point, namely, that it was only through explicit modelling and application to naturalistic data that competing versions of a model could be compared.

Turning to our second theme, conceptualisation or property selection, Dale and Reiter (1995) developed the Incremental Algorithm (IA), which selects the content for a descriptive referential noun phrase. The algorithm was a response to two developments. The first was the view, based on an interpretation of Grice (1975) and promoted in early psycholinguistic work (e.g. Ford & Olson, 1975; Olson, 1970), that producers aim to maximise efficiency in reference production through descriptions which contain no more information than required. However, REG models based on this 'full brevity' strategy (e.g. Dale, 1989) were found to be intractable because the identification of the smallest set of properties that jointly identify a referent requires a search through a space of possibilities that grows exponentially (see Reiter, 1990, for discussion). The second development was the empirical finding, by Pechmann (1989) among others, that humans frequently produce 'overspecified' descriptions as a result of the incremental nature of language production (Levelt, 1989). Dale and Reiter's algorithm explicitly incorporated a model of incrementality, resolving the intractability of the 'full brevity' approach, while also approximating the overspecification observed among speakers. Apart from its practical utility (because of its efficiency), such a model also raises questions that are of psycholinguistic interest. For example, can the model accurately predict the choices human producers make, especially in view of its deterministic behaviour (Gatt, van Gompel, Krahmer, & van Deemter, 2011)? Are more recent models that emphasise efficiency in communication (e.g. Frank & Goodman, 2012) better predictors of speaker choices?

As we have noted above, the choices modelled by Centring and by the IA are not independent of each other, in so far as properties need to be selected once a referential form has been determined. Krahmer and Theune (2002) have proposed an extension of Dale and Reiter's (1995) IA to handle anaphora, using ideas from Centring Theory. They propose that in selecting properties, a referent need only be distinguished from distractors that are more salient, where salience is computed based on grammatical role as in Centring Theory, decreasing with every intervening utterance. Pronouns are generated in case the reference is to the most salient entity in the discourse.

How should models be evaluated?

Cognitive models stand or fall by their ability to accurately predict experimental results as measured by metrics that evaluate goodness of fit between predicted and attested outcomes (Lewandowsky & Farrell, 2011).

On the other hand, the practice in REG and other subfields of computational linguistics varies depending on the aims of the system under development. To the extent that a model of reference production aims to produce 'comprehensible' referring expressions, it should be evaluated through a task that is 'addressee-oriented', for example, by estimating the time it takes for listeners to

identify referring expressions based on descriptions produced by an algorithm (e.g. Gatt & Belz, 2010). On the other hand, if the model is intended to mimic producer behaviour, the method of choice is comparison against production data, for example, a corpus (e.g. van Deemter, Gatt, van der Sluis, & Power, 2012; Viethen & Dale, 2007). Recent work in the context of a series of NLG shared tasks, in which participants are required to design algorithms that are developed and tested against a common data-set to enable comparison, has shown that results from these two perspectives may diverge significantly (Belz, Kow, Viethen, & Gatt, 2010; Gatt & Belz, 2010). For instance, an algorithm's choice of content for referential descriptions may be very similar to the choices humans make, as shown by its degree of match to corpus data, but this does not imply that the resulting description will be easily resolved by human listeners. This echoes the finding in some psycholinguistic studies that there may be a mismatch between the strategies used by speakers to produce references and those that listeners find easiest to process (see Engelhardt, Bailey, & Ferreira, 2006; Engelhardt, Demiral, & Ferreira, 2011, for example). This is a theme to which we will return in the section on Conceptualisation below.

An important issue in evaluation concerns the extent to which a model generalises beyond a specific data-set. A cognitive model is often motivated by data from a specific experimental task, although a goodness-of-fit metric will take into account the extent to which its predictive power relies on characteristics of the data distribution. By contrast, REG models have frequently been couched in general-purpose terms. For instance, the IA described above (Dale & Reiter, 1995) was not specifically tied to a particular scenario or domain. A recent example of the tension that may arise between general-purpose and domain-specific approaches comes from the work of Guhe (2012), which addresses conceptualisation in the context of the iMAP dialogues (Louwerse et al., 2007), where interlocutors negotiate a route through a maze while seeking to align their mutual knowledge of the spatial layout. Guhe describes two property selection models couched in the ACT-R framework (Anderson et al., 2004), one based on the 'general purpose' IA and the other a simpler model based explicitly on the iMAP domain. The latter model outperforms the IA (in the sense that its predictions account for a greater proportion of the variance in the data) but does so at the expense of generalisability beyond iMAP-like domains. A similar concern with generalisability is also evident in work using data-driven, machine-learning methods, where a model which is trained on annotated corpus data is usually evaluated on held-out, previously unseen data using measures such as precision and recall (e.g. Di Fabbrizio, Stent, & Bangalore, 2008; Viethen, Dale, & Guhe, 2011).

Despite the diverging goals of models developed in the computational and psycholinguistic communities, there are compelling reasons to inculcate a greater mutual awareness within these two fields. To the extent that REG algorithms are intended to mimic a human capacity (over and above their practical considerations), they would benefit from psycholinguistic insights, not only on the nature of the underlying processes but also on the extent to which such processes vary across conditions and individuals. On the other hand, psycholinguistic results on reference production can also benefit from computational approaches, especially since this often sheds light on underspecified aspects of a theoretical model.

Against the background of this general discussion, let us now turn to some of the main questions that have engaged the attention of researchers on reference production in recent years.

Choice of anaphoric expression

In functional linguistics, there has been a long-standing interest in the question of why language producers choose particular anaphoric forms such as pronouns, demonstratives, repeated and non-repeated definite noun phrases, depending on the context in which they occur. One general assumption is that there is a direct relationship between the form of a referring expression and the accessibility (or giveness or topicality) of the referent in the addressee's discourse model (e.g. Ariel, 1990, 2001; Givón, 1983; Gundel et al., 1993; Gundel, Hedberg, & Zacharski, 2012). When the referent is highly accessible, the use of a short anaphoric form with limited semantic information suffices, but for inaccessible referents, a form with more semantic content is needed, giving rise to the second type of referential choice outlined in the Introduction, namely, which content to select, which we turn to in the following section.

The most elaborate hierarchy has been proposed by Ariel (1990) and is partially shown in (2) below, where accessibility decreases as one moves rightwards.

(2) zero pronoun >unstressed pronoun >stressed pronoun >proximate demonstrative >distal demonstrative >first name >last name >definite description >full name

Accessibility is assumed to be influenced by various factors including recency and frequency of mention and the number of semantically similar competitor discourse entities (Ariel, 1990; Givón, 1983); this is also compatible with the view in Centring Theory that pronouns tend to be used to refer to highly accessible referents, as we have seen.

Several studies have shown that language producers tend to use more pronouns and fewer repeated names or noun phrases when referring to a subject antecedent in the preceding clause than when referring to an object

antecedent (Arnold, 2001; Brennan, 1995; Fletcher, 1984; Stevenson, Crawley, & Kleinman, 1994) consistent with linguistic theories that assume that the subject generally refers to more accessible entities (e.g. Chafe, 1976; Keenan & Comrie, 1977; Li & Thompson, 1976). Recent data by Fukumura and van Gompel (2014) suggest that it is, indeed, the grammatical role of the antecedent and not its position in the sentence that affects the choice of anaphor. Furthermore, Rohde and Kehler (2013, this volume) showed that language producers use more pronouns when they refer to the subject of a passive than active sentence, suggesting that the non-canonical passive structure marks subjects as more topical. All these studies compared subject and object referents that were in the same clause. When referents in different clauses are compared, language producers' use of pronouns is influenced by recency (Ariel, 1990; Arnold, Bennetto, & Diehl, 2009; Givón, 1983). In contrast, the likelihood with which people refer to a referent does not appear to play a role in people's choice of anaphor. In a sentence completion study, Fukumura and van Gompel (2010) found that participants referred much more often to the entity with the stimulus (Gary) than the experiencer role (Anna) in sentences such as (3).

(3) (a) Gary scared Anna after the long discussion ended in a row. This was because …
(b) Anna feared Gary after the long discussion ended in a row. This was because …

However, which anaphoric form they chose (pronoun or repeated name) was only affected by the grammatical role (subject or object) of the antecedent. Rohde and Kehler (2013, this volume) show the same pattern of results with gender-ambiguous pronouns. These findings provide evidence against the view that referents that are more frequently referred to are more accessible (Arnold, 2001, 2008) as well as models that assume that speakers produce more reduced expressions when reference is more predictable (e.g. Jaeger, 2010). Instead, they fit better with accounts that assume that what speakers refer to and how they do this are independent (e.g. Kehler, Kertz, Rohde, & Elman, 2008; Rohde & Kehler, 2013; Stevenson et al., 1994). For example, Rohde and Kehler (2013, this volume) argue for a Bayesian model according to which the choice of a specific referent given a particular anaphoric form and the choice of a specific form given a particular referent are affected by different factors.

Choice of referential form is also affected by non-linguistic factors, especially visual salience. Fukumura, van Gompel, and Pickering (2010) found that participants produced more pronouns and fewer repeated noun phrases to refer to a person when another person was present in the visual context than when there was not. However, Vogels, Krahmer, and Maes (2013b) found no evidence that the visual salience of the referent (whether it was foregrounded or backgrounded) had an effect on speakers' choice of anaphor. If choice of NP is affected by the presence of competitors, but not the salience of the referent itself, then the visual salience effect might be due to competition, whereby competitors reduce referent accessibility, resulting in more explicit expressions.

Competition has also been linked to other factors. Arnold and Griffin (2007) found that speakers produced fewer pronouns and more repeated names when the preceding sentence mentioned a competitor with the same gender than with a different gender, suggesting that two characters with the same gender compete. Fukumura and van Gompel (2010) observed that language producers used fewer pronouns when the referent and competitor were both animate or inanimate than when they were different in animacy, while Fukumura, van Gompel, Harley, and Pickering (2011) showed that speakers produced fewer pronouns when the referent and the competitor were visually similar (e.g. both were sitting on a horse) than when they were not. One possibility is that all these effects are due to ambiguity avoidance rather than due to competition, for example, when the referent and the competitor have the same gender, speakers may avoid pronouns because they would be gender-ambiguous. However, Fukumura, Hyönä, and Scholfield (2013) found that even in Finnish, where pronouns are not marked for gender, speakers produced fewer pronouns when the referent and the competitor had the same gender than when their gender was different. Because Finnish pronouns are ambiguous regardless of whether two characters in the preceding context have the same gender, these results suggest that when the referent and the competitor are semantically similar, they compete for activation, lowering the accessibility of the referent and resulting in more explicit anaphoric expressions.

Language producers also use more pronouns with animate than inanimate antecedents (Fukumura & van Gompel, 2011), in line with the idea that animate referents are more salient than inanimate referents (e.g. Comrie, 1989; Bock & Warren, 1985). Normally, lexical information indicates animacy, but results by Vogels, Krahmer, and Maes (2013a) suggest that lexical animacy can be overridden by visual context. In their study, abstract figures moved in a way that suggested that they were either animate (e.g. jumping up and down irregularly, suggesting voluntary control) or inanimate (e.g. rolling down a slope, suggesting externally controlled movement). Participants used more pronouns and fewer repeated noun phrases when they referred to objects with animate-like than inanimate-like movements, whereas lexical animacy (prior reference with either an animate or inanimate word) did not have an effect in such cases. Interestingly, in another study, Vogels et al. (2013a) showed that language producers do not always produce

more reduced expressions for animate than inanimate referents. Dutch language producers used fewer reduced pronouns (e.g. *ze*) for animate than inanimate antecedents, possibly because grammatical gender marking in Dutch is gradually disappearing, and Dutch producers may be avoiding highlighting the gender of inanimates (Vogels et al., 2013a). It is also possible that they use full forms to indicate that the referent is likely to be important in the upcoming discourse; animate referents tend to play a more central role in the discourse than inanimate entities. Importantly, these results are inconsistent with anaphor hierarchies that assume that speakers choose more reduced expressions for more accessible referents (Ariel, 1990; Givón, 1983; Gundel et al., 1993). It suggests that anaphoric reduction is not just affected by the accessibility of the referent, but that other factors also play a role.

There is further evidence that the assumptions underlying anaphor hierarchies may be too simplistic. Ariel's (1990) and Givón's (1983) hierarchies consist of single scales, where all forms are affected by the same accessibility factors. However, several language comprehension experiments have shed doubt on this (e.g. Brown-Schmidt, Byron, & Tanenhaus, 2005; Kaiser, 2011; Kaiser & Trueswell, 2008). For example, in the experiments of Kaiser and Trueswell (2004), participants tended to interpret a personal pronoun (*hän*) as referring to the subject of a preceding sentence, whether it was in Subject-Verb-Object (SVO) or Object-Verb-Subject (OVS) order, whereas a demonstrative (*tämä*) was interpreted as referring to the second-mentioned entity. The same pattern of results was observed in a sentence completion experiment. This suggests that the two pronouns are influenced by different accessibility factors.

Functional-linguistic theories generally also assume that speakers use particular anaphoric forms to signal the degree of accessibility of the referent to the addressee (e.g. Ariel, 2001; Givón, 1983), with more information signalling less accessibility. Thus, these accounts assume that speakers are allocentric and try to facilitate comprehension. However, an alternative account is that speakers are egocentric, making their choices based solely on their own discourse model, though this often results in anaphors that are comprehensible to the addressee because interlocutors' discourse context is usually shared.

There is little doubt that speakers are sensitive to the addressee's needs to some extent: if they chose anaphoric expressions that were easiest for themselves to produce, then they would presumably always produce pronouns because they are shorter and more frequent than more explicit expressions. However, the question is whether they take into account more fine-grained information about the addressee, such as whether the referent is accessible to the addressee. Fukumura and van Gompel (2012) presented speakers with context sentences in which both the referent and the competitor were mentioned. Linguistic salience was manipulated by either mentioning the referent or the competitor in the second sentence, which was either audible or inaudible to the addressee. Following this second sentence, speakers produced more pronouns and fewer repeated noun phrases when the referent was made salient by the second sentence than when it was not, but crucially, their choice of anaphor was unaffected by whether the addressee could hear the second sentence. Thus, speakers did not take into account how accessible the referent was to their addressee but based their anaphoric choice entirely on their own discourse model.

In a related vein, Bard, Hill, Foster, and Arai (2014, this volume) report an experiment where participants described routes on maps to each other. They manipulated whether or not participants could see each other's computer mouse or eye gaze projected onto their own screen. Furthermore, in some conditions, one participant was asked to be the director and the other the follower, whereas in other conditions, the participants were not assigned different roles. Several findings in this study suggested that speakers took into account course-grained information, for example, when the mouse was visible to the participants, speakers used fewer definite noun phrases and more deictic expressions (e.g. this and that), which are most felicitous in the presence of a pointing gesture. They also used fewer definites and more deictics when they were assigned the same roles in the task than when they had different roles, presumably because the dialogue was easier when they had the same roles. Furthermore, speakers used more pronouns when addressees hovered their mouse over the referent than when they did not, and they produced more definites when addressees looked at the referent than when they did not. But crucially, these effects occurred regardless of whether the speaker could see the addressee's mouse or eye-gaze projection, indicating that they were not sensitive to this fine-grained information about the addressee.

In sum, research on the choice of anaphoric expression has shown that language producers generally use more reduced expressions when the referent is accessible than when it is inaccessible and that various linguistic and non-linguistic factors affect the choice of anaphor. However, different types of anaphor may be sensitive to different accessibility factors. Furthermore, current evidence suggests that speakers' choice of anaphor is based on course-grained information about the addressee, but not fine-grained information. There are relatively few computational models that account for this large body of empirical work; notable exceptions include Rohde and Kehler (2013, this volume) and implementations and evaluations of Centring Theory (Krahmer & Theune, 2002; Passonneau, 1997; Poesio et al., 2004). There is a need for more computational work that addresses the increasingly nuanced picture that emerges from the empirical results we have reviewed here. To date, the lion's share of

computational models of reference production has focussed on the second type of choice we outlined in the Introduction. It is to these models that we now turn.

Conceptualisation: determining the content of referential NPs

In addressing the factors influencing language producers' choice of content for referential NPS, researchers in both the computational and the psycholinguistic community have tended to rely on visual domains such as those shown in Figure 1. Our example in the Introduction hinted at a central issue in these studies, namely, the extent to which speakers' choices are motivated by communicative efficiency, typically interpreted along the lines suggested by Grice's Maxim of Quantity and emphasised in early theoretical work (Olson, 1970) as well as in empirical investigations among children (Deutsch & Pechmann, 1982; Ford & Olson, 1975; Matthews et al., 2012; Sonnenschein, 1982, 1984) and adults (Pechmann, 1989, among many others). Under this view, if the producer's purpose is to identify an entity, the combination of properties selected should be such as to maximise her interlocutor's chances of finding the entity as quickly as possible. Thus, in Figure 1(a), both the aeroplane's size and its colour would be required to distinguish it from the other distractors (resulting in a description such as *the large blue aeroplane*); by contrast, in Figure 1(b), its size would suffice (*the large aeroplane*). As we have also seen, REG algorithms have also been couched in these terms (e.g. Dale, 1989; Gardent, 2002).

It turns out, however, that adult speakers tend not to exhibit the behaviour predicted by a simple reading of the Gricean maxim.[3] Pechmann (1989) found that speakers tend to *overspecify*. Thus, our hypothetical language producer would be more likely to refer to the target referent in Figure 1(b) as *the large blue aeroplane*, even though the referent's size alone would suffice.

Pechmann suggested that this is due to the incremental nature of speech production, whereby speakers select properties in sequence and formulate their utterance accordingly, prioritising those that are most salient or highly 'preferred', before having completely scanned the domain. The tendency to overspecify by including certain properties (such as colour) has proven to be remarkably robust and has been replicated several times (e.g. Arts, 2004; Engelhardt et al., 2006; Koolen, Gatt, Goudbeek, & Krahmer, 2011). Some studies have also suggested that the basis for the preference for properties like colour might lie in 'early' perceptual, rather than conceptual or linguistic factors (e.g. Belke, 2006; Belke & Meyer, 2002), although it has been shown to be affected by linguistic factors as well, for example, whether basic-level terms can be used for a referent's colour (Viethen, Goudbeek, & Krahmer, 2012).

Another factor that modulates overspecification is the nature of the communicative task and the domain. For example, Paraboni, van Deemter, and Masthoff (2007) found that overspecification increases in spatially complex domains, while Arts (2004) found more overspecification when references were being produced in an instruction-giving task. Jordan (2002) and Jordan and Walker (2005) also demonstrate a relationship between content determination and communicative intention in a dialogue setting. For example, in the context of a dialogue in which interlocutors are negotiating to buy furniture for an apartment, a speaker may use ostensibly redundant properties in order to confirm that she has understood and accepted her partner's previous utterance.

Since the Incremental Algorithm (IA; Dale & Reiter, 1995) was developed in part to address the empirical observations of Pechmann and others, and it has proven to be highly influential, it is worth looking more closely at some of the implications of this and related models.

The IA is in fact one of a family of algorithms which model content selection as an incremental process, iteratively selecting properties that help to distinguish a referent from its distractors until a fully distinguishing description is found. The basic assumption underlying these models is that if a speaker mentions, for example, that an object is blue, she implies that at least some of the distractors are not blue; these are ruled out by mentioning this property. The models based on this incremental procedure differ primarily in how a property is selected at any stage (see Krahmer, van Erk, & Verleg, 2003, for a general framework based on graphs that can accommodate these different heuristics by interpreting them as cost functions).

The IA itself makes a selection based on a predetermined (and possibly domain-specific) preference order. For instance, faced with the domain in Figure 1(b), and given a preference order which places colour before size, the IA would first select the referent's colour (because it excludes the non-blue distractors), but then would still need to select size because colour alone does not distinguish the referent. The outcome is an overspecified description.

An alternative incremental procedure is incorporated in the Greedy Heuristic (Dale, 1989), which tries out properties in order of their discriminatory power: properties that rule out many distractors are preferred over properties that rule out only a few. This algorithm would immediately select size in Figure 1(b) and terminate on finding that the referent has been fully distinguished.

One feature of the family of incremental models we have just outlined is their *serial dependency* (Viethen et al., 2011), whereby a property, once included in a description, alters the set of distractors[4] and therefore also the likelihood with which remaining properties will be selected. In a study using machine-learning techniques to

compare different REG strategies using corpora of human-produced referring expressions in dialogue, Viethen et al. found that serial dependency *lowers* the goodness of fit to human data. This suggests that there are at least two possible interpretations of Pechmann's views on preference and incrementality: the serial dependency model as captured in the IA and an alternative model where property selection is performed incrementally, but the selection of any property in the available set is independent of all the others. More research is required to distinguish between these two possibilities.

Another prediction that is implicitly made by this family of models is that overspecification only occurs if producers happen to choose a property that rules out at least one distractor at that stage in the content selection process. In other words, even in an incremental framework, property choices are motivated by the overarching intention to identify the referent (i.e. exclude all other possible entities). By contrast, Koolen, Goudbeek, and Krahmer (2013) found that speakers, when confronted with a blue ball and a blue square (and no objects with a different colour), would still on occasion produce a description like *the blue ball* (though less frequently than when the square was not blue), suggesting that preference for colour is not completely overruled by discriminatory power even in such simple situations.

None of the models we have surveyed were originally developed or tested in a dialogue setting, although they were intended as 'general-purpose'. Yet, as the work of Jordan and Walker (2005), Viethen et al. (2011) and Guhe (2012) suggests, these models often make the wrong predictions when a language producer is engaged in a purposeful exchange with an interlocutor. In particular, they ignore the possibility that language producers' choices of content may be influenced by discourse history and by the addressee's state of knowledge as evidenced in the course of the dialogue. In this connection, Viethen, Dale, and Guhe (2014, this volume) compare classic REG algorithms and models which explicitly take into account alignment with an interlocutor. They find evidence that REG models which select content purely on the basis of discriminatory value do not capture the variation in human dialogue data, compared to implementations of psycholinguistic models that emphasise alignment at various linguistic levels.

Current debates on communication in dialogue cluster around two principal theoretical poles. On the one hand, language producers may be viewed as explicitly making their choices based on a common ground that is shared with an interlocutor. Under this view, they are ultimately engaged in conscious collaborative strategies to achieve alignment (Clark, 1996). Evidence of such strategies comes from explicit negotiation on how referents are to be described (Clark & Wilkes-Gibbs, 1986), as well as lexical entrainment, the process whereby speakers 'agree' on how to verbalise properties of a referent (Brennan & Clark, 1996). A related, though somewhat contrasting view on dialogue, is offered by the Interactive Alignment model of Pickering and Garrod (2004), who suggest that the alignment phenomena observed in several studies are ultimately due to priming mechanisms, whereby speakers eventually converge on their representations at multiple levels. This more mechanistic view, if correct, would obviate the need for explicit reference to collaboration and negotiation in models of dialogue. Indeed, some recent work by Goudbeek and Krahmer (2012) has shown that speakers' use of properties which are typically dispreferred, as well as their tendency to overspecify, increase when they have been exposed to descriptions which are either overspecified or contained dispreferred properties (see Gatt, Goudbeek, & Krahmer, 2011, for a computational model that seeks to account for these results).

Several recent studies of reference production have raised interesting questions on the extent to which language producers engage in 'audience design', with consequences for classical models of content selection that emphasise preference and overspecification. Engelhardt et al. (2006) showed that speakers tend to produce overspecified descriptions, although they also tended to judge them more negatively than non-overspecified ones in an offline task (but see Davies & Katsos, 2013, for a criticism of the conclusions reached in this paper). More recently, studies using Event-Related Potentials (ERPs) by Engelhardt et al. (2011) have suggested that overspecification may not be optimal for listeners. This is in line with a prediction that emerges from eye-tracking research on reference comprehension in the Visual World paradigm (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Sedivy, 2003; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In these studies, listeners have been shown to entertain different referential possibilities in the visual domain only up to the point where the description disambiguates the target referent; further information is then redundant. In the present volume, Engelhardt and Ferreira (2014) shed new light on these questions from a production perspective, with evidence that modifiers which are strictly unnecessary for identification tend to be articulated differently (for example, they have shorter duration) than contrastive modifiers. This suggests that speakers do distinguish between discriminatory and redundant properties at some level, though this is not necessarily under conscious control.

The idea that language producers may not explicitly take into account their addressees' discourse models has been at the heart of the so-called 'egocentricity' debate (Horton & Keysar, 1996; Keysar, Barr, Balin, & Baruner, 2000; Keysar, Lin, & Barr, 2003), with evidence suggesting that there are limits in the extent to which producers can entertain potentially conflicting constraints in making

their choices. As we saw in our discussion of choice of anaphors, recent research suggests that producers do take their addressee's knowledge into account in choosing a referential form, but perhaps not in a fine-grained manner. In relation to conceptualisation, Wardlow Lane, Groisman, and Ferreira (2006) show that speakers sometimes select properties to distinguish a target referent even though they are aware that the distractors from which they seek to distinguish the target are not visible to their interlocutors. This evidence suggests that speakers are subject to constraints that are 'speaker-internal' and that these may override more 'addressee-oriented' processes under certain conditions (Arnold, 2008).

A rather different approach to interlocutor sensitivity is offered by Garoufi and Koller (2014, this vulume), Paraboni and van Deemter (2013, this volume) and Foster, Giuliani, and Isard (2014, this volume). All are concerned with computational models that seek to maximise utility, in the sense that they facilitate comprehension for potential addressees. Garoufi and Koller (2014) combine a model for planning the content of referring expressions with heuristics, which are found to be helpful to addressees. These heuristics are identified using machine learning methods (Maximum Entropy or logistic regression models) in a corpus of dialogues in which addressees need to follow instructions to navigate virtual environments (Gargett, Garoufi, Koller, & Striegnitz, 2010). Garoufi and Koller (2014) show that the expressions produced by this model are more helpful than those of baseline REG models. Interestingly, the model also turns out to be more humanlike. Paraboni and van Deemter (2013) focus on reference in spatial and hierarchical domains (e.g. a document divided into sections and subsections), where overspecification has been found to facilitate identification of a referent to a degree that is not predicted by standard REG algorithms (Paraboni et al., 2007). Their new findings, which are incorporated in a new algorithm, suggest that overspecification plays a facilitative role for a listener under conditions where finding a path through a space may be problematic. Foster et al. (2014) are also concerned with the task of reference generation in situated context, describing a study to evaluate a context-sensitive REG algorithm used in an interactive humanoid robot. While the objective evaluation measures used in their study do not distinguish this algorithm from the baselines against which it is compared, subjective evaluations do suggest that users prefer interactions when these involve the context-sensitive strategy.

Thus, models which incorporate some degree of alignment or sensitivity to the addressee perform better than their classic, more 'egocentric', counterparts. Note, however, that models such as those of Garoufi and Koller (2014) and Paraboni and van Deemter (2013) were evaluated not only against production data but also in terms of their utility where addressees are concerned. This is an instance of the distinction between 'producer-oriented' and 'addressee-oriented' evaluation mentioned in the Introduction. To the extent that models which incorporate addressee-oriented heuristics match production data better than those which do not, they suggest that producers are not entirely egocentric. However, current accounts are not sufficiently nuanced to enable a clear picture to emerge of the extent to which producers are allocentric.

To summarise the overview in this section, there is a significant degree of convergence between psycholinguistic and computational work on conceptualisation in reference production. In both fields, research has addressed the importance of efficiency and overspecification. Recent empirical work has suggested that classic computational models do not generalise to non-monologue situations, with new approaches emerging that address some of the findings in the psycholinguistic literature on dialogue. These models are also raising new questions, especially in relation to the egocentricity or otherwise of producers' choices of content. This is a field that is ripe for further exploration.

Conclusion

This paper has given an overview of work on reference production, in both monologue and dialogue settings, as an introduction to the present volume on *Models and Empirical Data for the Production of Referring Expressions*. It has focussed on two of the choices faced by the language producer: (1) choice of referential form, especially the conditions under which speakers choose to use reduced noun phrases and pronouns; and (2) conceptualisation, that is, which properties of a referent are included in a descriptive noun phrase. We have looked at these questions both from an empirical, psycholinguistic perspective and the computational modelling perspective, arguing that there is substantial overlap in the interests of these two communities, though there is also much scope for further cross-fertilisation.

Our hope is that more interdisciplinary work of the kind presented here and elsewhere in this volume will come to light. This will require more synergy among researchers with an interest in computational modelling, and those with an interest in experimental methods. The success of some recent initiatives, such as the organisation of a series of workshops on Production of Referring Expressions, held in conjunction with the Annual Meetings of the Cognitive Science Society in 2009, 2011 and 2013, suggest that there is a growing awareness of this need. We hope the present volume will serve to highlight fruitful avenues for research and collaboration.

Acknowledgements

The authors gratefully acknowledge the support of the Cognitive Science Society for the organisation of the Workshop on Production of Referring Expressions: Bridging the Gap between Cognitive and Computational Approaches to Reference, from which this special issue originated.

Funding

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* (grant number 277-70-007).

Notes

1. In what follows, we will sometimes use the term cognitive modelling, with the understanding that the models under consideration have been explicitly developed with a view to model the results of experimental findings related to human speech and reference production and to characterise or explain the underlying cognitive processes.
2. A classic example – one among many application areas in NLG – is the generation of a report that summarises raw data (such as meteorological or clinical data) to facilitate human access to relevant information (e.g. Gatt et al., 2009; Goldberg et al., 1994; Portet et al., 2009; Reiter et al., 2005). Such systems incorporate REG algorithms to generate references to domain entities.
3. The extent to which Grice intended his maxims as 'rules' governing conversation is, of course, debatable. What follows should not be read as a suggestion that speakers somehow 'violate' Gricean rules; rather, we are interested here in the predictions that stem from a specific interpretation of the Maxim of Quantity, applied to simple referential tasks.
4. For example, once the IA selects colour in Figure 1(b), only distractors with a different colour remain to be excluded and this motivates the choice of the next property to be included in the description.

References

Abbott, B. (2010). *Reference*. Oxford: Oxford University Press.

Anderson, J., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036–1060. doi:10.1037/0033-295X.111.4.1036

Ariel, M. (1990). *Accessing noun phrase antecedents*. London: Routledge.

Ariel, M. (2001). Accessibility theory: An overview. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects*. (pp. 29–88). Amsterdam: John Benjamins.

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, *31*, 137–162. doi:10.1207/S15326950DP3102_02

Arnold, J. E. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, *23*, 495–527. doi:10.1080/01690960801920099

Arnold, J., Bennetto, L., & Diehl, J. (2009). Reference production in young speakers with and without autism: Effects of discourse status and processing constraints. *Cognition*, *110*(2), 131–146. doi:10.1016/j.cognition.2008.10.016

Arnold, J., & Griffin, Z. (2007). The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, *56*, 521–536. doi:10.1016/j.jml.2006.09.007

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

Bard, E., Hill, R., Foster, M., & Arai, M. (2014). Tuning accessibility of referring expressions in situated dialogue. *Language, Cognition and Neuroscience*, this volume.

Belke, E. (2006). Visual determinants of preferred adjective order. *Visual Cognition*, *14*, 261–294. doi:10.1080/13506280500260484

Belke, E., & Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. *European Journal of Cognitive Psychology*, *14*, 237–266. doi:10.1080/09541440143000050

Belz, A., Kow, E., Viethen, J., & Gatt, A. (2010). Generating referring expressions in context: The GREC shared task evaluation challenges. In E. Krahmer & M. Theune (Eds.), *Empirical methods in natural language generation* (vol. 5980, pp. 294–327). Berlin and Heidelberg: Springer.

Bock, J., & Warren, R. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, *21*(1), 47–67. doi:10.1016/0010-0277(85)90023-X

Boden, M. (2008). *Mind as machine: A history of cognitive science* (vol. 1 & 2). Oxford: Oxford University Press.

Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, *10*, 137–167. doi:10.1080/01690969508407091

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

Brennan, S., Friedman, M., & Pollard, C. (1995). A centering approach to pronouns. In *Proceedings of the 25th annual meeting of the association for computational linguistics (acl'95)* (pp. 155–162). Stroudsburg, PA: Association for Computational Linguistics.

Brown-Schmidt, S., Byron, D., & Tanenhaus, M. (2005). Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language*, *53*, 292–313. doi:10.1016/j.jml.2005.03.003

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. In C. Li (Ed.), *Subject and topic* (pp. 25–53). New York, NY: Academic Press.

Clark, H. (1996). *Using language*. Cambridge, MA: Cambridge University Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. doi:10.1016/0010-0277(86)90010-7

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. Chicago, IL: University of Chicago Press.

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). Stroudsburg, PA: Association for Computational Linguistics.

Dale, R., & Reiter, E. (1995). Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*, 233–263. doi:10.1207/s15516709cog1902_3

Davies, C., & Katsos, N. (2013). Are speakers and listeners 'only moderately Gricean'? an empirical response to Engelhardt et al. (2006). *Journal of Pragmatics*, *49*(1), 78–106. doi:10.1016/j.pragma.2013.01.004

Deutsch, W., & Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition*, *11*, 159–184. doi:10.1016/0010-0277(82)90024-5

Di Fabbrizio, G., Stent, A. J., & Bangalore, S. (2008). Trainable speaker-based referring expression generation. In *Proceedings of the 12th conference on computational natural language learning (conll'08)* (pp. 151–158). Stroudsburg, PA: Association for Computational Linguistics.

Eberhard, K., Spivey-Knowlton, M., Sedivy, J., & Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*, 409–436. doi:10.1007/BF02143160

Engelhardt, P., Bailey, K., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, *54*, 554–573. doi:10.1016/j.jml.2005.12.009

Engelhardt, P., Demiral, S. B., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, *77*, 304–314. doi:10.1016/j.bandc.2011.07.004

Engelhardt, P., & Ferreira, F. (2014). Do speakers articulate over-described modifiers differently from modifiers that are required by context?: Implications for models of reference production. *Language, Cognition and Neuroscience*, this volume.

Fletcher, C. R. (1984). Markedness and topic continuity in discourse processing. *Journal of Verbal Learning and Verbal Behavior*, *23*, 487–493. doi:10.1016/S0022-5371(84)90309-8

Ford, W., & Olson, D. (1975). The elaboration of the noun phrase in children's description of objects. *Journal of Experimental Child Psychology*, *19*, 371–382. doi:10.1016/0022-0965(75)90068-5

Foster, M., Giuliani, M., & Isard, A. (2014). Task-based evaluation of adaptive referring expressions in human-robot dialogue. *Language, Cognition and Neuroscience*, this volume.

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998. doi:10.1126/science.1218633

Frege, G. (1892). Über sinn und bedeutung [On sense and reference]. *Zeitschrift für Philosophie und Philosophische Kritik*, *C*, *100*, 25–50.

Fukumura, K., Hyönä, J., & Scholfield, M. (2013). Gender affects semantic competition: The effect of gender in a non-gender marking language. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*, 1012–1021. doi:10.1037/a0031215

Fukumura, K., & van Gompel, R. (2011). The effect of animacy on the choice of referring expression. *Language and Cognitive Processes*, *26*, 1472–1504. doi:10.1080/01690965.2010.506444

Fukumura, K., & van Gompel, R. (2012). Producing pronouns and definite noun phrases: Do speakers use the addressee's discourse model? *Cognitive Science*, *36*, 1289–1311. doi:10.1111/j.1551-6709.2012.01255.x

Fukumura, K., & van Gompel, R. (2014). Effects of order of mention and grammatical role on anaphor resolution. *Journal of Experimental Psychology: Learning, Memory and Cognition*, under revision.

Fukumura, K., van Gompel, R., Harley, T., & Pickering, M. (2011). How does similarity-based interference affect the choice of referring expression? *Journal of Memory and Language*, *65*, 331–344. doi:10.1016/j.jml.2011.06.001

Fukumura, K., van Gompel, R., & Pickering, M. (2010). The use of visual context during the production of referring expressions. *The Quarterly Journal of Experimental Psychology*, *63*, 1700–1715. doi:10.1080/17470210903490969

Fukumura, K., & van Gompel, R. P. (2010). Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language*, *62*(1), 52–66. doi:10.1016/j.jml.2009.09.001

Gardent, C. (2002). Generating minimal definite descriptions. In *Proceedings of the 40th annual meeting of the association for computational linguistics (acl'02)*. Stroudsburg, PA: Association for Computational Linguistics.

Gargett, A., Garoufi, K., Koller, A., & Striegnitz, K. (2010). The give-2 corpus of giving instructions in virtual environments. In *Proceedings of the 7th international conference on language resources and evaluation (lrec'10)*. Valletta: European Language Resources Association (ELRA).

Garoufi, K., & Koller, A. (2014). Generation of effective referring expressions in situated context. *Language, Cognition and Neuroscience*, this volume.

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 and Heidelberg: Springer.

Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Attribute preference and priming in reference production: Experimental evidence and computational modeling. In *Proceedings of the 33rd annual meeting of the cognitive science society (CogSci'11*, pp. 2627–2632*)*.

Gatt, A., Portet, F., Reiter, E., Hunter, J., Mahamood, S., Moncur, W., & Sripada, S. (2009). From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *AI Communications*, *22*, 153–186.

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 (PRE-CogSci'11)*. http://pre2011.uvt.nl

Givón, T. (1983). *Topic continuity in discourse*. Amsterdam: John Benjamins.

Goldberg, E., Driedger, N., & Kittredge, R. I. (1994). Using natural language processing to produce weather forecasts. *IEEE Expert*, *9*(2), 45–53. doi:10.1109/64.294135

Gordon, P., & Chan, D. (1995). Pronouns, passives, and discourse coherence. *Journal of Memory and Language*, *34*, 216–231. doi:10.1006/jmla.1995.1010

Gordon, P., Grosz, B., & Gilliom, L. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, *17*, 311–347. doi:10.1207/s15516709cog1703_1

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

Grosz, B., Joshi, A., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st annual meeting on association for computational linguistics* (pp. 44–50). Stroudsburg, PA: Association for Computational Linguistics.

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

Guhe, M. (2012). Utility-based generation of referring expressions. *Topics in Cognitive Science*, *4*, 306–329. doi:10.1111/j.1756-8765.2012.01185.x

Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*, 274–307. doi:10.2307/416535

Gundel, J., Hedberg, N., & Zacharski, R. (2012). Underspecification of cognitive status in reference production: Some empirical predictions. *Topics in Cognitive Science*, *4*, 249–268. doi:10.1111/j.1756-8765.2012.01184.x

Horton, W., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91–117. doi:10.1016/0010-0277(96)81418-1

Jaeger, T. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, *61*(1), 23–62. doi:10.1016/j.cogpsych.2010.02.002

Jordan, P. (2002). Contextual influences on attribute selection for repeated descriptions. In K. van Deemter & R. Kibble (Eds.), *Information sharing: Reference and presupposition in natural language generation and understanding*. (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.

Kaiser, E. (2011). Focusing on pronouns: Consequences of subjecthood, pronominalisation, and contrastive focus. *Language and Cognitive Processes*, *26*, 1625–1666. doi:10.1080/01690965.2010.523082

Kaiser, E., & Trueswell, J. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition*, *94*(2), 113–147. doi:10.1016/j.cognition.2004.01.002

Kaiser, E., & Trueswell, J. (2008). Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes*, *23*, 709–748. doi:10.1080/01690960701771220

Keenan, E., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, *8*(1), 63–99.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. (2008). Coherence and coreference revisited. *Journal of Semantics*, *25*(1), 1–44. doi:10.1093/jos/ffm018

Keysar, B., Barr, D., Balin, J., & Baruner, J. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38. doi:10.1111/1467-9280.00211

Keysar, B., Lin, S., & Barr, D. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41. doi:10.1016/S0010-0277(03)00064-7

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*, 3231–3250. doi:10.1016/j.pragma.2011.06.008

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, *37*, 395–411. doi:10.1111/cogs.12019

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In K. van Deemter & R. Kibble (Eds.), *Information sharing: Reference and presupposition in language generation and interpretation*. Stanford, CA: CSLI Publications.

Krahmer, E., & van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*, 173–218. doi:10.3758/BF03195480

Krahmer, E., van Erk, S., & Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics*, *29*, 53–72. doi:10.1162/089120102317341765

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

Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition*. London: Sage.

Li, C., & Thompson, S. (1976). Subject and topic: A new typology of language. In C. Li (Ed.), *Subject and topic*. (pp. 457–489). New York, NY: Academic Press.

Louwerse, M., Benesh, N., Hoque, M., Jeuniaux, P., Lewis, G., Wu, J., & Zirnstein, M. (2007). Multimodal communication in face-to-face computer-mediated conversations. In *Proceedings of the 28th annual conference of the cognitive science society (cogsci'07)* (pp. 1235–1240). Austin, TX: Cognitive Science Society.

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*, *4*, 184–210. doi:10.1111/j.1756-8765.2012.01181.x

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38. doi:10.1111/j.1756-8765.2008.01003.x

Newell, A. (1973). You can't play 20 questions with nature and win. In W. Chase (Ed.), *Visual information processing* (pp. 283–308). New York, NY: Academic Press.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, *77*, 257–273. doi:10.1037/h0029436

Paraboni, I., & van Deemter, K. (2013). Reference and the facilitation of search in spatial domains. *Language and Cognitive Processes*, this volume.

Paraboni, I., van Deemter, K., & Masthoff, J. (2007). Generating referring expressions: Making referents easy to identify. *Computational Linguistics*, *32*, 229–254. doi:10.1162/coli.2007.33.2.229

Passonneau, R. J. (1997). Interaction of discourse structure with explicitness of anaphoric noun phrases. In M. Walker, A. K. Joshi, & E. Prince (Eds.), *Centering in discourse*. (pp. 327–358). Oxford: Oxford University Press.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110. doi:10.1515/ling.1989.27.1.89

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

Poesio, M., Stevenson, R., Di Eugenio, B., & Hitzeman, J. (2004). Centering: A parametric theory and its instantiations. *Computational Linguistics*, *30*, 309–363. doi:10.1162/089120100750105966

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., & Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, *173*, 789–816. doi:10.1016/j.artint.2008.12.002

Reiter, E. (1990). The computational complexity of avoiding conversational implicatures. In *Proceedings of the 28th annual meeting on association for computational linguistics (acl'90)* (pp. 97–104). Stroudsburg, PA: Association for Computational Linguistics.

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

Reiter, E., Sripada, S., Hunter, J., Yu, J., & Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, *167*, 137–169. doi:10.1016/j.artint.2005.06.006

Rohde, H., & Kehler, A. (2013). Grammatical and information-structural influences on pronoun production. *Language and Cognitive Processes*, this volume.

Russell, B. (1905). On denoting. *Mind*, *14*, 479–493. doi:10.1093/mind/XIV.4.479

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23. doi:10.1023/A:1021928914454

Sonnenschein, S. (1982). The effects of redundant communications on listeners: When more is less. *Child Development*, *53*, 717–729. doi:10.2307/1129385

Sonnenschein, S. (1984). Why young listeners do not benefit from differentiating verbal redundancy. *Child Development*, *55*, 929–935. doi:10.2307/1130144

Stevenson, R., Crawley, R., & Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, *9*, 519–548. doi:10.1080/01690969408402130

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634. doi:10.1126/science.7777863

van Deemter, K., Gatt, A., van der Sluis, I., & Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, *36*, 799–836. doi:10.1111/j.1551-6709.2011.01205.x

van Deemter, K., Gatt, A., van Gompel, R., & Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics in Cognitive Science*, *4*, 166–183. doi:10.1111/j.1756-8765.2012.01187.x

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

Viethen, J., Dale, R., & Guhe, M. (2011). Serial dependency: Is it a characteristic of human referring expression generation? In *Proceedings of the workshop on production of referring expressions: Bridging the gap between computational and empirical approaches to reference (PRE-CogSci'11)*. Austin, TX: Cognitive Science Society. http://pre2011.uvt.nl

Viethen, J., Dale, R., & Guhe, M. (2014). Referring in dialogue: Alignment or construction? *Language, Cognition and Neuroscience*, this volume.

Viethen, J., Goudbeek, M., & Krahmer, E. (2012). The impact of colour difference and colour codability on reference production. In *Proceedings of the 34th annual meeting of the cognitive science society (CogSci'12)*. (pp. 1084–1089). Austin, TX: Cognitive Science Society.

Vogels, J., Krahmer, E., & Maes, A. (2013a). When a stone tries to climb up a slope: The interplay between lexical and perceptual animacy in referential choices. *Frontiers in Psychology*, *4*, 1–15. doi:10.3389/fpsyg.2013.00154

Vogels, J., Krahmer, E., & Maes, A. (2013b). Who is where referred to how, and why?: The influence of visual saliency on referent accessibility in spoken language production. *Language and cognitive processes*, *28*, 1323–1349. doi:10.1080/01690965.2012.682072

Wardlow Lane, L., Groisman, M., & Ferreira, V. (2006). Don't talk about pink elephants!: Speakers' control over leaking private information during language production. *Psychological Science*, *17*, 273–277. doi:10.1111/j.1467-9280.2006.01697.x