

# Natural Language Generation for Embodied Agents<sup>1</sup>

**Emiel Krahmer**

Communication and Cognition  
Faculty of Arts  
Tilburg University  
P.O. Box 90153  
NL-5000 LE Tilburg  
The Netherlands  
e.j.krahmer@uvt.nl  
fdlwww.uvt.nl/~krahmer

**Paul Piwek**

Department of Computing  
Faculty Mathematics & Computing  
The Open University  
Walton Hall  
MK7 6AA, Milton Keynes  
United Kingdom  
p.piwek@open.ac.uk  
mcs.open.ac.uk/

## Abstract

Embodied Agents (Virtual Characters, Talking Heads, Humanoids, Animated Agents, Lifelike Computer Characters, ...) have in the past 10 years become a new alternative paradigm in user interfaces for a wide variety of application domains including training, education, presentation, sales and entertainment. The appeal of embodied agents derives from the fact that they aim to communicate with human users in the mode of interaction which is arguably most natural for humans: face-to-face conversation. In face-to-face conversation, we, humans, almost effortlessly produce complex messages, simultaneously using various channels of communication: we speak, make gestures, change our posture, use facial expressions, etc. This tutorial is an introduction into research on algorithms, evaluation strategies and applications for embodied agents. We take a widely used reference architecture for text-based Natural Language Generation as our starting point and discuss a number of extensions and modifications for the generation of multimodal behaviours for embodied agents.

## 1 Introduction: NLG for Embodied Agents

[...] face-to-face conversation is the basic setting for language. It is universal, requires no special training, and is essential in acquiring one's first language. Other settings lack the immediacy,

---

<sup>1</sup>These notes provide an extended abstract for a tutorial delivered at The seventh Agent Systems Summer School (EASSS 2005), held in Utrecht, The Netherlands on July 18, 2005. The tutorial consists of eight chunks (each approximately 30 minutes long) addressing the eight themes described below, interleaving general introductions into aspects of natural language generation for embodied agents with specific case studies.

medium, or control of face-to-face conversation, so they require special techniques or practices. [...] (Cited from Clark [7], page 11)

The ability to communicate is an essential ability of any software agent. Software agents need this ability to exchange information with other software agents and/or human users. In this tutorial, we provide an overview of theories and technologies which allow agents to communicate information naturally to human users.

The paradigm of natural human-human interaction is face-to-face dialogue. In face-to-face conversation, humans almost effortlessly produce complex messages, simultaneously using various channels of communication. They speak, make gestures, change their posture, facial expression, etc. In order for an agent to engage in such a conversation, it needs to be **embodied** and should be able to both **interpret** and **generate** communicative acts exploiting this wide range of means.

Both interpretation and generation of language and other modalities are research fields in their own right. In this tutorial, we will focus on the latter as it has the largest impact on the appearance and usefulness of embodied agents as information presentation applications.<sup>2</sup> Generation for embodied agents has seen a rise in interest in the past five to ten years, partly due to the now widely available technologies for computer-animated rendering of agents and text-to-speech synthesis (TTS).

A variety of **applications** for embodied agents exists in a wide range of domains including:

- Training and education/tutoring,
- Entertainment (including computer games),
- Product presentations,
- Customer relations management, and
- Sales.

A number of **companies**, such as PlanB, Cantoche, NoDNA, Charamel, Oddcast and Zoesis, and **projects** such as MagiCster, EPOCH, MRE and NECA (see Figure 1) have been working on applications.

---

<sup>2</sup>The natural language generation community has its own biannual (alternating) international conferences and European workshops. These are held under the auspices of the Special Interest Group on Generation of the Association for Computational Linguistics, see [www.aclweb.org/siggen](http://www.aclweb.org/siggen).



Figure 1: Screen captures of the NECA Socialite and eShowroom Demonstrators

There is an ongoing debate whether the use of embodied agents has a beneficial effect on human-computer interaction. Traditionally, the notion of **antropomorphic** interfaces met with some resistance in the user interface community, for instance, because of the idea that a human-like appearance/behaviour might give rise to false expectations (e.g., Norman [23]; Shneiderman [31]). But more recently people have argued that, for instance, computer assistants need to acknowledge the social and emotional aspects of interaction (e.g., Ball et al. [3]), and that in certain situations embodied agents may indeed lead to a noticeable improvement of the interaction (this is referred to as **the persona effect**, Dehn & Van Mulken [11]). However, this persona effect can only arise when the embodied agent has an added value for a given application and if the agent provides output that is of high quality, both verbally and non-verbally. The remainder of this tutorial offers insights in how this high output quality may be achieved.

## 2 What is Natural Language Generation?

When an embodied agent only has to communicate a few utterances to the user, simple string manipulation techniques may suffice (i.e., the agent only outputs predefined messages). Naturally, this approach is limited and highly inflexible. Once the spoken utterances should be more complex and varied, predefined strings will no longer suffice and “real” **natural language generation** (NLG) is required.

Natural language generation may be contrasted with natural language understanding; according to McDonald [22], natural language understanding

can be characterized as **hypothesis management** (given an utterance, find the most plausible interpretation for it given the current context), whereas natural language generation can be characterized as **making choices** (given a particular communicative goal, choose the means to achieve that goal).

Various **architectures** for NLG systems have been proposed, of which the one by Reiter & Dale [28] is generally considered to be the reference architecture, and is adopted by many existing NLG systems (see e.g., Evans, Piwek and Cahill [12]). Recently, standardized **data structures** for NLG systems have been proposed.<sup>3</sup>

The NLG reference architecture consists of a pipeline with three processing stages.

**Document Planner** At this stage, the content and structure of the communicative act is determined. Note that in the reference architecture (Reiter & Dale [28]) there is talk about “the document” rather than the communicative act. This indicates that the model is rather biased towards conventional written text. This bias needs to be removed for embodied agents applications.

**Microplanner** Concerns the selection of words, grammatical structures, etc. It has been argued that this can be done in a shallow or in a deep way, but this distinction is controversial (van Deemter et al [34]).

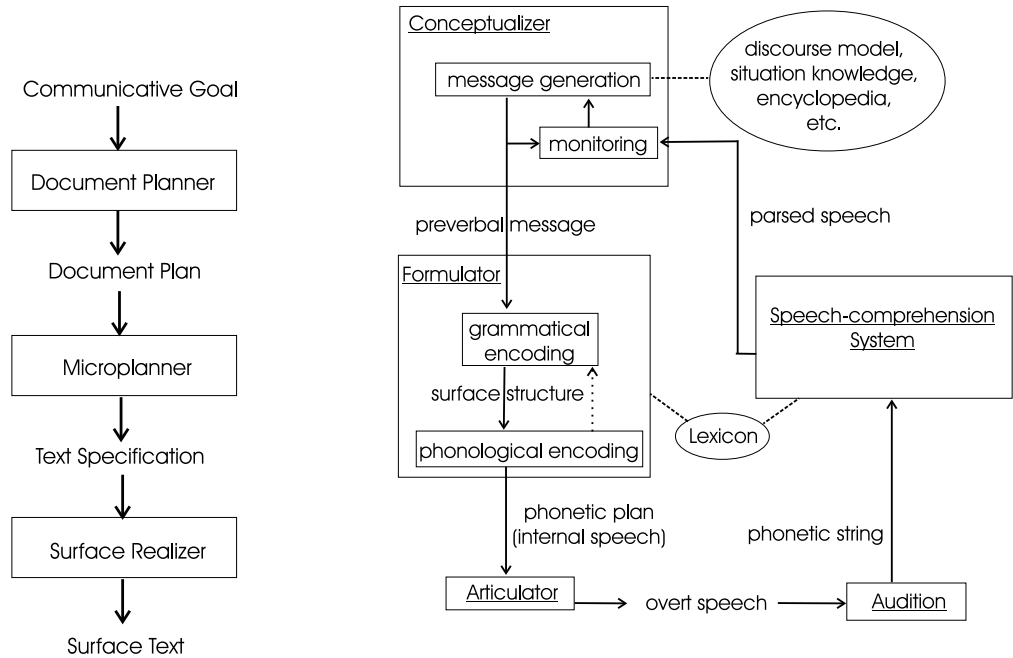
**Surface Realizer** In this stage, the abstract representations of the content and form of the document which have been produced by the preceding modules are mapped to concrete text.

There are a number of **alternatives** for the reference architecture which may be useful when considering NLG for embodied agents (such as integrated planning, revision-based and multi-solution generate-and-test pipelines).<sup>4</sup> It is also worth observing that there are some interesting parallels (and differences) between the reference architecture –which was proposed from an engineering point of view– and **psycholinguistic models** of human language production (e.g., Levelt [21]; see Figure 2). Both models consist of three similar modules, where the Conceptualiser in Levelt’s model may be likened to the Document Planner in Reiter and Dale’s architecture, and the Formulator and Articulator are similar to the Microplanner and Surface Realiser respectively.

---

<sup>3</sup>See, e.g., <http://www.itri.brighton.ac.uk/projects/rags/>.

<sup>4</sup>Currently, statistical approaches to NLG are gaining ground. Such generate-and-test approaches do fit into the pipeline architecture and this allows for the integration of statistical and rule-based components (e.g., Knight & Hatzivassiloglou [15]).



NLG Architecture (from Reiter & Dale, 2000:60)

Blueprint for the speaker (From Levelt, 1989:9)

Figure 2: An engineering and a psycholinguistically motivated architecture for NLG

### 3 Case: Generation of Referring Expressions (GRE)

To discuss the various stages of the NLG reference architecture in more detail we focus on one specific NLG task, namely the generation of **referring expressions**. Generation of referring expressions (GRE) is the task of producing speech acts for identifying objects, e.g., the utterance of a definite description such as “the man with the funny moustache” to identify a particular male individual. GRE is one of the central tasks of Natural Language Generation, and is addressed in nearly every NLG system (albeit in various degrees of detail). Some have even argued that many tasks in sentence planning beyond noun phrase generation can be viewed as a variety of GRE, e.g., tense can be seen as reference to time (*cf.* Stone & Doran [32]).

Basically, GRE algorithms take as input a single object  $v$  (**the target object**) for which a referring expression is to be generated, and a set of objects (**the distractors**) from which the target needs to be distinguished.

The task of a GRE algorithm is to decide which set of properties is needed to single out the target (this is known as the **content determination** problem). On the basis of this set of properties a **distinguishing description** can be generated. Arguably, the Incremental Algorithm (Dale & Reiter [8]) is the state-of-the-art in this field. This algorithm basically tries properties in a predetermined (empirically motivated) order. For each property it encounters, the algorithm determines whether adding this property to the properties selected so far rules out any of the remaining distractors. If so, it is included in the referring expression under construction. This process is repeated until either all distractors are ruled out (success) or all properties have been tried (failure).

The Incremental Algorithm as proposed by Dale & Reiter is limited in that it only applies to singular objects, without taking context into account, and does not offer a principled approach to vague properties (such as “large” or “small”) or relations (such as “next to” or “belonging to”). In the last years, various researchers have proposed extensions and/or variations of the Incremental Algorithm removing one or more of the aforementioned limitations.<sup>5</sup>

Recently a generalized version of the Incremental Algorithm has been proposed, which represents domain information about targets and distractors as a labelled graph and reformulates the task of referring expression generation as a graph construction problem (Krahmer et al. [18]). One advantage of this approach is that it makes it easier to extend the coverage of the Incremental Algorithm (e.g., by incorporating different proposals in a single algorithm). Below, it will be argued that gestures can also be included in this way.

## 4 NLG for Embodied Agents: Architectures

The traditional reference architecture for NLG systems discussed above is not directly applicable to generation for embodied agents. For one thing, it is formulated in terms of “documents”, although this can be remedied relatively easily. But more principled extensions are required: the generated language will need to be converted into audiovisual speech and should be accompanied by facial expressions and gestures (an embodied agent that

---

<sup>5</sup>A good overview of research in GRE can be found on the website for the TUNA project (“Towards a unified algorithm for the generation of referring expressions”), which includes an annotated bibliography. See: <http://www.itri.brighton.ac.uk/projects/tuna/TUNA-index.html>.

only speaks but is otherwise static is perceived as very unnatural). Various extended architectures have been proposed (including those of de Carolis et al. [9], Cassell et al. [4] and Cassell et al. [6]).

These new architectures raise a number of issues, which have not yet been satisfactorily resolved:

- How is information pertaining to the different modalities (speech, gestures, etc.) represented at various stages of processing?
- How is information from different modalities integrated?
- What model drives the choice between different modalities for the realization of content?

It is interesting to observe that in the psycholinguistic community comparable issues are currently being discussed; there is an ongoing debate, for instance, on how Levelt's model for speaking should be extended to include gestures as well. One particular controversy concerns the question *where* gestures originate, with some researchers arguing that the source of gestures is the conceptualiser, while others maintain that the gestures are planned *before* the conceptualiser is activated (see e.g., Krauss et al. [19] and de Ruiter [10]).

## 5 Case: Generating Referring Expressions for Embodied Agents

As an example of how language and gestures can be generated in tandem for use in an embodied agent, we take a closer look at the generation of multimodal referring expression, i.e., linguistic referring expressions with **pointing gestures**. Such expressions are motivated from at least two reasons: (1) in various situations a purely linguistic description may be too complex, for instance because the domain consists of many very similar objects, and (2) in human-human communication such multimodal referring expressions occur very frequently. Various algorithms have been proposed for the generation of multimodal referring expressions, such as André & Rist [1] and Lester et al. [20]. Most of these earlier approaches assume that pointing gesture is precise and unambiguous. As soon as such a gesture is included in the referring expression under construction, this immediately rules out all of the remaining distractors. As a result, the generated descriptions tend to be relatively simple. In addition, the decision to include a gesture is often

based on a straightforward, context-independent criterion (e.g., always add a gesture, or only add a gesture if this is required for unique identification).

An alternative approach, proposed by Krahmer and Van der Sluis [17], is based on the assumption that pointing can be of different **degrees of precision** (ranging from very precise to very imprecise), and thus will not always eliminate all distractors. The decision to point is not hard-coded, but is based on a trade-off between the costs of pointing and the costs of a linguistic expression. Both kinds of costs can be derived empirically: the costs of linguistic properties are based on human preferences (more preferred, i.e., used more often, is cheaper), the costs of gestural properties are derived from **Fitts' law**, a fundamental psychomotoric law indicating the effort for the human motor-system to reach a target (Fitts [13]). The model has been implemented using the graph-based approach to GRE which was mentioned above. Arguably, this approach not only applies to deictic referring expressions (i.e., including pointing gestures), but also to more general combinations of speech and gesture, where, for instance, a circular movement may indicate that the target object is round.

## 6 Realization: Audiovisual Speech and Gestures

An important difference between NLG for embodied agents and NLG for textual documents is that in the former case the generated language will need to be realized via **(audiovisual) speech**: the agent will utter the generated speech (typically via TTS) and this has consequences for the visual appearance as well (e.g., lip movements). This calls for two extensions. First, the generation algorithm will have to determine which words should be **emphasized** and where **pauses** should be placed. Various approaches to this problem of **prosody prediction** exist, but it turns out to be beneficial to determine the placement of pitch accents and pauses during the actual generation process, since much of the required information (about syntax and context) is available there (see e.g., Theune et al. [33]). Second, it needs to be decided how the emphasis and pauses should be realised. Typically, this can be done both auditory (in the speech signal) and visually (using facial expressions and gestures of the embodied agent), and the result is sometimes referred to as **audiovisual prosody**.<sup>6</sup>

Some of the issues related to NLG architectures for embodied agents resurface here at a lower level. For instance, prominence of a word can be

---

<sup>6</sup>For more information about this field, see e.g., the web-page for the "Functions of Audiovisual Prosody" project: <http://foap.uvt.nl/>.



indicated in the speech (typically with a pitch accent), but also visually (often with an eyebrow movement, but other cues such as head nods are used as well). This raises various empirical questions, such as when should a word be emphasized with a pitch accent and when with, say, an eyebrow movement? It has been shown that when *every* word that receives a pitch accent also carries an eyebrow movement (as has been suggested), this leads to rather neurotic looking embodied agents. In general, there is no consensus among developers of embodied agents about these issues, which is partly explainable from the fact that we still do not sufficiently understand when, how and why *human* speakers signal prominence visually. This also raises the question how people *process* visual cues to prominence. Are they just as important as auditory cues or not? And are they used in a similar fashion in different languages. Krahmer & Swerts [16]) argue that for answering these questions experimental evaluation is essential, both with real humans (**analysis-by-observation**) and with virtual ones (**analysis-by-synthesis**).

## 7 Case: NLG for Teams of Embodied Agents

For some purposes, it is useful to present information through a **team** of embodied agents. Such a team might, for instance, *perform* a dialogue which entertains or informs the audience. We are all familiar with such dialogues from TV commercials, plays in the theatre, movies, etc. The automated generation of such presentations was pioneered by André et al. [2], but see also Cassell et al. [4] and Hayes-Roth et al. [14]. Whereas André et al. [2] relied mainly on the use of canned text, more recently, in the NECA project<sup>7</sup> NLG technologies were introduced to increase the believability and variation of the performances (see Piwek [24] and Piwek et al. [25]).

Generation for teams of agents has its own issues; one of the problems is, for instance, the coordination of the behaviours of the agents in the team. One approach is to use a central script authoring agent which controls the actions of the performing agents. This approach has interesting possibilities for controlling global properties of such presentations (e.g., length and style) as discussed in Piwek & Van Deemter [26]. If we start from a *global constraint* on generated discourse, it is possible to shed light on a range of approaches

---

<sup>7</sup>NECA stands for Net-environment for Embodied Emotional Conversational Agents. The project ran from 2001 until 2004 and involved ITRI at the University of Brighton, DFKI, the Institute of Phonetics at the University of the Saarland, OEFAI and Sysis Interactive Simulations AG.

to constraint satisfaction that can be found in existing natural language generation systems. Using these constraints, we can make explicit the – often implicit – choices which the designers of such systems make. A system for the generation of dialogues for teams of agents is used to illustrate the choices that are involved in developing a generation system that makes use of global and soft constraints on its output. The approach we advocate puts well-known techniques from decision theory and game theory to a novel use, thereby exemplifying a recent trend in formal and computational linguistics (Rubinstein [29]).

## 8 Evaluating NLG for Embodied Conversational Agents

Evaluation of NLG for embodied agents is (or should be) a *sine qua non*: it is important to test whether the algorithms indeed achieve what was intended. In the case of embodied agents, we can build in part on experience with the evaluation of user interfaces or NLG applications, but it raises a number of specific problems as well (see e.g., Cassell et al. [5], Ruttkay & Pelachaud, [30], Prendinger & Ishizuka [27] and Dehn & Van Mulken [11]). Various evaluation methods have been applied, such as:

- Observation
- Experiment
- Benchmark and comparative tests
- Usage data
- Questionnaires/interviews
- Biomedical data

These methods may be used to either test how effective an embodied agent is for a particular task or to make claims about the general benefits of embodied agents. Dehn & Van Mulken [11] offer an interesting meta-analysis of evaluation studies involving embodied agents, which provides insights in both the pitfalls in evaluating embodied agents and in the general benefits from adding embodied agents to an application.



## References

- [1] E. Andre and T. Rist. Coping with temporal constraints in multimedia presentation planning. In *Proceedings of the 13th National Conference on Artificial Intelligence*, pages 142–147, 1996.
- [2] E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. The automated design of believable dialogues for animated presentation teams. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, editors, *Embodied Conversational Agents*, pages 220–255. MIT Press, Cambridge, Massachusetts, 2000.
- [3] G. Ball, D. Ling, D. Kurlander, J. Miller, D. Pugh, T. Skelly, A. Stankosky, D. Thiel, M. van Dantzich, and T. Wax. Lifelike computer characters: The persona project at microsoft. In J. Bradshaw, editor, *Software Agents*, pages 191–222. AAAI Press/The MIT Press, 1997.
- [4] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer Graphics and Interactive Techniques*, pages 413–420, 1994.
- [5] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. MIT Press, Cambridge, Massachusetts, 2000.
- [6] J. Cassell, H. Vilhjalmsen, and T. Bickmore. BEAT: the behavior expression animation toolkit. In *Proceedings of SIGGRAPH'01*, Los Angeles, CA, 2001.
- [7] H. Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.
- [8] R. Dale and E. Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263, 1995.
- [9] B. de Carolis, V. Carofiglio, and C. Pelachaud. From discourse plans to believable behavior generation. In *Proceedings of the Second International Conference on Natural Language Generation (INLG02)*, New York, USA, 1–3 July 2002 2002.

- [10] J.P.A de Ruiter. *Gesture and Speech Production*. PhD thesis, Max Planck Institute, Nijmegen, 1998.
- [11] D. Dehn and S. van Mulken. The impact of animated interface agents: a review of empirical research. *Int. J. Human-Computer Studies*, 52:1–22, 2000.
- [12] R. Evans, P. Piwek, and L. Cahill. What is NLG? In *Proceedings of International Natural Language Generation Conference (INLG02)*, New York, USA, 2002.
- [13] P. Fitts. The information capacity of the human motor system in controlling amplitude of movement. *Journal of Experimental Psychology*, 47:381–391, 1954.
- [14] B. Hayes-Roth, R. van Gent, and D. Huber. Acting in character. In R. Trappl and P. Petta, editors, *Creating personalities for synthetic actors*, pages 92–112. Springer, Berlin, 1997.
- [15] K. Knight and V. Hatzivassiloglou. Two-level, many-paths generation. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL-95)*, Boston, MA, 1995.
- [16] E. Krahmer and M. Swerts. More about brows. In Zs. Rutkay and C. Pelachaud, editors, *From brows to trust: Evaluating Embodied Conversational Agents*. Kluwer Academic Publishers, 2004.
- [17] E. Krahmer and I. van der Sluis. A new model for the generation of multimodal referring expressions. In *Proceedings European Workshop on Natural Language Generation (ENLG2003)*, Budapest, Hungary, 2003.
- [18] E. Krahmer, S. van Erk, and A. Verleg. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72, 2003.
- [19] R. Krauss, Y. Chen, and P. Chawla. Nonverbal behavior and non-verbal communication: What do conversational hand gestures tell us? In M. Zanna, editor, *Advances in experimental social psychology*. Academic Press, Tampa, 1996.
- [20] J. Lester, J. Voerman, S. Towns, and C. Callaway. Deictic believability: Coordinating gesture, locomotion and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*, 13(4-5):383–414, 1999.

- [21] Willem J.M. Levelt. *Speaking: From Intention to Articulation*. The MIT Press, Cambridge, Massachusetts, 1989.
- [22] D. McDonald. Natural language generation. In *Encyclopedia of Artificial Intelligence*, pages 983–997. John Wiley and Sons, 1992.
- [23] D. Norman. How might people interact with agents. In J. Bradshaw, editor, *Software Agents*, pages 49–55. AAAI Press/The MIT Press, 1997.
- [24] P. Piwek. A flexible pragmatics-driven language generator for animated agents. In *Proceedings of EACL (Research Notes)*, Budapest, Hungary, 2003.
- [25] P. Piwek, R. Power, D. Scott, and K. van Deemter. Generating multimedia presentations: From plain text to screen play. In O. Stock and M. Zancanaro, editors, *Multimodal Intelligent Information Presentation*, volume 27 of *Text, Speech and Language Technology*. Springer, 2005.
- [26] P. Piwek and K. van Deemter. Dialogue as discourse: Controlling global properties of scripted dialogue. In Reva Freedman and Charles Callaway, editors, *Natural Language Generation in Spoken and Written Dialogue: Papers from the 2003 AAAI Spring Symposium*, pages 118–124. AAAI Press, 2003.
- [27] H. Prendinger and M. Ishizuka, editors. *Life-Like Characters: Tools, Affective Functions, and Applications*. Cognitive Technologies Series. Springer, Berlin, 2004.
- [28] E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, 2000.
- [29] A. Rubinstein. *Economics and Language: Five Essays*. The Churchill Lectures in Economic Theory. Cambridge University Press, 2000.
- [30] Zs. Ruttkay and C. Pelachaud, editors. *From Brows to Trust: Evaluating Embodied Conversational Agents*. Kluwer Academic Publishers, Dordrecht, 2004.
- [31] B. Shneiderman. Direct manipulation versus agents. In J. Bradshaw, editor, *Software Agents*, pages 97–106. AAAI Press/The MIT Press, 1997.

- [32] M. Stone and C. Doran. Sentence planning as description using tree-  
adjoining grammars. In *Proceedings of the Annual Conference of the  
Association for Computational Linguistics (ACL)*, Madrid, Spain, 1997.
- [33] Mariet Theune, Esther Klabbers, Jan Odijk, Jan Roelof de Pijper, and  
Emiel Krahmer. From data to speech: a general approach. *Natural  
Language Engineering*, 7(1):47–86, 2001.
- [34] K. van Deemter, E. Krahmer, and M. Theune. Real vs. template-  
based natural language generation: A false opposition? *Computational  
Linguistics*, 31(1), 2005.