Generation of multimodal dialogue for net environments

Brigitte Krenn, Hannes Pirker
Austrian Research Institute for Artificial Intelligence ÖFAI
Vienna, Austria
{brigitte,hannes}@ai.univie.ac.at

Martine Grice, StefanBaumann
Department of Phonetics, University of the Saarland
Saarbrücken, Germany
{mgrice,baumann}@coli.uni-sb.de

Paul Piwek, Kees van Deemter
Information Technology Research Institute ITRI, University of Brighton,
Brighton, United Kingdom
{paul.piwek,kees.van.deemter}@itri.brightion.ac.uk

Marc Schröder, Martin Klesen
German Research Center for Artificial Intelligence DFKI,
Saarbrücken, Germany
{schroed,klesen}@dfki.de

Erich Gstrein
Sysis interactive simulations ag,
Vienna, Austria
eg@sysis.at

Abstract

In this paper an architecture and special purpose markup language for simulated affective face-to-face communication is presented. In systems based on this architecture, users will be able to watch embodied conversational agents interact with each other in virtual locations on the internet. The markup language, or Rich Representation Language (RRL), has been designed to provide an integrated representation of speech, gesture, posture and facial animation.

1 Introduction

In this paper, we address aspects of the design and implementation of so called *net environments*, i.e., virtual environments in the internet where embodied conversational characters act on behalf of their users. The focus in our net environments is on the situatedness of the agents in a particular application domain, and on the simulation of affective face-to-face life-like interaction between agents[1] watched by the user.

A *net environment* is defined as a multi-user application for the internet

1. where the users are represented by avatars which engage in social behaviour in a virtual location. This behaviour may have a practical purpose, like seeking information about a product, or a social purpose, like making friends;
2. where the user is able to design her/his avatar with respect to its graphical representation, its personality traits and emotional disposition, as well as its interests. The amount of freedom the user has for defining her/his avatar is application-specific;
3. where the agents are autonomous once they have been created by the user; the user may have a degree of influence over the agents by giving them advice. However, the agents may or may not take this advice, depending on their personality and mood, or on parameters set within the application.

Our characters are Embodied Conversational Agents (ECAs) for the internet. According to Cassell et al. (2000), ECAs are computer-

[1] In the following we refer to agents as avatars or agents: avatars because they are comparable to templates filled by the user; agents because (i) net environments are also inhabited by agents which are fully system defined, and (ii) each avatar, after creation, has its autonomous existence in the net environment.

generated cartoon-like characters presenting life-like properties in face-to-face conversation. They need to be able to manage turn-taking, giving appropriate cues when holding or relinquishing the floor, and to signal discourse-related aspects such as whether information is given or new and whether a new discourse topic is being introduced. They need to be able to express personality and emotion through verbal and nonverbal communication channels, using high quality speech synthesis and animation of facial expressions, posture and gesture. By engaging in face-to-face interaction, the characters simulate making contact and establishing social relationships with other agents (and possibly with the user who is engaging in social interaction by proxy). The need to integrate acoustic and visual aspects of communication has led to the definition of a Rich Representation Language. See section 4 where the functionality of the RRL is described with respect to the information required/available at the individual interfaces between the components of the core system. In section 2 concrete application scenarios for our architecture are presented. In section 3 an overview of the architecture is given. The work described in this paper is part of the NECA project.[2]

2 Example Applications

In order to provide a clearer picture of the sort of applications our architecture aims for, we will describe the two application scenarios which are to be implemented as demonstrators within the NECA project.

2.1 eShowRoom

In the eShowRoom scenario a car sales dialogue between a seller and one or more buyers is simulated. The purpose of this demo is (basically) to entertain the site visitor and to embed product information into a narrative context similar to TV commercials nowadays. User interaction is restricted to setting general parameters prior to the display. These parameters include the user's preferences in respect to different value dimensions, e.g. on how important aspects like sportiness, prestige or environmental issues are for the user. After having specified these preferences, a scene is generated which takes these settings into account: The agents/interlocutors will put special emphasis on conveying information about those aspects, which have been classified as being of importance for the user. The visitor can also specify the personality traits of the agents, i.e., their agreeableness and politeness, to influence the style and the course of their conversation. This option aims to provide a means of entertaining the user by experimenting with different (possibly absurd) settings.

2.2 Socialite

The Socialite demonstrator implements a multi user web-application in the social domain. The users create their personal avatar, endow it with personality traits and preferences and send it to the virtual environment in order to meet other avatars. The overall goal is to be accepted in the community, to reach a certain degree of popularity within this environment.
In this setting the user is not permanently logged on. The avatar/agent will report back to the user about encounters with other avatars when the user logs in the next time. This report is presented in the form of monologues, which are alternated with displays of dialogues between avatars, much in the style of the rendering of retrospectives in older movies. The user is then queried for choosing new instructions for the avatar from a given set of possibilities and sends the avatar off to its environment again.

3 The NECA Architecture

The NECA architecture consists of the following main components: a scene generator, a multimodal natural language generator, a text/concept-to-speech synthesis, and a gesture assignment module (see Figure 1).
In the preparation phase of an application the user is able to provide the system with information on her information needs and preferences ("User Input") in order to adapt the generated scenes accordingly. For instance, the user is allowed to select the dialogue topic, or to define the personality traits of the agents/interlocutors.
In our application scenarios the *scene generator* takes the role of a playwright, generating a script

[2] IST-2000-28580, http://www.ai.univie.ac.at/NECA/

for the characters that become actors in a scene[3]. In the script, dialogue and presentation acts to be carried out are specified as well as their temporal coordination. The parts of the dialogue are not generated in an incremental way, as would be the case in a conventional interactive system, but an entire scene is generated taking into account the user's input provided beforehand. The information given by the user is also included in the scene description, which specifies the semantic content, type, temporal order, and associated emotion of the communicative acts that the characters will perform. All this information is encoded in an abstract representation scheme which is part of our Rich Representation Language as described in the next section.

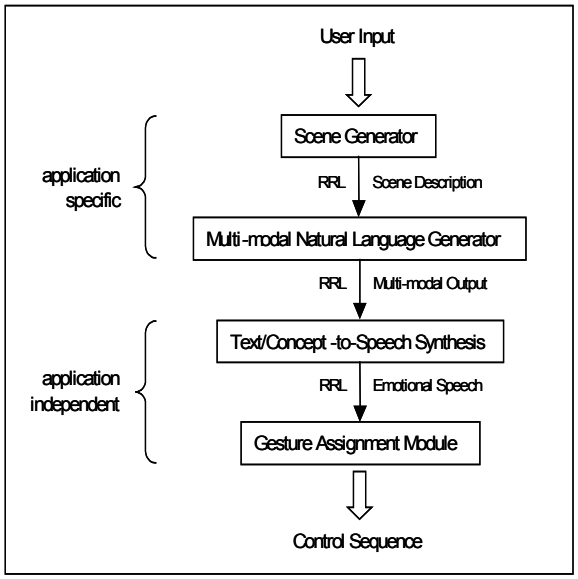

Figure 1 Overview of the NECA Architecture

The scene description is then handed over to the *multimodal natural language generator*, which transforms the formal specification of the communicative acts into text, annotated with syntactic, semantic, and pragmatic features. The component is also (partially) responsible for nonverbal behaviour, such as selecting gestures. The *multimodal output* is a list of communicative acts with a fully specified temporal ordering. The task of the *text/concept-to-speech synthesis* is then to convey, through adequate voice quality and prosody, the intended meaning of the text as well as the emotion with which it is uttered. It also provides information on the exact timing of utterances, syllables and phonemes, which is indispensable for the *gesture assignment module*. This module is responsible for fine-tuning the synchronisation of speech and gesture, including proper alignment of verbal and nonverbal output. It also schedules physiologically based animations (e.g. eye blinking and breathing) in accordance with the constraints imposed by the content-bearing gestures, so as to make the characters more life-like. The output of this process is a *control sequence* comprising the synchronised verbal and nonverbal behaviour of all the characters in the scene. In a last step this control sequence is converted into a data stream that can be processed by a specific player, e.g. Macromedia Flash[4] or Microsoft Agent[5]. In Figure 1 the scene generator and the multimodal natural language generator are characterised as *application specific* whereas the text/concept-to-speech synthesis and the gesture assignment modules are *application independent*. This clearly shows the amount of adaptation that would be required in order to construct a specific application scenario.

In the remainder of this section we will compare NECA with other systems that have been used to create embodied conversational agents. As mentioned before, NECA focuses on communication between animated agents to be observed by the user. Thus during script execution, there is no interaction with the characters, i.e., the user cannot engage in a face-to-face conversation with the agents, as in the REA system (Cassell et al. (1999)) or the animated pedagogical agent, STEVE (Rickel & Johnson (1998)). As we are not concerned with multimodal input, there is no need to recognise and respond to verbal and nonverbal input, nor to deal with conversational functions such as turn taking, feedback and re-

[3] "Scene I Theatr. 1 A subdivision of (an act of) a play, in which the time is continuous and the setting fixed, marked in classic drama by the entrance or departure of one or more actors and in non-classic drama often by a change of setting; the action and dialogue comprised in any one of these subdivisions." (source: Electronic New Shorter Oxford English Dictionary, 1996)

[4] http://www.macromedia.com/software/flash/

[5] http://www.microsoft.com/msagent/

pair mechanisms towards the user. These discourse functions, however, are used during the interaction between the characters in a scene by *simulated* dialogue, thus back channeling (e.g. a nod or an interjection like "uh") and clarification dialogue parts are *generated* by the system in a similar way to the other parts of the dialogue.

Application scenarios that can be developed with the NECA platform include the automated generation of believable dialogues comparable to the work described by André et al. (2000). In their plan-based approach a dialogue script is generated based on a set of text templates and character-specific animation sequences. In our system there is a clear distinction between the generation of semantic content and the surface realisation of an utterance. This allows us to convey the same semantic content in a way specific to individual personalities and cultures. Our system aims at producing output similar to the Behavior Expression Animation Toolkit (BEAT) by Cassell et al. (2001). The BEAT system allows animators to input typed text that they wish to be spoken by an animated human figure, and to obtain as output appropriate and synchronised nonverbal behaviour and synthesised speech in a form that can be sent to a number of different animation systems. The nonverbal behaviour is assigned on the basis of actual linguistic and contextual analysis of the typed text, relying on rules derived from research into human conversational behaviour. In our system however, the gesture assignment module (section 3.4) is not responsible for nonverbal behaviour that is part of the discourse structure. These content-bearing gestures are provided by the multimodal natural language generation component.

4 The RRL

The RRL is an XML compliant special purpose markup language which has been designed for the description of agent behaviour in our net environments. The RRL represents a wide range of expert knowledge required at the interfaces between the different components in the NECA architecture.

Existing markup languages have either been designed for the representation of information at individual levels of description or provide a combination of markups at different levels of representation for multimedia annotation. A good deal of work has been done on the former, especially on speech synthesis markup and facial animation coding, see, e.g. the W3C Speech Synthesis Markup Language and MPEG4 FAPs (Facial Animation Parameters).[6] For a recent survey of facial and gesture coding schemes see the ISLE Report D9.1. (Wegener-Knudsen et al., (2002)). Markup languages for multimedia annotation include VHML, SMIL, MPML and TVML.[7] VHML especially, aims at unifying a confederation of existing special purpose markup languages. The RRL differs from other multimedia markup languages in that these are typically designed to support a fairly text-based annotation of multimodal input to media players, ideally in a rather generalised and standardised way, whereas the RRL is in addition capable of representing expert knowledge which may be created by a processing component rather than a human author (for instance, detailed information on the linguistic structure). In developing the RRL we are able to draw on existing standardisation efforts and build on well-defined cores of XML-based markup languages, especially in the field of speech synthesis and facial animation.

In the following, we describe the RRL on a conceptual level. For a realisation in XML see http://www.ai.univie.ac.at/NECA/RRL/RRL_docs/RRL_Specification-0.2.pdf .

4.1 Scene Descriptions

The Scene Generator takes as its input a database containing facts about the application domain and an assignment of roles and personality characteristics to the agents involved in the scene. Based on this information it produces a high-level specification of the scene: the scene description.

The scene description contains the following information: (1) Information on the *agents*, in particular, their personality and the role they play in the scene. (2) The *common ground* (CG; e.g. Clark (1996)) of the agents at the outset of

[6] http://www.w3.org/TR/speech-synthesis/, http://www.es.com/mpeg4-snhc/index.html

[7] http://www.vhml.org/, http://www.w3org/AudioVideo/, http://www.miv.t.u-tokyo.ac.jp/HomePageEng.html, http://www.strl.nhk.or.jp/TVML/index.html

the scene. It contains primarily information about objects in the application domain that the agents assume to be common. The CG plays, for instance, a role in determining the content and form of referring expressions. (3) The *history* of the scene, that is the set of acts which occur in the scene and their temporal ordering. We distinguish between 'dialogue acts' (any communicative act whether expressed by linguistic or non-linguistic means) and 'non-communicative acts' (e.g. walking/moving from one location to another). Each dialogue act has six attributes: its dialogue act type, its speaker, its set of addressees, its semantic content, the act which it is a reaction to (e.g. an assertion can be a reaction to a preceding question) and finally a set of emotions. We distinguish between the emotions which the speaker feels when performing the act and the ones which s/he expresses in performing the act. Additionally, both the emotion felt by the hearer when the act is performed and the emotions expressed by her or him as a direct result of the dialogue act are part of the dialogue act description.

Scene descriptions are the input to the Multimodal Natural Language Generator (M-NLG). They are exchanged between modules in the form of an XML document. Conceptually, they can, however, be thought of as network representations. Network representations are used for a number of reasons. Most importantly, they allow for the uniform representation of different types of information. For instance, paralinguistic information (e.g. emotion as characterised by Ortony et al.(1988) or Cowie et al. (2001)) and semantic information (we use an encoding of Discourse Representation Structures – Kamp & Reyle (1993)) can be integrated more easily in a network representation. For details, we refer to Piwek (2002).

4.2 Multimodal NLG

In terms of Reiter and Dale's (2000) architecture for NLG systems, the Scene Generator corresponds roughly to their "document planner" which specifies the discourse/dialogue content and structure. Thus the M-NLG covers primarily what is known as "microplanning" and "surface realisation". In NECA, microplanning will consist of content elaboration (e.g. the introduction of indirect speech acts, irony and implicature triggers on the basis of pragmatic considerations such as the role/status of the agents) and content organisation (e.g. aggregation, repetition and dialogue act internal ordering of information[8]). Microplanning also includes the lexicalisation of the organised and elaborated content. Lexicalisation results in the specification of utterances: it delivers an abstract representation of sentences (an abstract syntactic representation containing information on mood, base lexemes, parts of speech, etc.), gestures (see the section on gesture assignment) and information structure.[9]

Referring expression generation is dealt with in a separate submodule. This module determines the content, abstract syntactic form and gestures of referring expressions.[10] A surface realiser takes the output of the microplanner and produces a specification of the scene in the form of an annotated text. For this purpose, the abstract syntactic representations need to be turned into word sequences (linearisation) taking agreement into account. The text is annotated with information on gestures, emotion, information structure, syntactic structure and dialogue structure.

4.3 Speech Synthesis

The task for speech synthesis is to convey, through adequate voice quality and prosody, the intended meaning of the text as well as the emotion with which it is uttered.

Since global syntactic structure as well as semantics and information structure are known to be important factors influencing prosody (Ladd (1996), Steedman (2000)), the speech synthesis system has to be able to make use of such information. However, it also has to be able to

---

[8] The scene generator determines the order in which dialogue acts occur. The content of a dialogue act can, however, contain several pieces of information whose ordering is not determined by the scene generator. To determine this ordering the M-NLG uses criteria for judging the coherence of the resulting ordering (e.g. in terms of centering theory), see Kibble & Power (1999).

[9] E.g., the distinction between new and given information which influences accent placement. See, e.g., Van Deemter (1998).

[10] Our algorithm for the combination of gestures (in particular, pointing acts) and linguistic acts in referring expressions will make use of empirical research, e.g., Piwek & Beun (2001).

handle input in which no linguistic annotation is available (e.g. where orthographic text is entered by hand). This means that a scalable representation language is needed, which can represent little more than plain text in the minimal case, and a full specification of linguistic structure in the maximal case.

In the minimal case, the M-NLG (or a human developer) only provides text with no linguistic annotation to the speech synthesis component, along with information about the speaker of a dialogue act and the emotion with which it is to be spoken. The text-to-speech part of the synthesis component utilises the sparser representation to drive the synthesiser's default text-to-speech rules. Appropriate prosody and voice quality are determined based on the specified emotion dimensions (Schröder et al. (2001)).

In the maximal case, the M-NLG additionally provides detailed linguistic information about the text structure. This includes the part-of-speech category for each word, and optionally a phonetic transcription for irregular words not in the lexicon. The syntactic structure is fully specified, providing the full syntactic tree of phrase nodes and their grammatical functions. In addition, the information structure (in terms of theme and rheme) as well as the informational status of individual referents (in terms of givenness and contrast) is specified.

A challenge for the RRL is the simultaneous specification of syntactic structure, information structure and prosodic structure, since there is the possibility of overlap (Steedman (2000)), corresponding to crossing edges in the respective tree structures Crossing edges are not permitted in XML, which requires a strictly embedded tree structure. This issue is partly resolved by the fact that prosodic structure is encoded using a "flat" version of the tonal annotation system GToBI (Grice et al., to appear) which is already part of NECA's TTS system, MARY (Schröder & Trouvain (2001)). We currently assume that syntax and information structure can still be represented in one tree structure by applying only slight restrictions to the possible configurations for information structure.

4.4 Gesture Assignment

A central goal within NECA is to provide agents with appropriate and correctly synchronised gestures. The term "gesture" is used in a broad sense here, referring to all different forms of non-verbal behaviour, i.e., facial expression, gesture proper and posture. In the NECA architecture the task of Gesture Assignment is distributed over three levels in the information flow during processing.

(1) Within the M-NLG *candidate gestures* are assigned to dialogue acts on the basis of the scene to be generated, semantic content of utterances, turn-taking etc. Within NECA, the M-NLG is responsible for specifying both linguistic and non-linguistic behaviour. In addition to producing richly annotated text which will feed into the speech synthesis component, it determines non-linguistic aspects of characters' behaviour such as iconic and emblematic gestures, a number of facial expressions, and emotive interjections (also referred to as affect bursts).[11] Because information on the exact timing of utterances, syllables and phonemes is still lacking at this stage but is indispensable for synchronisation, the final specification of non-linguistic acts has to be postponed. In our architecture the M-NLG thus has the role of mainly providing candidates, while the selection, final specification and scheduling is performed later on. (See the BEAT-system for a related approach; Cassell et al., (2001)). Preliminary decisions on the selection of candidate gestures comprise non-verbal actions such as *emblematic gesture*s (conventionalised gestures such as yes/no); *iconic gesture*s (mimicking the form of an object or action, such as imitating a telephone receiver by stretching thumb and little finger between ear and mouth); *deictic gestures* (pointing gestures with arm and hand); *contrast gesture* (usually literally a "on the one hand… on the other hand" gesture); *back channeling* (e.g. nodding, frowning but also such gestures combined with interjections such as "aha").

[11] Emotive interjections, such as sigh, yawn, laughing etc. add to the emotional believability of the agents' utterances, and are produced holistically (as opposed to being generated from smaller units in the speech synthesis). Technically speaking this can be handled by including tags specifying interjections within the text to be spoken, as no overlap/parallelism with speech can occur.

In the RRL every gesture is attributed a priority (aiding behaviour selection), intensity, direction, and degree of stretch/size.
(2) After speech synthesis, i.e., when exact timing information is available, the Gesture Assignment Module (GA) proper is invoked. At this stage of processing gestures are selected from the set of candidates and exact timing is specified. In order to facilitate the synchronisation of gestures and speech, the GA module is provided with information derived from the speech-synthesis module, in particular:
Phones: information as to the name and exact temporal position of each speech sound is used both for specifying the visemes for lip-synchronous animation and for calculating all further timing information
Syllable and word-boundaries: for instance eye movements are tightly synchronised with syllables, beats synchronise with emphasised words.
Syllables bearing word stress: stressed syllables are the preferred anchor point for deictic gestures, eyebrow raises and head nods.
Position and type of sentence accents: for instance stroke gestures preferentially coincide with syllables bearing a pitch accent.
Location of prosodic phrase boundaries: prosodic phrase boundaries are landmarks for eyebrow raising, head nods, and eye blinking.
Pause Positions: are used for the timing of posture changes, breathing movements, head nods.
As a final step in the GA module, physiologically based animations, e.g. *physiological eye blinking* and *physiological breathing* are added and scheduled in accordance with the constraints imposed by the content based animations.
(3) gestural specifications have to be transformed from the format used in the RRL to player-specific formats, such as MPEG-4. This task is not addressed here.

5 Conclusion and Outlook

We have presented an architecture and markup language for the design and implementation of conversational characters. The focus of the work is on multimodal output, in particular the simulation of multi-agent communication. It describes the core of a platform of dedicated components for the implementation of net environments, virtual spaces for the internet which are populated by life-like agents. Currently two demonstrators are under implementation, and initial prototypes with restricted functionality are due at the end of 2002. In the first demonstrator, animated presentation teams discuss a product along a set of value dimensions that are important for the user. How the information is presented depends on the personality, the emotional state and the (user specified) preferences of the virtual actors. In the second demonstrator, users define their avatar representatives and send them off into a virtual social environment. This is a game situation where the users help their avatars to be socially successful according to the terms of the particular environment. Currently a core version of the architecture is implemented, and a first version of the RRL is available.

Acknowledgements and Disclaimer

The authors thank Neil Tipper and an anonymous reviewer for providing helpful comments on an earlier version of this paper.
This research is supported by the EC Project NECA IST-2000-28580. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

References

André, E., Rist, T., van Mulken, S., Klesen, M. and Baldes, S. (2000). The Automated Design of Believable Dialogues for Animated Presentation Teams. In: Cassell et al. Embodied Conversational Agents, MIT Press.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K. and Vilhjálmsson, H., Yan, H.(1999). "Embodiment in Conversational Interfaces: Rea", ACM CHI 99 Conference Proceedings, Pittsburgh, PA.

Cassell J., Sullivan J., Prevost S. and Churchill E (2000), (eds) Embodied Conversational Agents. The MIT Press, Cambridge, MA.

Cassell, J., Vilhjálmsson H. and Bickmore T. (2001). BEAT: the Behavior Expression Animation Toolkit. *Computer Graphics* (Proceedings of SIGGRAPH), Los Angeles, CA.

Clark, H. (1996). Using Language. Cambridge University Press, Cambridge.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J.

(2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1):32-80.

Grice, M., Baumann, S. and Benzmüller, R. (to appear). German Intonation in Autosegmental-Metrical Phonology. In: Jun, S.-A. (ed.) *Prosodic Typology*. OUP.

Kamp, H. and Reyle, U. (1993). From Discourse to Logic. Kluwer Academic Publishers, Dordrecht.

Kibble, R. and Power, R. (1999). Using centering theory to plan coherent texts. Proceedings of the 12th Amsterdam Colloquium. Amsterdam.

Ladd, D.R. (1996). *Intonational Phonology*. CUP.

Ortony, A., Clore, G. and Collins, A. (1988). The Structure of Emotions. Cambridge University Press. Cambridge MA.

Piwek, P. and Beun, R.J. (2001). Multimodal referential acts in a dialogue game: from empirical investigation to algorithms. Proceedings of International Workshop on Information Presentation and Natural Multimodal Dialogue (IPNMD-2001), Verona, Italy, 127-131.

Piwek, P. (2002). Specification of Scene Descriptions for the NECA domains. NECA deliverable D3a. http://www.ai.univie.ac.at/NECA/RRL/

Reiter, E. and Dale, R. (2000). Building natural language generation systems. Cambridge University Press.

Rickel, J. and Johnson, W.L. (1998). STEVE: A Pedagogical Agent for Virtual Reality. In *Proceedings of the Second International Conference on Autonomous Agents*, Minneapolis/St. Paul, ACM Press.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M. and Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In: Proceedings of Eurospeech, Aalborg, Denmark. Volume 1, 87-90.

Schröder, M. and Trouvain, J. (2001). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. 4th ISCA Workshop on Speech Synthesis, Blair Atholl, Scotland. http://mary.dfki.de.

Steedman, M. (2000). Information structure and the syntax-phonology interface. Linguistic Inquiry, 31(4): 649-689.

Wegener-Knudsen M., Martin J.-C.and Dybkjaer L. (eds.) (2002). Survey of Multimodal Annotation Schemes and Best Practice. EAGLES/ISLE Natural Interactivity and Multimodality Working Group, http://www.ilc.pi.cnr.it/EAGLES96/isle/nimmwg_doc/ISLE_D9.1.zip

Van Deemter, K. (1998). Towards a Blackboard Model of Accenting. Computer Speech and Language 12 (3)