

Research proposal

2. Description of the proposed research

A. Research topic

1 Overall aim, scientific background, key objectives

1.1 Overall aim

Consider this simple scene and suppose you want to point out one of these persons to an addressee:

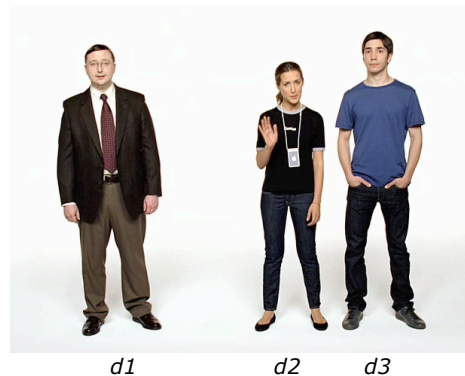


FIGURE 1: Example scene

There are many ways in which you could do this, but most speakers have no difficulty in quickly deciding which information to include and how to realize this information verbally and non-verbally via **referring expressions** like “the man in a suit,” “the woman” and “the younger-looking man”. How do speakers do this?

Both in psycholinguistics and computational linguistics, researchers have addressed aspects of this question, but we still have an incomplete understanding of how the human production of referring expressions works. This proposal argues that for a better, more complete understanding of this process, it is important to bridge the gap between the two disciplines. Such a bridge has both theoretical and methodological advantages. Psycholinguistics has important insights to offer in the human production of referring expressions and a methodology involving carefully constructed and controlled experiments; computational linguistics a well-established approach involving corpus analysis and computational modeling. Together they can be harnessed to achieve new scientific insights in the human production of referring expressions which in turn may result in a better understanding of speech production in general, especially where the interactions between speaker and addressee, between planning and realization, between verbal and non-verbal communication, and between scene perception and speech production are concerned. This also has societal and technological relevance, since such insights are important for developing software characters with human-like appearances and with verbal and non-verbal communicative capabilities (“virtual humans”).

1.2 Scientific background

1.2.1 The computational perspective

In computational linguistics, the production of referring expressions has been studied primarily in the subfield known as **Natural Language Generation (NLG)**. NLG is the process of automatically converting non-linguistic information (e.g., from a database)

into coherent natural language text, which is useful for many practical applications ranging from automatically generated weather forecasts to summarizing medical information in a patient-friendly way (Reiter & Dale 2000). Generating referring expressions is a core task in NLG, and has received much attention recently (Bateman 1999, Dale & Reiter 1995, van Deemter 2006, Gardent 2001, Gatt 2006, Horacek 2005, Krahmer et al. 2003, Siddharthan & Copestake 2004, among others). In NLG, the focus is usually on distinguishing descriptions: given a scene with various objects, describe one target object (say, *d1* in our example) by singling it out from the other objects (the distractors, *d2* and *d3*). This is essentially a problem of choice: there are many ways in which *d1* could be described ("the man on the left", "the man with glasses", etc.), and it has to be decided which is optimal in the given context.

NLG researchers have proposed different interpretations of what makes a referring expression optimal. A popular interpretation is that minimal references are optimal, containing precisely enough information to single out the target (e.g., Viethen & Dale 2006), in line with Grice's (1975) Maxim of Quantity ("be as informative as necessary"). However, finding such minimal, "Gricean" references is computationally expensive (Garey & Johnson 1979), and may take considerable time for realistic domains. Moreover, there is a growing awareness that the descriptions produced by these algorithms are rather different from the ones produced by human speakers (e.g., van der Sluis & Krahmer 2007). In addition, virtually all NLG algorithms focus on the production of (written) text, which limits their applicability for the development of virtual humans. To address these problems, more knowledge is required of how real humans realize referring expressions through speech, and support what they are saying with visual cues.

1.2.2 The human perspective

The human production of referring expressions is studied primarily in **Psycholinguistics**. Arguably, the computational, mechanistic view described above corresponds to the "language-as-product" view (Clark 1996) in "traditional" psycholinguistics. Clark (1996, 1997) contrasts this with the "language-as-process" view, which emphasizes that utterances are produced in a specific context, in cooperation between speaker and addressee. The gradual shift in recent years from the product to the process view has resulted in new insights in the human production of referring expressions. For instance, human speakers tend to produce *overspecified* rather than minimal, Gricean expressions (describing *d3* as "the younger-looking man on the right", although *d3* is the only younger-looking man; Maes et al. 2004, Engelhardt et al. 2006). Arguably, human speech is such an efficient medium that uttering a few potentially superfluous words is hardly a loss of time, but why and how speakers overspecify remains unclear. It might be that speakers overspecify because it simplifies their own search process (e.g., Engelhardt et al. 2006), but it is also conceivable that they overspecify because it helps the addressee in interpreting the referring expression (e.g., Paraboni et al. 2007).

The latter option is consistent with another finding of the process-view typically not captured in current generation algorithms, namely that speakers take the addressee into account when referring (an instance of "audience design", Clark & Murphey 1983). Again, the details are unclear; some argue that speakers only have limited capabilities for considering the addressee's perspective (Horton & Keysar 1996, Keysar et al. 2003), while others offer empirical evidence that referring is indeed an interactive process, with speaker and addressee forming a "conceptual pact" on how to refer to some object (Clark & Wilkes-Gibbs 1986, Brennan & Clark 1996, Metzging & Brennan 2003). Making such a conceptual pact can be seen as a general instance of what Pickering and Garrod (2004) call *alignment*, in their mechanistic theory of dialogue (essentially a proposal to reunite the language-as-product and language-as-process views).

It is interesting to observe that generating overspecified expressions is computationally cheaper than producing minimal ones (Dale & Reiter 1995), and, in a similar vein, it can be argued that audience design and alignment can reduce the search space of the generation algorithm, since they limit the number of possibilities that have to be considered. This suggests that some of the aforementioned problems for NLG algorithms can be addressed simultaneously by paying heed to human reference. The situation is more complex for another problem of the computational perspective: the restriction to text. Traditionally, psycholinguists focused almost exclusively on language (textual or spoken). However, when humans produce referring expressions, they do more than just produce language; e.g., they may produce gestures to express some property of the target (e.g., “the man standing like *this*” [uttered with hands in pockets]) or they may single out the target with a pointing gesture (“*this* one”). Besides such representational gestures, speakers may also produce non-representational gestures (such as beats) to emphasize words, and facial expressions (eyebrow movements, head nods, etc.) may be used for these purposes as well. In recent years researchers have started exploring “visual speech” (including facial expressions and gestures), and its relation to “auditory speech” (e.g., Alibali et al. 2000, Barkhuysen et al. 2007, Clark & Krych 2004, Kita and Özyürek 2003, Krahmer & Swerts 2005, 2007, Munhall et al 2004, de Ruiter 2000, Srinivasan & Massaro 2003, Swerts & Krahmer 2005, 2007). However, many open questions remain about the relation between auditory and visual speech.

Even though psycholinguistics has produced many interesting findings about human reference, using these for the development of virtual humans is challenged by two factors. First, psycholinguistic theories say little about how speakers quickly decide which properties, from the large set of potential ones, to use in a referring expression. Second, these theories often rely on intuitive but complex notions like audience design or alignment, and a common criticism is that they would greatly benefit from “explicit computational modeling” (Brown-Schmidt & Tanenhaus 2004). Solving choice problems and explicit modeling are precisely what the computational perspective has to offer.

1.3 Problem statement and key objectives

Both computational linguistics and psycholinguistics have yielded important insights in the production of referring expressions, but the different approaches also have their own sets of limitations. Interestingly, the limitations and strong points appear to be largely complementary, so bridging the gap between the disciplines offers an important step forward in answering this proposal’s research question: How do human speakers go from the intention to point out an object to an addressee to the realization of this intention through spoken language, supported with appropriate gestures and facial expressions?

2. Bridging the gap: Original and innovative aspects of the topic

It is commonly assumed that human speech production involves several modules (e.g., Barrett & Kurzban 2006). Psycholinguists have postulated various modular models for the speech production process, of which Levelt’s (1989, 1999) detailed Blueprint for the speaker is the prime example. These models may be informed by production data (e.g., self-corrections, temporal information) but also by experimental results or by on-line measurements including eye-tracking. Most researchers in NLG likewise opt for a modular approach (Reiter & Dale 2000). In part this is for practical reasons (allowing re-use of existing software-modules; Reiter 2000), but it also allows researchers to compare the performance of alternative modular models (e.g., van den Bosch 1997).

Even though details differ and researchers may disagree in which module a certain

process takes place, a common distinction is made between “deciding-what-to-say” and “deciding-how-to-say-it”. In Levelt’s Blueprint, the former is done by the Conceptualizer and the latter by the Formulator, while in the typical NLG architecture these tasks are performed by the Planning and Realizer modules respectively (Figure 2). Throughout this project, both deciding-what-to-say and deciding-how-to-say-it will be systematically addressed, giving rise to two broad research questions:

- 1. Deciding-what-to-say (Planning)** How do speakers decide which information to include in a referring expression, taking the context including their addressee into account, and how can this process be made computationally explicit?
- 2. Deciding-how-to-say-it (Realization)** How do speakers realize this information, both verbally and non-verbally, and how can this process be made computationally explicit?

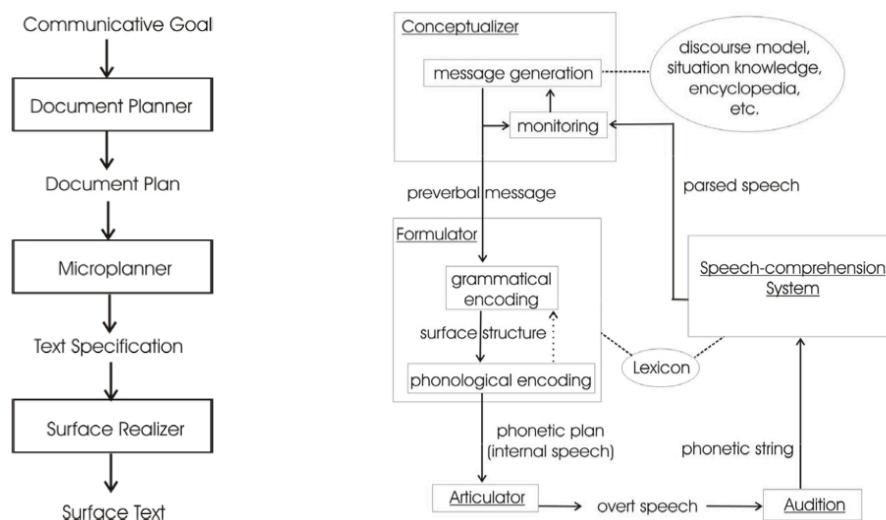


FIGURE 2: Left: common NLG architecture. Right: Blueprint for the speaker.

2.1 On Planning

In one of the first studies addressing the automatic generation of referring expressions, Appelt (1985) describes a plan-based system in which generation is modeled as an interactive process between system/speaker and addressee, with the system reasoning about its own goals as well as the derived goals from the addressee. Interestingly, this work (and more recent follow-up work such as Kronfeld 1986 and Heeman & Hirst 1995) is similar in spirit to the "language-as-process" view from psycholinguistics (although, like most "early" computational linguistics, it is logic-based and not empirically oriented).

Even though referring acts have been incorporated in such planning-based models, they generally do not address determining the content of referring expressions. This aspect does play a central role in more recent proposals for the generation of referring expressions, of which the Incremental Algorithm (Dale & Reiter 1995) is the algorithm-of-choice in many applications, due to its flexibility and conceptual transparency. The Incremental Algorithm considers potential properties in a specific order, based on the idea that speakers prefer certain properties when referring to objects. For instance, most speakers first try to describe an object in terms of its "type" (is it a man or a woman?). If that does not suffice, additional properties are tried. Following Pechman (1989), it is assumed that absolute properties ("wearing glasses") are preferred over relative ones

("younger-looking"). The intuition is that relative properties require inspection of the entire scene (a person is "younger-looking" compared to others) while absolute properties only require inspection of the target. The Incremental Algorithm checks properties in the predetermined order, selecting a property if it rules out one of the distractors not previously ruled out.

In this way, the Incremental Algorithm can account for relatively simple object descriptions, but human speakers, of course, have a wider repertoire of possible references at their disposal. Much of the recent interest in the generation of referring expressions has focused on extending the coverage of the Incremental Algorithm, to include, for instance, pointing gestures ("this one" + pointing), negations ("the man without glasses"), and relations ("the man left of the woman"). Arguably, this results in "an embarrassment of riches": there are so many ways in which an object can be referred to, that finding the optimal reference becomes increasingly difficult and computationally expensive (van Deemter and Krahmer 2007). Moreover, even though the coverage has been extended substantially, neither the Incremental Algorithm nor any of its competitors is capable of producing referring expressions with the flexibility and naturalness of humans in actual conversations.

It is interesting to observe that, complex as this task may be in theory, it is one that human speakers perform quickly and mostly without conscious effort. This suggests that human speakers use certain strategies that are not incorporated in algorithms such as the Incremental one. I argue that computational accounts of strategies related to (1) scene perception, (2) overspecification and (3) audience design and adaptation are needed. This gives rise to the following three planning (P) issues.

P1: How does scene perception influence the production of referring expressions?

Not all objects in a scene are equally prominent; for instance, large objects at the center stand out more than small ones in the periphery. It is likely that this influences the production of referring expressions; e.g., it may take more effort to refer to the peripheral objects than to the salient objects (Krahmer & Theune 2002, Paraboni et al. 2007). In recent years, much has been learned about human scene perception, and these results can have a major impact on how generation algorithms operate. Our example scene is very simple, but most real-life scenes contain a multitude of objects that could be referred to or that may act as potential distractors to be ruled out. Just look up from the page and look around; notice that virtually any object in your field of vision could be referred to. It is highly unlikely that humans take all these objects into account when drafting a referring expression.

In real-life, it appears that humans employ two kinds of strategies when perceiving a scene (Itti & Koch 2000, Henderson 2003, Kelleher et al. 2005, Wooding et al. 2002). Wooding and colleagues analyzed eye-movement patterns from thousands of people watching paintings, and found that statistical properties of the image, such as changes in intensity and local contrasts, to a large extent determine view patterns. Besides such bottom-up strategies, determined by properties of the scene, top-down strategies also play a role: for instance, areas that are currently under discussion are looked at more frequently and for longer periods of time.

There is growing evidence that the visual system and the speech production system are closely intertwined (e.g., Meyer et al. 1998, Griffin & Bock 2000, Hanna & Brennan 2007, Spivey et al. 1999), but little is known about how scene perception influences the

production of referring expressions. Arguably, both bottom-up and top-down strategies make the problem of finding adequate referring expressions easier for speakers, because not all objects present in a scene need to be taken into account and the search space is considerably reduced. In this project we investigate how these recent findings from human scene perception can be modeled, and what the effects are of integrating these insights in the speaking process. For this a combination of eye-tracking studies and computational modeling will be used, outlined in section B (Approach).

P2: Why and how do speakers overspecify their references?

Human speakers do not necessarily produce minimal, “Gricean” references. Rather they frequently produce overinformative references, including more information than is strictly speaking necessary. Why?

Various explanations exist. As said, it might benefit the speaker, or the hearer (or both). Orthogonal to this, it has been argued that task factors may lead to overspecification as well. For instance, speakers tend to produce more overspecified references when the chances for misunderstanding certain properties are relatively high (van der Sluis and Kraemer 2007) or when the task is fault-critical, i.e., when producing references that may cause misunderstandings would be very costly (Arts 2004, Maes et al. 2004). In addition, taking real world scene perception into account may also explain some instances of overspecification; speakers presumably do not inspect the entire scene before starting to produce their referring expressions and hence cannot be certain whether a property is distinguishing or not, which may lead to unintended overspecifications. Moreover, it is unclear *how* human speakers overspecify. Interestingly, the Incremental Algorithm does occasionally produce overspecified references as a side-effect of the incremental search strategy. However, it does not offer a systematic account of human overspecification.

Arguably, such a systematic account of overspecification could simplify the search process (*not* overspecifying implies a costly search for the shortest expression). In this project, we use experimental studies and computational modeling to find out which factors contribute to speakers producing overspecified references, and to devise an automatic strategy capable of generating such references as human speakers do.

P3: How do audience design and adaptation work and how to model it computationally?

Human speakers do not produce referring expressions in a vacuum, but they do so for a particular addressee. It is commonly assumed that human speakers do engage in audience design when referring, although the extent to which they do this is still under discussion. Moreover, there are still gaps in our understanding of these processes that so far have hindered a precise explication of them.

In particular, even though the evidence of Brennan and Clark (1996), Metzging and Brennan (2003) and others shows that speakers adapt to their addressees by agreeing on the usage of certain words and phrases, it is unclear *how* adaptation works precisely. The experimental evidence generally takes the following form: pairs of participants take turns in referring to objects, depicted on cards (Brennan and Clark 1996) or placed in an open grid positioned between the participants (Metzging and Brennan 2003). Results show that once speaker and addressee have agreed on a way to refer to an entity (a conceptual pact has been reached), they have a tendency to re-use this or similar expressions later on even though shorter variants may exist (which, incidentally, offers

another potential explanation of overspecification phenomena). It is unclear, however, how participants reach the conceptual pact precisely (which factors play a role? how does information about the context and the addressee influence this process?). In addition, the experiments are done with a limited number of simple objects (e.g., shoes, dogs, etc.) so that the number of choices to be considered during planning is limited.

For a better understanding of these processes, a series of experiments will be conducted with participants referring to objects in realistic scenes. It is a reasonable assumption that factors such as co-presence (are interlocutors in the same room?) and visibility (can they see each other?) influence the adaptation process (Mol et al. 2007, Hanna & Brennan 2007, Bard et al. 2007). Here, the settings of the experiments will be varied in such a way that the impact of such potential factors on adaptation can be studied. Based on this, automatic adaptation strategies will be implemented and integrated in the model.

2.2 On Realization

Besides deciding which information to include in a referring expression, a human speaker also has to decide how to realize this information. This involves at least a conversion of selected properties into natural language and determining an intonation contour (e.g., decide which words express important information and might be emphasized in speech by a pitch accent), and this "articulatory specification" (phonetic plan in Levelt's model) can then be uttered. However, there are at least two complications with this picture: (1) a speaker may decide to realize a certain property of the target (e.g., its shape or location) using a representational gesture, and (2) a speaker may also decide to emphasize a certain word with a non-representational gesture or with a facial cue such as a head nod or eyebrow movement.

Interestingly, both Reiter & Dale's NLG-architecture and Levelt's Blueprint only account for output in a single modality (text and speech, respectively). However, such a unimodal view of communication is not representative for the most archetypical communicative situation: a face-to-face setting in which speaker and addressee see and hear each other, and continuously pay attention to both auditory and visual cues. Many researchers nowadays assume that producing gestures and facial expressions is an integral part of utterance production, and that they should be studied together to fully understand human speech production (e.g., Alibali et al. 2000, Cassell et al. 2001, Clark & Krych 2004, Iversen & Goldin-Meadow 1998, McNeil 1992, Krahmer & Swerts 2007). Since the goal of this project is to model how speakers plan *and* realize referring expressions in conversation (both verbally and non-verbally), three realization (R) issues naturally arise.

R1: How to model the co-production of gestures and speech?

Speakers produce gestures and speech in tandem. Recent evidence even suggests that when speakers produce a gesture, this has a noticeable impact on acoustic speech production, indeed indicating a tight coupling between speech and gesture (Bernardis and Gentilucci 2006, Krahmer & Swerts 2007, Swerts & Krahmer 2006). But the jury is still out on *how* speakers co-produce speech and gesture.

This can be illustrated by comparing various models for the combined production of spoken language and manual gestures that were recently proposed (Kita and Özyürek 2003, Krauss et al. 1996, and de Ruyter 2000; all extensions of Levelt's blueprint). What these proposals have in common is the addition of a separate gesture stream, which has a shared source with the speech stream. The main difference between the proposed models lies in the location where the two streams (speech and gesture) part. According to Krauss and co-workers, for instance, this happens *before* conceptualization, while both

de Ruiter and Kita and Ozyürek argue that the separation takes place in the Conceptualizer. McNeill and Duncan (2000) take a markedly different view, arguing that speech and gesture should not be delegated to different streams, but rather are produced in close connection with each other. Thus, even though all agree that speech and gestures are related, they disagree on the locus of this relation.

In this project, we will contribute to this discussion, working from the hypothesis that the nature of the relation is different for different kinds of gestures. It is conceivable that representational gestures arise earlier in the speaking process than non-representational ones, which would suggest that the latter are more tightly connected to speech than the former. This can be tested through a combination of acoustic and perceptual measurements, as proposed by Krahmer and Swerts (2007) for beat gestures. A closely related question is how to computationally model the co-production of speech and gesture (cf. Cassell et al. 2001), which will be studied by experimentally comparing alternative models.

R2: How are visual cues timed and produced during referring expression production?

When humans speak, they frequently produce visual cues such as eyebrow movements or head nods. One important function of such cues is to indicate which words in the expression are important (e.g., because they express new information). In this respect, visual cues are similar to other aspects of intonation such as pitch accents (e.g., Bolinger 1985). In fact, it has recently been shown that placing such visual cues on important words indeed speeds up processing, and placing them on unimportant word hinders it (Krahmer & Swerts 2006, Swerts & Krahmer 2007), mimicking earlier results on the correct and incorrect placement of pitch accents (e.g., Terken & Nooteboom 1987).

Still speakers certainly do not associate every pitch accent with a corresponding visual cue. This raises the question how human speakers actually time and produce visual cues. Pelachaud et al. (1996) note that "there is a lack of empirical information on when an accent or other intonational components are accompanied by a facial action". The situation has not substantially changed in the last 10 years. To deal with this persisting lack of knowledge, an alternative strategy that is currently being explored is deriving this information from audiovisual corpora (e.g., Foster & Oberlander 2006, Nakano et al. 2003). In many subfields of computational linguistics, data-driven approaches have gained a lot of mileage in the past decade.

In this project we will investigate whether such a data-driven approach helps for deciding which visual cues should be realized while producing a referring expression. Alternative data-driven models for the timing and production of visual cues will be developed, and it will be tested which yield the best performance. Special attention will be given to the realization of visual cues for consecutive references to objects. Bard et al. (2000) have shown that a speakers' articulatory effort decreases for repeated references (first references are more intelligible, when presented in isolation, than later ones), and a specific hypothesis we will test is whether a similar decrease of effort is noticeable when focusing on visual cues (i.e., first references more visually marked than later ones).

R3: To what extent do speakers adapt their non-verbal realizations of referring expressions?

A central question in this research proposal is how speakers adapt to their addressees while producing referring expressions. The evidence gathered by Brennan, Clark and

colleagues indicates that speakers and addressees indeed adapt to each other while deciding what to say. Pickering and Garrod (2004) argue that adaptation occurs on all levels of communication, but adaptation on the non-verbal level has so far received relatively little attention, and the work that has been done mostly focuses on adaptation in speech (e.g., alignment of pitch level or volume).

Still, there is reason to believe that these effects may also occur on the level of visual cues and gestures. It is well known that seeing someone yawn triggers a yawn impulse, and, likewise, seeing someone smile may cause you to smile as well. In general, there are suggestive findings showing non-verbal mimicry among humans (Bailenson & Yee 2006, Chartrand & Bargh 1999, Sato & Yoshikawa 2007). This naturally raises the question how and to what extent human dialogue participants align their non-verbal behaviors while producing referring expressions. Non-verbal adaptation is different from, say, lexical adaptation, in at least two interesting ways. First, lexical adaptation is a more or less discrete process in which dialogue partners share words and phrases. Non-verbal adaptation is a more gradual process comparable to, say, phonetic alignment (cf. Krauss & Pardo 2004). Second, lexical adaptation presumably makes understanding each other's references easier, which may be less obvious for non-verbal adaptations. On the other hand, psychological research on mimicry clearly shows that unconscious non-verbal adaptations, serve other, social purposes (e.g., waitresses that mimic their customers receive higher tips; van Baaren et al. 2003).

In this project we will experimentally investigate how and to what extent dialogue participants, adapt their non-verbal communication to each other while referring, with gestures and with facial expressions. This knowledge forms an important counterpart to the insights obtained for issues R1 and R2, and taken together would greatly extend our knowledge of audiovisual speech realization.

B. Approach

1. Outline

The proposed approach starts from a corpus of audiovisual recordings of human speakers in natural conversations, supplemented with experimental data addressing various issues described above. On this basis, a computational model will be developed, mimicking the human production of referring expressions. The model will be demonstrated by embedding it in different virtual characters. Evaluation forms an integral part of the project, both against natural corpus data and in experiments with human participants.

2. Audiovisual data collection and experimental methodology

Starting point is an existing corpus of audiovisual recordings of human speakers, made within the Tilburg FOAP project, containing audiovisual speech from more than 300 different speakers of Dutch, and covering various communicative settings, ranging from answering questions to re-telling stories. This data-collection offers an excellent starting point for a data-driven analysis of timing and production of facial cues and gestures and their relation to speech (issue R2). To gain more insight in the human production of referring expressions (What is the influence of scene-perception (P1)? Why and how do speakers overspecify (P2)? How does gesturing relate to speech production (R1)? How do speaker and addressee adapt to each other, both verbally (P3) and non-verbally (R3)?), the corpus will be extended with additional audiovisual recordings. For this purpose a series of experiments will be conducted in which a human speaker communicates with an addressee, while referring to objects in a shared visual scene.



FIGURE 3: An example of a real life scene (from an amazon.com advertisement). Notice, for instance, "the small grey box on the upper shelf".

The experiments build on earlier experimental paradigms developed for studying shared references (Brennan & Clark 1996, Metzinger & Brennan 2003), but with a number of crucial differences. First, rather than only referring to simple objects, we will also study references to targets in real-life scenes, for which speakers will need to draft more complex references. Typically, these will be realistic depictions of neutral scenes, such as street views and home interiors (cf. Figure 3), which have been widely used in scene perception research (Oliva et al. 2004, Henderson & Ferreira 2004). Using such scenes, it is possible to systematically vary the presence of objects in the scene —as well as their properties— to allow for a wide variety of referring expressions. During these experiments naive participants perform identification tasks, for which they naturally need to refer to objects in the scenes. Eye-tracking tools will be used to find out which parts of these scenes are attended to by speaker and addressee while referring, to gain insights in how scene perception interacts with the production of referring expressions. To gain a better understanding of how conceptual pacts arise, the experimental setting of the experiments will be varied as well. By manipulating factors such as presence (are interlocutors in the same room?) and visibility (can they see each other?) (cf. Mol et al. 2007), we can gain a better understanding of the audience design and adaptation issues. In a similar vein, by manipulating task factors such as time pressure (are interlocutors forced to produce their references quickly?) and importance (how crucial is a proper understanding of references for the task?), we can improve our understanding of how such factors influence referential overspecification. Finally, to test more specific hypotheses related to adaptation, additional experiments will be conducted in which naive participants interact with an accomplice of the experimenter rather than with another naive participant (a common method in social psychology, also increasingly popular in psycholinguistics, e.g. Keysar et al. 2003). In this way, for example, we can gain a better understanding of non-verbal adaptation, by letting the accomplice use specific gestures in one condition and no gestures in the other.

3. Approach to Planning

Based on the collected data, computational models will be developed, modeling the human production of referring expressions. The models focus on producing correct, distinguishing references to objects in real world scenes (both initial and subsequent ones). The planning component takes as input a target object and yields a non-linguistic representation of the information the speaker wants to convey.

As one of our starting points for this component, we take the graph-based approach to generation proposed by Krahmer et al. (2003). In this approach scenes are formalized as

labeled directed graphs, expressing properties of objects present in a scene ("x is-male") and of relations between these objects ("x is-to-right-of y"). Planning is formalized as the search problem of finding a sub-graph of the scene-representation uniquely characterizing the target. There are many well-understood search algorithms for graph structures (e.g., Chartrand and Oellermann 1993, Cormen et al. 1990), and these can directly be applied. For the study described in Krahmer et al. (2003), an efficient Java-implementation was developed, which serves as the basis for the present project.

An important advantage of the graph-based perspective is that planning is conceptualized as a two-stage procedure: first, the algorithm computes what the possible referring expressions are for a given object, and subsequently it uses cost functions to rank these alternatives, where the cheapest, top-ranked solution is the one selected by the algorithm. What the cheapest solution is depends on the cost function used; it can be the minimal one, but more fine-grained solutions exist as well. A first approximation of such a more fine-grained cost function obtained state-of-the-art performance in the first referring expression evaluation competition (Theune et al. 2007).¹ The approach is general and flexible, and hence highly suitable for modeling human references.

To deal with the planning issues described above, various extensions to the basic graph-based approach are required. Knowledge of both top-down and bottom-up human scene perception should be modeled. One option we will explore is to treat these as filters of the original graph representing the entire scene, thereby reducing the search space. Krahmer and Theune (2002) have described a method of assigning costs to objects, which limits the attention of the algorithm to objects that are salient since they were recently under discussion (these are "cheaper" to refer to than non-salient ones). They show that existing methods such as those proposed by Grosz et al. (1995) or Gundel et al. (1993) can be used to assign salience weights to objects and to generate appropriate contextual references (pronouns, anaphoric descriptions, etc.) on this basis. In this project it will be studied whether this method (marking some objects as more salient than others) can be generalized to visual salience in scene perception as well.

In addition, a prominent place will be given to modeling the speaker's assumptions about the addressee. This calls for a reinterpretation of the graph-based generation method, which should no longer only be based on a graph representing the speaker's perception of the scene, but should also take the speaker's knowledge of the addressee into account (What is the position of the addressee with respect to the scene? What can be derived from the addressee's earlier utterances about her perception of the scene? Etc.). In a way, this amounts to reuniting the currently popular approaches to generating referring expressions (the Incremental Algorithm and its ilk) with the earlier plan-based approaches of Appelt (1985), Kronfeld (1986) and others. On the basis of the addressee-model, and taking the collected experimental data into account, audience design and alignment strategies will be developed and integrated in the model. Cost functions will be useful here as well, e.g., for making aligned variants cheaper than non-aligned ones, which would imply that the planning component has a general preference for aligned over non-aligned variants. To integrate the various knowledge sources (related to the scene, the context and the addressee) in single models, the usefulness of constraint satisfaction techniques will be investigated (e.g., Tsang 1993, Piwek & van Deemter 2006), where statistical information from the collected corpus will be used to derive the relevant weights.

4. Approach to Realization

The realization component takes a non-linguistic meaning representation (here: a graph

¹ <http://www.csd.abdn.ac.uk/research/evaluation/>

structure characterizing the target) as input, and converts it into a spoken referring expression, supported with appropriate gestures and facial expressions.

In this phase, a first decision that needs to be made is whether a certain non-linguistic potentially overspecified property should be realized through language or through a (representational) gesture. It has been argued that the graph-based algorithm is well equipped to automatically make this decision for pointing gestures (see Kraemer and van der Sluis 2004, van der Sluis and Kraemer 2007), and it will be investigated whether a similar strategy could be applied to other representational gestures as well. Once it has been decided which information should be realized through gestures and which through language, the actual "surface" realization can start, resulting in a linguistic referring expression, with a specification for gestures and facial expressions at appropriate places.

The emphasis will not be on grammatical and morphological realization (for which we will rely on existing techniques, see e.g., Belz 2005), but on the audiovisual realization. The data collection will be the starting point for modeling the timing and production of visual cues produced while referring. On the basis of corpus frequencies of different visual cues realized while referring, predictions can be made about the inclusion of different cues in new references. Various possibilities exist, ranging from modeling audiovisual behavior from one person to modeling average behavior of an entire group of speakers (Foster and Oberlander 2006), and these variants will be tested in the context of referring expression generation. Special attention will be given to the distribution of visual cues over repeated references, to test the hypothesis that the number of visual cues associated with a reference decreases over time.

The data obtained using the experimental methods outlined above will provide important insights into alignment of non-verbal cues, and taking the results from these different studies together, will allow us to make well-informed choices about the audiovisual realization of referring expressions. To demonstrate and test the resulting model, a proof-of-concept implementation will be developed, for which existing tools will be used: Ruth (de Carlo et al. 2004) and Greta (Pelachaud & Poggi 2002) (see Figure 4). For a given target, these characters then produce spoken referring expressions, accompanied by appropriate gestures and visual cues.



FIGURE 4: Stills of the virtual characters Greta (left) and Ruth (right).

5. Evaluation

Throughout the project, various alternative models for the human production of referring expressions will be developed and experimentally compared. Here, two evaluation methods will be used. The first is to compare the automatically generated audiovisual expressions for a target with human produced references for the same target. For this, standard methodology of corpus-based evaluations is applicable, plus additional metrics like Dice and PRP (Gatt et al. 2007). In the current setting, this kind of evaluation is

complicated by the fact that there is generally not one perfect solution: there will often be more than one way to efficiently refer to a target, or to realize a given expression (cf. Belz & Reiter 2006, Foster & Oberlander 2006, Viethen & Dale 2006, 2007). This implies that a corpus-based evaluation may not be sufficient: if the model predicts a different referring expression for a target than found in the corpus, this does not necessarily mean that the predicted output is wrong.

To compensate for this, the second evaluation method lets human participants process referring expressions in experimental settings. In a sense, this evaluation follows the opposite route: from an audiovisual referring expression to a target. On-line methods (eye-tracking, reaction time measurements) can be used to measure how human participants interpret a particular referring expressions and locate the intended target in the visual scene. Additionally, in this way it can be checked whether participants process aligned or overspecified references differently from non-aligned or minimally specified ones (reaction times might reveal that such references are processed quicker, eye-tracking may show that fewer potential distractors are inspected).

6. Knowledge dissemination

The team members will make their findings available to both the computational linguistics and psycholinguistics communities, at conferences such as ACL, AMLaP, CogSci, CUNY, LREC, and Interspeech, and in journals as *Computational Linguistics* (2.676), *Journal of the Acoustical Society of America* (1.677), *Journal of Memory and Language* (2.815), *Journal of Phonetics* (1.891), *Language and Cognitive Processes* (1.962), *Language and Speech* (1.138) and *Speech Communication* (1.178).² These journals are among the top ranking ones in the fields addressed by the project proposal. The proof-of-concept implementation will be another important tool for knowledge dissemination, allowing easy demonstration of the proposed model. Finally, the results of the project will be described in a monograph to be published at the end of the project.

C. Innovation / Significance & contributions to science, technology and society

Science. The research proposed here is aimed to considerably extend our scientific knowledge of human speech production, and its innovations are both methodological and theoretical. Bridging the gap between computational linguistics and psycholinguistics in the way proposed, has not been done before. In recent years there has been a growing interest in what has been called Computational Psycholinguistics (e.g., Dijkstra & de Smedt 1996, Jurafsky 2002), but when work in this field addresses speech production at all, it mostly focuses on the final stages of production (e.g., Levelt et al.'s 1999 WEAVER++ model for lexical access). The proposed combination of psycholinguistics and computational linguistics offers a more complete view of human production of referring expressions than currently available in either of the individual disciplines.

A further innovation is that planning and realization are studied in tandem. Studying these aspects in isolation may cause problems (Krahmer and Theune 2002, Horacek 2005, Stone et al. 2003), for instance, when information selected during planning cannot be realized adequately (e.g., because of linguistic constraints). Also the combination of speech and gesture favours an integrated approach.

Since many of the aspects of the project proposal do not exclusively address referring expressions, it paves the way for new insights in human speech production in general.

² Impact factors from the Journal Citation Reports (2005), Thomson ISI Web of Knowledge.

The proposal zooms in on referring expressions, since these form a well-delimited subpart of the speaking process, are ubiquitous in communication, and have been studied extensively both in computational linguistics and psycholinguistics. It has been argued that the processes involved in generating referring expressions are also applicable to other aspects of speech production, such as aggregation ("structure sharing"; Bateman 1999). The production of referring expressions is essentially a choice process, and a good understanding of how speakers make choices when drafting referring expressions is likely to help our understanding of how speakers make choices in other parts of the speaking process. In general, the project will lead to new scientific insights in the interaction between speaker and addressee, between verbal and non-verbal communication, and between scene perception and speech production, all of which are highly relevant for our understanding of the speaking process.

Technology. The proposal facilitates the technical development of "virtual humans", combining real-time language and speech technology with computer graphics (Cassell et al. 2001, DeCarlo et al. 2002, Gratch et al. 2002, Krahmer & Swerts 2004, Pelachaud et al. 1996). Due to technical progress and increasing computer power, virtual humans have become a practical reality in recent years. "Real" humans offered an important source of inspiration for their development, but developers soon realized that existing knowledge of how humans produce verbal and non-verbal behavior is underspecified in crucial respects. At the same time, evaluation studies of virtual humans reveal that real humans are very sensitive to the quality of the communicative behavior: if the verbal or non-verbal behavior of a virtual human is not adequate, people are negative in their evaluation of the character (e.g., Ruttkay & Pelachaud 2004). Moreover, for most practical applications, it would be beneficial if a virtual human could both automatically decide what to say and decide how to say it. Currently, these aspects are seriously understudied, and most virtual humans lack a detailed and computationally explicit model of audiovisual speech production.

Society. The current interest in virtual humans is largely motivated from the assumption that they may result in more natural and intuitive human-computer interaction, since they allow users to communicate with a computer as if they are communicating with another person. The alleged beneficial effect of virtual humans is known as the Persona-Effect (Dehn and van Mulken 2000), and successful applications have indeed been developed with virtual humans for gaming, domotica ("smart homes"), marketing and education (with the virtual human in the role of teacher/instructor). Yet, Dehn and van Mulken (2000) argue, on the basis of a meta-study of research on the benefits of virtual humans, that merely plugging a virtual human in an existing application is not sufficient for the persona-effect to arise. On top of that, the verbal and non-verbal behavior of the virtual human should be well developed, and this is exactly what the current project sets out to achieve.

D. Plan of work

1. Project team and Research Plan

The project will be carried out in close collaboration by a multi-disciplinary team of researchers, consisting of two PhDs and two PostDoc-researchers, under supervision of the applicant. To maximize overlap in staffing, one PhD student and PostDoc will have a background in psycholinguistics with a proven interest in computational modelling, the other PhD student and PostDoc will be computational linguists with expertise in human speech production; orthogonal to this, one PhD-PostDoc pair will work on planning issues (P1-P3), the other (starting a year later) will work on realization issues (R1-R3):

	Planning What to say?	Realization How to say it?
Human perspective	PostDoc1	PhD2
Computational perspective	PhD1	PostDoc2

To facilitate collaboration, the respective PostDocs are appointed simultaneously with the PhD students and will bring in their expertise throughout the project. PostDoc1, additionally, is in charge of the audiovisual corpus collection (both free and semi-controlled data) to which the PhDs also contribute. PostDoc2 is responsible for the integration, implementation and evaluation of the model in a virtual character. Moreover, there is support from student assistants for the entire duration of the project. They will help conducting the experiments, annotating data and implementing and evaluating the model. The applicant does daily project management, fostering multi-disciplinary collaboration within the team. His own research will address all the planning and realization issues described in this proposal, In addition, he has a special interest in developing and testing general models of human audiovisual speech production, based on the insights obtained in the project. He will be the main responsible one for the final monograph describing the findings of the project.

2. Practical timetable

The responsibilities of team members are schematized below; publishing is an integral part of the project and is not separately mentioned. Senior researchers are expected to publish at least one article in a journal mentioned in B.5 each year.

	200y	200y+1	200y+2	200y+3	200y+4
Applicant	Research (P1-3, R1-3) / Management	Research (P1-3, R1-3) / Management	Research (P1-3, R1-3) / Management	Research (P1-3, R1-3) / Management / Monograph	Research (P1-3, R1-3) / Management / Monograph
PhD1	Computational modeling / Scene perception (P1)	Computational modeling / Overspecification (P2)	Computational modeling / Adaptation (P3)	Write thesis	
PhD2		Speech and gesture (R1)	Timing and production of visual cues (R2)	Non-verbal Adaptation (R3)	Write thesis
PostDoc1	AV corpus collection / Scene perception (P1)	AV corpus collection / Overspecification (P2)	AV corpus collection / Adaptation (P3)		
PostDoc2		Implementation / Evaluation / Speech and gesture (R1)	Implementation / Evaluation / Timing and production of visual cues (R2)	Implementation / Evaluation / Non-verbal adaptation (R3)	
Support	Technical / Experimental Support	Technical / Experimental Support	Technical / Experimental Support	Technical / Experimental Support	Technical / Experimental Support

3. Collaboration

3.1 Local

The project will be carried out within the Department of Communication and Information Sciences (CIS) of the Faculty of Arts, in the Multimodality and Cognition (M&C) research programme: a young, enthusiastic programme (founded in 2004), bringing together experienced researchers from different backgrounds. The mission of this group is to find

out how humans produce and process multimodal information, and how such insights can help improve digital information presentation. The current proposal fits in very well with the mission and aims of this research programme. In a recent, independent Research Quality Assessment (December 2006), the programme scored a 4.5 (excellent / very good). The CIS department consists of an attractive mixture of experimental researchers (prof.dr. M. Swerts, prof.dr. A. Maes) and computational linguists (dr. A. van den Bosch, dr. E. Marsi, prof.dr. H. Bunt), with which the applicant frequently collaborates. The department has its own research laboratory, with advanced soft- and hardware (for eye-tracking, reaction time measurements, etc.) and excellent facilities for recording, storing, processing and analyzing digital video. For the current project, also the recent collaborations with the Psychology department are highly relevant (prof.dr. A. Vingerhoets, prof.dr. J. Vroomen).

3.2 National

Through participation in the national NWO-IMIX (Interactive Multimodal Information Extraction) programme, the applicant collaborates with researchers in Linguistics and Computer Science departments from the Universities of Groningen, Twente, Amsterdam, Nijmegen and Tilburg. There are particularly strong links with the Human-Media Interaction research group at Twente University (dr. M. Theune, drs. W. Bosma), through the joint IMIX-IMOGEN (Interactive Multimodal Output Generation) project. Within this project, an automatic NLG module is developed, in the context of a medical question-answering system, with special emphasis on information presentation through different modalities. With dr. M. Theune and dr. Zs. Ruttkay (Twente University) I have ongoing collaborations on the generation of referring expressions, and on the development and experimental evaluation of virtual humans. Via the Stan Ackermans Institute for User-System Interaction at the Eindhoven University of Technology, I collaborate with former colleagues from the Institute of Perception Research on the design and evaluation of efficient human-computer interaction.

3.3 International

For my general international activities (organizations, visits, etc.) I refer to section 4 below; here I limit myself to the most relevant collaborations for the present proposal. With dr. K. van Deemter (Aberdeen University), I have closely collaborated since the mid-nineties. From 2004 we were both involved in the TUNA EPSRC project addressing algorithms for the generation of referring expressions. This project, involving researchers from the University of Brighton, the Open University and Aberdeen University (all UK) as well as Tilburg University, provides some of the foundations for the current project proposal. The TUNA project recently ended, but collaboration will continue in the context of the present project proposal, in particular concerning (1) hearer-based evaluation and (2) referring expression generation in interaction. Part of the travel budget will be used for exchange-visits of researchers. Another relevant collaboration is with Antwerp University (prof.dr. W. Daelemans) in the STEVIN Daeso project (2006-2009). In this project automatic alignment techniques for detecting semantic overlap between sentences will be developed. Finally, two relevant collaborations are with dr. M. Stone (Rutgers University, U.S.A.) and prof.dr. C. Pelachaud (University of Paris-8, France), two of the leading researchers working on the development of virtual humans. With them I will collaborate on embedding the results of the present project in the virtual characters Greta and Ruth, respectively (new releases of each are scheduled for fall 2007).

[Word count: 7997; MAX 8000 ON 16 PAGES]

E. References

- Alibali, M., S. Kita, and A. Young (2000), Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes* 15, 593-613.
- Appelt, D. (1985), Planning English Referring Expressions, *Artificial Intelligence*, 26(1), 1-33.
- Arts, A. (2004), *Overspecification in instructive texts*, PhD Thesis, Tilburg University.
- van Baaren, R.B., Holland, R.W., Steenaerts, B., & van Knippenberg, A. (2003), Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, 39, 393-398
- van den Bosch, A. (1997), *Learning to Pronounce Written Words: A Study in Inductive Language Learning*, PhD Thesis, Maastricht University.
- Bailenson, J. and N. Yee (2006), Digital chameleons. Automatic assimilation of non-verbal gestures in immersive virtual environments, *Psychological Science*, 16, 814-819.
- Bard, E., A. Anderson, C. Sotillo, M. Aylett, G. Doherty-Sneddon and A. Newlands (2000), Controlling the Intelligibility of Referring Expressions in Dialogue, *Journal of Memory and Language*, 42, 1-22.
- Bard, E., A. Anderson, Y. Chen, H. Nicholson, C. Havard and S. Daizel-Job (2007), Let's you do that: Sharing the cognitive burdens of dialogue, *Journal of Memory and Language*, in press.
- Barkhuysen, P., E. Krahmer and M. Swerts (2005), Problem detection in human-machine interactions based on facial expressions of users, *Speech Communication*, 45(3), 343-359.
- Barkhuysen, P., E. Krahmer and M. Swerts (2007), The interplay between auditory and visual cues for end of utterance detection, under revision for *Journal of the Acoustical Society of America*.
- Barrett, H. and R. Kurzban (2006), Modularity in cognition: Framing the debate, *Psychological Review*, 113(3), 628-647.
- Bateman, J. (1999), Using aggregation for selecting content when generating referring expressions, *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL)*, Maryland
- Belz, A. (2005), Statistical generation: Three methods compared and evaluated, in: *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG)*, pp. 15-23.
- Belz, A. and E. Reiter (2006), Comparing automatic and human evaluation of NLG systems, in: *Proceedings of the European Chapter of the Association of Computational Linguistics (EACL)*, Trento, Italy.
- Bernardis, P. and M. Gentilucci (2006), Speech and gesture share the same communication system. *Neuropsychologia*, 44, 178-190.
- Bolinger, D. (1985), *Intonation and its Parts*, London: Edward Arnolds.
- Brennan, S. and H. Clark (1996), Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22, 1482-1493.
- Brown-Schmidt, S. and M. Tanenhaus (2004), Priming and alignment: Mechanism or Consequence? *Behavioral and Brain Sciences* 27, 193-194.
- Cassell, J., J. Sullivan, S. Prevost and E. Churchill (2000), *Embodied Conversational Agents*, MIT Press, Cambridge, MA.

- Cassell, J., Vilhalmsson, H. and Bickmore, T. (2001), BEAT: the Behavior Expression Animation Toolkit, *Proceedings of SIGGRAPH '01*, pp. 477-486, Los Angeles, CA.
- Chartrand, G. and O. Oellermann (1993), *Applied and Algorithmic Graph Theory*, McGraw-Hill, New York.
- Chartrand, T. and J. Bargh (1999), The chameleon effect: The perception-behavior link and social interaction, *Journal of Personality and Social Psychology*, 76, 893-910.
- Clark, H. (1996), *Using Language*, Cambridge University Press, Cambridge UK.
- Clark, H. (1997), Dogmas of Understanding, *Discourse Processes*, 23, 567-598.
- Clark, H. and Murphy, G. L. (1983). Audience design in meaning and reference. In J. F. LeNy & W. Kintsch (Eds.), *Language and comprehension* (pp. 287-299).
- Clark, H. and D. Wilkes-Gibbs (1986), Referring as a collaborative process. *Cognition*, 22, 1-39.
- Clark, H. and M. Krych (2004), Speaking while monitoring addressees for understanding, *Journal of Memory and Language* 50, 62-81.
- Cormen, T., C. Leiserson and R. Rivest (1990), *Introduction to Algorithms*, MIT Press, Cambridge.
- DeCarlo, D., M. Stone, C. Revilla and J. Venditti (2002), Specifying and animating facial signals for discourse in embodied conversational agents, *Computer Animation and Virtual Worlds* 15(1): 27-38.
- Dale, R. and E. Reiter (1995), Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science* 19:233-263
- Van Deemter, K. (2002), Generating Referring Expressions: Boolean Extensions of the Incremental Algorithm. *Computational Linguistics* 28 (1), 37-52.
- Van Deemter, K. (2006), Computational generation of vague descriptions, *Computational Linguistics* 32(2), 195-222.
- van Deemter, K. and E. Krahmer (2007), Graphs and Booleans, in: *Computing Meaning* (vol. 3), H. Bunt and R. Muskens (eds.), Studies in Linguistics and Philosophy, Kluwer Academic Publishers, 397-422.
- Dehn, D and S. van Mulken (2000), The impact of animated interface agents: a review of empirical research, *International Journal of Human-Computer Studies*, 33, 1-22.
- Dijkstra, T. and K. de Smedt (1996), *Computational Psycholinguistics: AI and Connectionist Models of Human Language Processing*, London: Taylor & Francis.
- Engelhardt, P., K. Bailey and F. Ferreira (2006), Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54, 554-573.
- Foster, M.E. and J. Oberlander (2006), Data-driven generation of emphatic facial displays, *Proceedings of the European meeting of the Association of Computational Linguistics (EACL)*, Trento.
- Gardent, C. (2001), Generating minimal definite descriptions, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, pp. 96-103
- Garey, M. and D. Johnson (1979), *Computers and Intractability*, New York: Freeman.

-
- Gatt, A. (2006), Generating collective spatial references, *Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci-06)*.
- Gatt, A., van der Sluis, I., and van Deemter, K. (2007), Evaluating algorithms for the generation of referring expressions using a balanced corpus, in: *Proceedings of the 11th European Workshop on Natural Language Generation, (ENLG-07)*, Dagstuhl, Germany.
- Gratch, J., J. Rickel, E. André, N. Badler, J. Cassell, and E. Petajan (2002), Creating Interactive Virtual Humans: Some Assembly Required, *IEEE Intelligent Systems*: 54-63
- Grice, H.P. (1975), Logic and Conversation, in: *Syntax and Semantics (vol. 3)*, Cole, P. and J. Morgan (eds.), New York: Academic Press (pp. 41-58).
- Griffin, Z. and K. Bock (2000), What the eyes say about speaking, *Psychological Science*, 11, 274-279.
- Grosz, B., A. Joshi & S. Weinstein (1995), Centering: A Framework for Modeling the Local Coherence of Discourse, *Computational Linguistics*, 21, 203-225.
- Gundel, J., N. Hedberg & R. Zacharski (1993), Cognitive Status and the Form of Referring Expressions in Discourse, *Language*, 69, 274-307.
- Hanna, J. and S. Brennan (2007), Speaker's eye gaze disambiguates referring expressions early during face-to-face conversation, *Journal of Memory and Language*, in press.
- Heeman, P. and G. Hirst (1995), Collaborating on referring expressions, *Computational Linguistics*, 21(3), 351-382.
- Henderson, J. (2003), Human gaze control in real world scene perception, *Trends in Cognitive Science*, 7, 498-504.
- Henderson, J. and F. Ferreira (2004), Scene perception for psycholinguists, in: *The Interface of language, vision, and action: Eye movements and the visual world*, J. Henderson and F. Ferreira (eds.), New York, Psychology Press (pp. 1-58).
- Horacek, H. (2005), Generating referential descriptions under conditions of uncertainty, *Proceedings of the tenth European workshop on Natural Language Generation*, Aberdeen, UK.
- Horton W. & Keyser B. (1996) When do speaker take common ground? *Cognition*, 59, 91-117
- Itti, L. and C. Koch (2000), A saliency-based search mechanism for overt and covert shifts in visual attention, *Vision Research*, 40, 1489-1506.
- Iversen, J. and S. Goldin-Meadow (1998), Why people gesture when they speak? *Nature*, 396, 228-229.
- Jurafsky, D. (2002), Probabilistic modeling in Psycholinguistics: Linguistic Comprehension and Production, in: *Probabilistic Linguistics*, R. Bod, J. Hay and S. Jannedy (eds.), MIT Press.
- Kelleher, J., F. Costello and J. van Genabith (2005), Dynamically structuring, updating and interrelating representations of visual and linguistics discourse context. *Artificial Intelligence*, 167, 62-102.
- Keysar, B., Lin, S., and Barr, D. (2003), Limits on theory of mind usage in adults, *Cognition*, 89, 25-41.
- Kita, S. and A. Ozyurek (2003), What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and gesture. *Journal of Memory and Language*, 48, 16-32.

- Krahmer, E. and M. Theune (2002), Efficient context-sensitive generation of referring expressions, in: *Information Sharing: Givenness and Newness in Language Processing*, K. van Deemter and R. Kibble (eds.), CSLI Publications, Stanford, 223-264.
- Krahmer, E., S. van Erk and A. Verleg (2003), Graph-based Generation of Referring Expressions, *Computational Linguistics* 29(1): 53-72.
- Krahmer, E. and M. Swerts (2004), More about brows: a cross-linguistic analysis-by-synthesis study, in: *From Brows to Trust: Evaluating Embodied Conversational Agents*, C. Pelachaud and Zs. Ruttkay (eds.), Kluwer Academic Publishers, 191-216.
- Krahmer, E. and M. Swerts (2005), How children and adults produce and perceive uncertainty in audiovisual speech, *Language and Speech* 48(1): 29-54.
- Krahmer, E. and M. Swerts (2006), Testing the effect of audiovisual cues to prominence via a reaction-time experiment, *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh, PA, USA.
- Krahmer, E. and M. Swerts (2007), The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception, *Journal of Memory and Language*, in press
- Krauss, R., Y. Chen and P. Chawla (1996), Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In: M. Zanna (ed.), *Advances in Experimental Social Psychology* (pp. 389-450). San Diego: Academic Press.
- Krauss, R. and J. Pardo (2004), Is alignment always the result of automatic priming, *Behavioral and Brain Science*, 27, 203-204.
- Kronfeld, A. (1986). Donnellan's distinction and a computational model of reference. *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics (ACL-86)*, 186-191.
- Levelt, W. (1989), *Speaking: From Intention to Articulation*, Cambridge: MIT Press.
- Levelt, W. (1999), Producing spoken language: a blueprint of the speaker, in Brown, C. and P. Hagoort (eds.), *The Neurocognition of Language*, Oxford University Press, Oxford, (pp. 83-122).
- Levelt, W., A. Roelofs and A. Meyer (1999), A theory of lexical access in speech production, *Behavioral and Brain Sciences*, 22(1):1-38.
- Maes, A., A. Arts and L. Noordman (2004), Reference management in instructive discourse. *Discourse Processes*, 37(2), 117-144.
- Meyer, A. Sleiderink, A. & Levelt, W. (1998), Viewing and naming objects: eye movements during noun phrase production, *Cognition*, 66, B25-B33.
- Metzing, C. and S. Brennan (2003), When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions, *Journal of Memory and Language*, 49, 201-213.
- McNeill, D. (1992), *Hand and mind: What gestures reveal about thought*. Chicago: Chicago University Press.
- McNeill, D. and S. Duncan (2000), Growth points in thinking-for-speaking, in: D. McNeill (Ed.), *Language and Gesture*, Cambridge: Cambridge University Press.
- Munhall, K., J. Jones, D. Callan, T. Kuratate and E. Vatikiotis-Bateson (2004), Visual prosody and speech intelligibility, *Psychological Science* 15, 133-137.
- Nakano, Y., G. Reinstein, T. Stocky, and J. Cassell (2003), Towards a Model of Face-to-Face

-
- Grounding, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. July 7-12, Sapporo, Japan.
- Oliva, A., J. Wolfe and H. Arsenio (2004), Panoramic search: The interaction of memory and vision in search through a familiar scene, *Journal of Experimental Psychology: Human Perception and Performance*, 30, 1132-1146.
- Olsen, D. (1970), Language and thought: Aspects of a cognitive theory of semantics, *Psychological Review*, 77, 257-273.
- Oviatt, S.L. (2003), *Multimodal interfaces*, in: The Human-Computer Interaction Handbook, J. Jacko and A. Sears (Eds), Lawrence Erlbaum Assoc., Mahwah, NJ, 286-304.
- Paraboni, I., K. van Deemter and J. Masthoff (2007), Generating Referring Expressions: Making Referents Easy to Identify, *Computational Linguistics*, 33, 229-254..
- Pechman, T. (1989), Incremental Speech production and referential overspecification, *Linguistics*, 27, 98-110.
- Pelachaud, C., N. Badler and M. Steedman (1996), Generating Facial Expressions for Speech, *Cognitive Science*, 20, 1-46,
- Pelachaud, C. and I. Poggi (2002), Subtleties of facial expressions in embodied agents, *Journal of Visualization and Computer Animation*, 13, 301-312.
- Pickering, M. and S. Garrod (2004), Towards a mechanistic psychology of dialogue, *Behavioral and Brain Sciences* 27, 169-226.
- Piwek, P. and K. van Deemter (2006), *Constraint-based natural language generation: A survey*, Technical Report N0. 2006/03, Open University, Milton Keynes, UK.
- Prendinger, H. and M. Ishizuka (2004), *Life-like characters: tools, affective functions and applications*. Springer: Berlin.
- Reiter, E. (2000), Pipelines and size constraints, *Computational Linguistics*, 26, 251-259.
- Reiter, E. and R. Dale (2000), *Building Natural-Language Generation Systems*. Cambridge University Press.
- de Ruiter, J.P. (2000), The production of gesture and speech. In: D. McNeill (ed.), *Language and Gesture* (pp. 284-311). Cambridge: Cambridge University Press.
- Ruttkay, Zs. and C. Pelachaud (2004), *From brows to trust: Evaluating Embodied Conversational Agents*, Kluwer Academic Publishers.
- Sato, W. and S. Yoshikawa (2007), Spontaneous facial mimicry in response to dynamic facial expressions, *Cognition*, 104 (1), 1-18.
- Siddharthan, A. and A. Copestake (2004), Generating Referring Expressions in Open Domains, in *Proceedings of the 42th Meeting of the Association for Computational Linguistics Annual Conference (ACL 2004)*, Barcelona, Spain.
- Spivey, M., M. Tyler, K. Eberhard and M. Tanenhaus (2001), Linguistically mediated visual search, *Psychological Science*, 12 (4), 282-286
- Srinivasan, R. and D. Massaro (2003), Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English, *Language and Speech*, 46, 1-22.

- Stone, M. C. Doran, B. Webber, T. Bleam, and M. Palmer (2003), Microplanning with communicative intentions: the SPUD system, *Computational Intelligence*, 19, 311-381.
- Swerts, M. and E. Krahmer (2005), Audiovisual prosody and feeling of knowing, *Journal of Memory and Language*, 53(1), 81-94.
- Swerts, M. and E. Krahmer (2007), Facial expressions and prosodic prominence: Comparing modalities and facial areas, *Journal of Phonetics*, in press.
- Terken, J. and S. Nootboom (1987), Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information, *Language and Cognitive Processes*, 2, 145-163.
- Tsang, E. (1993), *Foundations of Constraint Satisfaction*, Academic Press, London.
- Van der Sluis, I. and Krahmer, E. (2004), Evaluating Multimodal NLG using Production Experiments. *Proceedings of the 4th International Conference on Language Resources and Evaluation*, May 26-28, Lisbon, Portugal, p. 209-212.
- Van der Sluis, I. and Krahmer, E. (2007), Generating Multimodal References, *Discourse Processes*, in press.
- Viethen, J. and R. Dale (2006), Algorithms for Generating Referring Expressions: Do They Do What People Do? *Proceedings of the Fourth International Natural Language Generation Conference*, Sydney, Australia, 63-70.
- Viethen, J. and R. Dale (2007), Capturing Acceptable Variation in Distinguishing Descriptions, *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG 2007)*, Schloss Dagstuhl, Germany.
- Wooding, D., Muggelstone, M. Purdy, K. and Gale, A. (2002), Eye movements of large populations, *Behavior Research Methods, Instruments and Computers*, 34, 509-517.