Abstract

In this paper, we report on an experiment showing how the introduction of prosodic information from detailed syntactic structures into synthetic speech leads to better disambiguation of structurally ambiguous sentences. Using modifier attachment (MA) ambiguities and subject/object fronting (OF) in German as test cases, we show that prosody which is automatically generated from deep syntactic information provided by an HPSG generator can lead to considerable disambiguation effects, and can even override a strong semantics-driven bias. The architecture used in the experiment, consisting of the LKB generator running a large-scale grammar for German, a syntax-prosody interface module, and the speech synthesis system MARY is shown to be a valuable platform for testing hypotheses in intonation studies.

1 Prosody and Generation

The inclusion of prosodic information is standardly believed to play a prominent role for the improvement of CTS and TTS applications, in terms of naturalness and intelligibility, see, e.g., McKeown and Pan (2000) and Olaszy & Nemeth (1997). Another added value of prosody lies with its potential for disambiguation: it is often observed that structural ambiguities found in written texts are absent from speech, which is prosodically structured. In order to assess this potential, we carried out an experiment to establish how and to what extent prosody can contribute to improved comprehension of automatically generated speech. We conjecture that disambiguating prosody will not only lead to better intelligibility, but also enhance overall naturalness, due to an improved correspondence between intended meaning and prosodic realisation.

Current TTS systems for German, such as MARY, typically only make use of shallow linguistic annotations like those provided by chunk parsers to control generation of prosody. Due to the limitations of shallow analysis, these TTS systems typically lack the kind of detailed and rich information that can be provided by deep parsers grounded in linguistic theory. By showing that substantial disambiguation effects can be obtained on the basis of prosody derived from HPSG trees, we believe to have made a case for the inclusion of deep syntactic analysis in TTS.

Finally, research at the syntax-prosody interface often involves construction of test data to verify hypotheses. With the help of HPSG processing connected to speech synthesis via a syntax-prosody module, construction of test stimuli can be greatly facilitated.

[†]We would like to thank Martine Grice, Stefan Baumann and Marc Schröder for fruitful discussion of several aspects of this work. We are also indebted to the audiences at the HPSG 2007 and ICPhS 2007 conferences for comments on and discussion of the ideas presented here, in particular Ivan Sag, Emily Bender, Tobias Kaufmann, and Carlos Gussenhoven. A great many thanks also to the anonymous reviewers for their invaluable comments.

The research reported on in this paper has been partially supported by the DFKI project Checkpoint, funded by the Federal State of Berlin and the EFRE programme of the European Union.

1.1 System architecture

The prosody component we present here is part of a system that implements an entire concept-to-speech (CTS) pipeline from deep semantic input in MRS format, through tactical chart-based HPSG generation, to speech output.

1.1.1 Deep syntactic generation

The tactical syntactic generator used in the experiments consists of a linguistically grounded large-scale HPSG grammar of German (GG; http://gg.dfki.de; Müller and Kasper, 2000; Crysmann, 2003, 2005), running on the LKB system (Copestake, 2001). Both the grammar and processing system are fully reversible, i.e., they are suitable for parsing, as well as generation. Generation with the German grammar GG has recently been evaluated on the Babel test suite (Müller, 2004), a phenomenon-oriented regression test suite for German: currently, 99.6% of all sentences that can be parsed, can also be generated by the grammar.

The LKB chart generator takes as input sentence-semantic representations in the form of Minimal Recursion Semantics (Copestake et al., 2005). Given the reversibility of both grammar and processing system, the current architecture may also be used in a text-to-speech (TTS) scenario by simply running the grammar in parsing mode (see below).

As output the generator produces surface strings, together with two isomorphic tree structures, one containing traditional category labels derived from the underlying AVM representation, the other encoding functional notions, such as head, subject, complement or modifier, corresponding to the composition principles of the grammar.

1.1.2 Syntax-Prosody Interface

The two tree representations provided by the generator are folded into a single XML tree representation, where functional and categorial labels are represented as attributes on the nodes.

The information contained in the syntax trees is transformed into prosodic markup by means of XSLT, crucially using XPATH regular expressions. The prosodic markup generated on the basis of the syntactic representations comprises tonal and phrasing information, represented as GToBI annotations (Grice and Baumann, 2002).

For the phenomena discussed in this paper, information provided by the HPSG rule backbone and the category labels was sufficient for prosody planning. We are fully aware of the fact that this may easily prove insufficient for prosodic phenomena more tightly linked to semantics or information structure. However, given that the entire sentential feature structure is always accessible when processing with HPSG grammars, the necessary information can easily be extracted.

Realisation of the prosody module as a separate component was a design decision, since it supports a clean separation of syntactic and phonological aspects suit-

able for distributed development. In the context of a test platform for hypothesis testing in prosody, the usefulness of such a modular organisation cannot be underestimated, as it ensures that most grammar-internal details are effectively hidden from the phonologist primarily concerned with the syntax-prosody interface.

1.1.3 Phonetic realisation

The prosodically annotated text is submitted to the MARY synthesis system (Schröder and Trouvain, 2003) for phonetic realisation using diphone synthesis. MARY is a highly flexible TTS system, supporting annotation of the input data, ranging from low-level control over physical parameters to high-level phonological specification. For the experiments, we made use of GToBI-style tones and break indices, while disabling MARY default prosody rules.

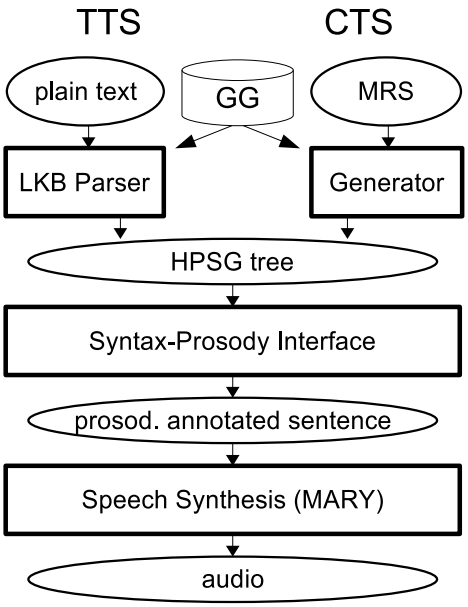

Figure 1: Architecture

1.2 Application scenarios

1.2.1 Platform for hypothesis testing at the syntax-prosody interface

For the purposes of the current experiment, input MRSs and corresponding surface realisations were chosen manually using the comparison tool provided by the LKB and [incr tsdb()] (Oepen, 2002)

This is probably the preferred procedure when using the HPSG-based syntax-prosody module as a tool for automatic and controlled generation of experimental stimuli, since precise control over the selected realisation is of utmost importance.

1.2.2 Text-to-Speech (TTS)

In a text-to-speech application scenario, however, manual selection is not really an option. Fortunately, though, the LKB and Pet (Callmeier, 2000) processing platforms both support maximum entropy discriminative parse selection models, on the basis of which syntactic disambiguation can be performed.

For the German grammar GG, exact match accuracy on dialogue data currently averages at around 81.3% (10-fold cross validation) compared to a random baseline of 25.4%. The model has been trained on a Redwoods-style treebank for German, derived from the Verbmobil corpus. Currently, the treebank consists of over 10,000 manually disambiguated trees. As features, the parse selection model uses local trees of depth 1 plus grand parenting. The parse selection results achieved for German are comparable to those reported by Oepen et al. (2002) and Toutanova and Manning (2002) for the English Resource Grammar (ERG) using similar data.

1.2.3 Concept-to-Speech (CTS)

The problem of realisation-ranking in a CTS scenario is related, though not identical to that of parse selection: here the task is to choose the most natural surface realisation given an input MRS. Fortunately again, models to perform this task can be derived quite cheaply, using a method suggested by (Velldal and Oepen, 2005): on the basis of a disambiguated parsing treebank they use the LKB generator to derive a corresponding generation treebank, taking the surface realisation found in the original corpus as gold standard. Combining a maximum entropy model trained on this generation bank with n-gram language models, they report an exact match accuracy in realisation ranking of 65%. Although we have not yet evaluated this for German, we expect to achieve similar results, given the comparatively similar performance of the two grammars in parse selection.

1.3 Background

1.3.1 Modifier attachment

Probably one of the most thoroughly studied types of structural ambiguity in human language are attachment ambiguities. While most research in this area has focused on written language, more recently, there has been a number of detailed studies of how prosody contributes to disambiguation, most notably the work of Schafer (1997) and Speer et al. (2003). Using task-oriented elicited speech, Schafer (1997) identified the prosodic parameters responsible for disambiguation of attachment ambiguities in English as follows: High attachments are perceived best when there is a prosodic break before the modifier, but not between the preceding object NP and the verb. Conversely, low attachment corresponded to the absence of a prosodic break between the modifier and the NP to which it is attached; the entire object NP, including the modifier, was preceded by a prosodic break. Speer et al. (2003) observe that, high attachment is characterised by an increased duration of the head

noun and following pause, which was verified perceptually.

1.3.2 Object fronting

In German, both subjects and objects can appear in sentence-initial topic position, preceding the finite verb. Since nominative and accusative case are not always distinct, local or even global ambiguity can arise with regard to grammatical function. Subjects in topic position are generally considered unmarked. In an eye-tracking experiment using resynthesised speech, Weber et al. (2006) showed that prosodic information leads to Early Effects with sentences involving local ambiguity. Using an L+H* contour on fronted objects followed by a steep fall achieved early disambiguation, even against a strong bias for subject topics.

2 Perception Experiment

In order to quantify the potential for prosodic disambiguation, we carried out a perception experiment, comparing how subjects interpret prosodically disambiguated stimuli as compared to their ambiguous textual counterparts. Furthermore, we used different candidate contours for each of the intended interpretations in order to measure which combination of tones and breaks will perform best.

2.1 Method

The experiment was designed as an online study; subjects were not observed. To make sure that the task was clearly explained, the main study was preceded by a pilot, involving 5 subjects.

The main study was carried out with 58 subjects (27 female, 31 male). They were aged from 17 to 54, and came from all parts of Germany.

Subjects had to assign an interpretation each stimulus in a self-paced forced-choice test. Each stimulus could be heard as often as required.

In order to control for semantic or pragmatic preferences, subjects first had to judge stimuli presented in text form. 4 different sentences were used for modifier attachment and 2 for object fronting. From these sentences we generated 4 different speech stimuli for modifier attachment and 3 for object fronting, yielding a total of 6 textual and 22 randomised speech stimuli per subject.

2.1.1 Stimuli for Modifier Attachment Experiment

The sentences involving modifier attachment (MA) ambiguities all followed the same basic syntactic pattern subject-verb-object-modifier (S-V-O-M).

(1) a. Rainer verfolgt den Mann mit dem Motorrad.
Rainer chases the man on the motorbike
'Rainer chases the man on the motorbike.'

b. Er begutachtete den Tisch vor dem Schrank.
He inspected the table before/in front of the cupboard
'He inspected the table before/in front of the cupboard.'

c. Ich sehe den Mann mit dem Fernglas.
I see the man with the binoculars
'I see the man with the binoculars.'

d. Er schlug die Frau mit dem Spazierstock.
He beat the woman with the walking stick
'He beat the woman with the walking stick.'

For the generation of disambiguating auditive stimuli, we used a combination of prosodic breaks and tones (Grice and Baumann, 2002). In order to determine which prosody gives the best results in terms of naturalness and disambiguation, we tested 4 different tonal patterns, 2 each for high and low attachment.

High Modifier Attachment S [[V O] M]

- H1:
  - L* on object head noun
  - H- before modifier
- H2:
  - L+H* on object head noun
  - L- before modifier

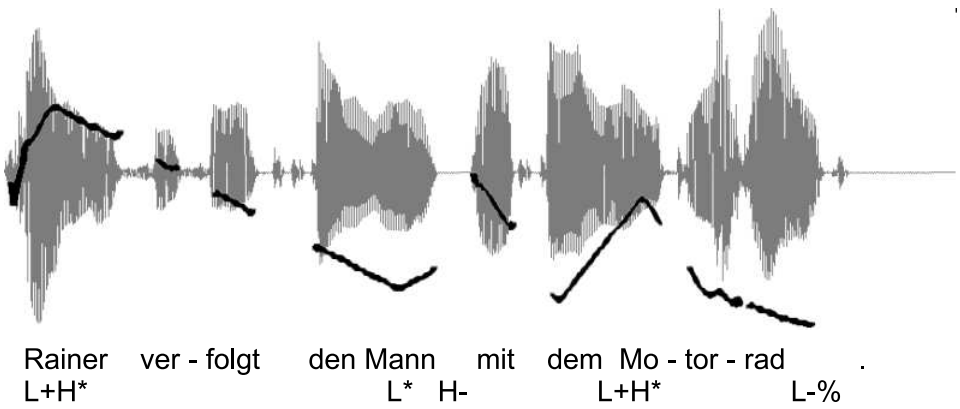

Figure 2: Tonal contour for high attachment: H1

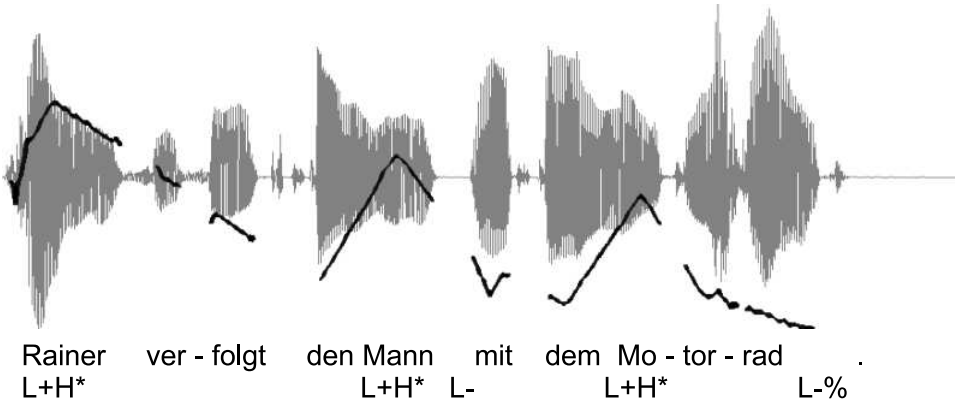

Figure 3: Tonal contour for high attachment: H2

Neither had any break before the direct object (O). The other two possible tone combinations, i.e., H* H- and L* L- sounded unnatural and were therefore discarded during the pilot study already.

Low Modifier Attachment S [V [O M]]

- L1: no break before object
- L2: H- before object

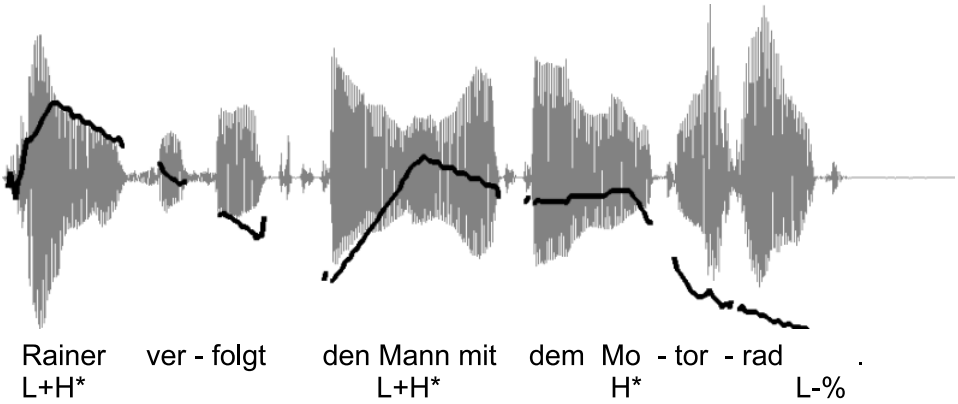

Figure 4: Tonal contour for low attachment: L1

Both versions contained an L+H* on the object and no break before the modifier.

2.1.2 Stimuli for Object Fronting Experiment

The disambiguating speech stimuli for the object vs. subject fronting subtask were based on Weber et al. (2006). Since timing was not an issue in our study, contrary

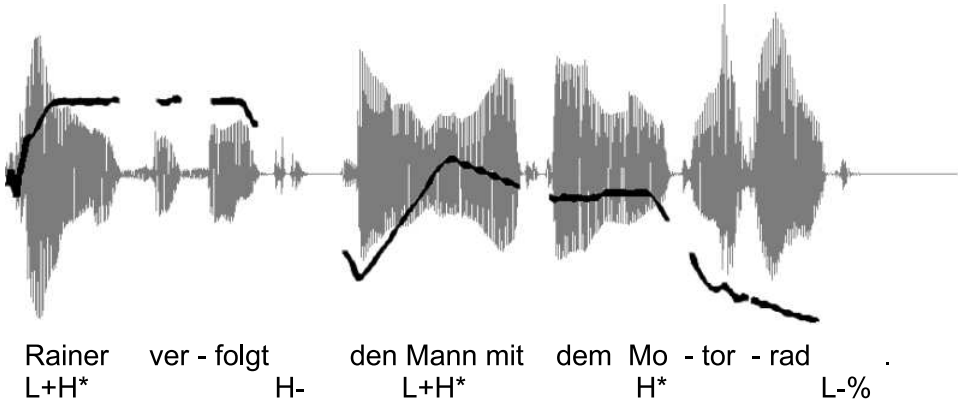

Figure 5: Tonal contour for low attachment: L2

to Weber et al. (2006), we inserted an additional intonation phrase break after the fronted object in OVS-sentences. Also in contrast to Weber et al. (2006), ambiguity was global, not local. The resulting utterances synthesized by use of the prosody module has the following prosody:

SVO no intonation phrase break after fronted subject

OVS
- OVS1: L- after fronted object
- OVS2: H- after fronted object

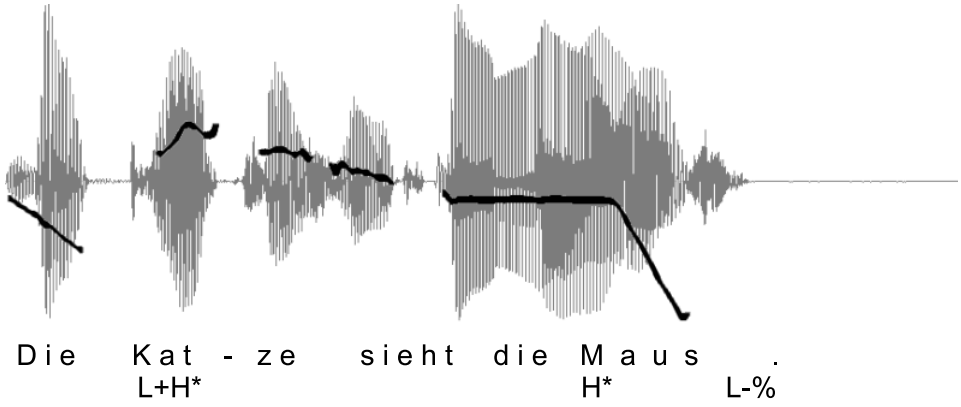

Figure 6: Tonal contour for subject fronting (SVO)

Both, SVO and OVS, ended in an L-% boundary tone and had an L+H* accent on the fronted constituent. In the OVS-versions this accent was additionally emphasized by raising the peak, thus strengthening accent-prominence. The tonal pattern used for SVO was based on the default contour in MARY.

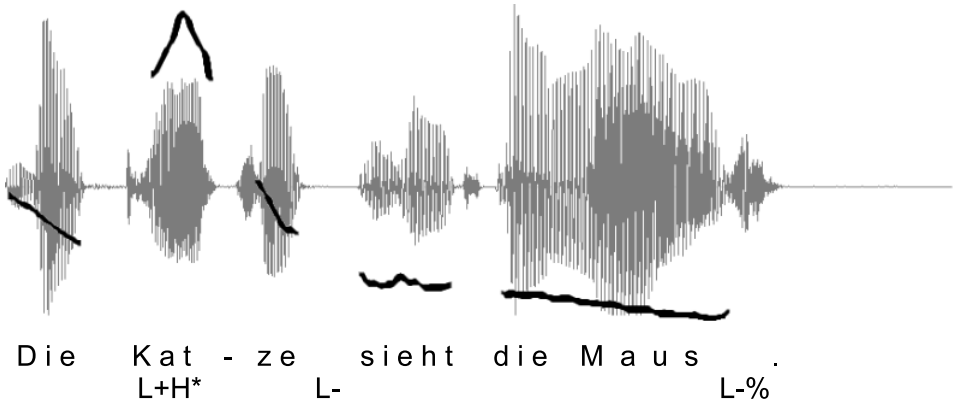

Figure 7: Tonal contour for object fronting (OVS1)

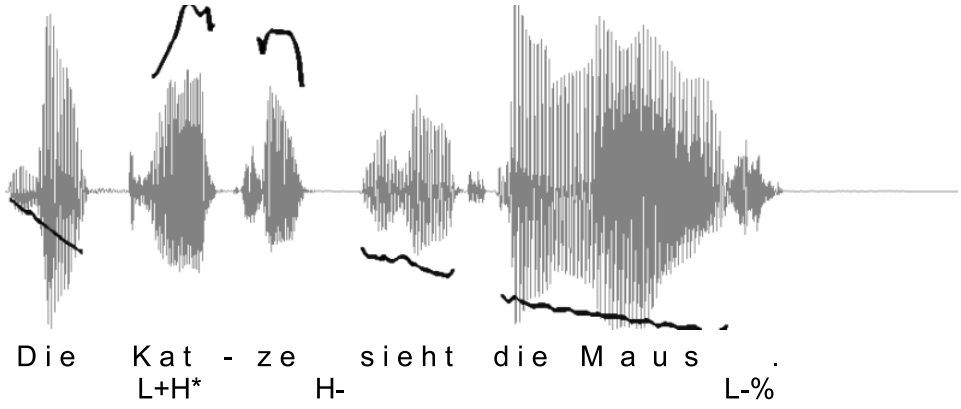

Figure 8: Tonal contour for object fronting (OVS2)

2.2 Results

The main experimental results are summarised in Figures 9 and 10. As compared to baseline obtained with textual stimuli (bias), the perception experiment shows a clear disambiguation effect with speech stimuli, for both modifier attachment and subject vs. object fronting. Our main claim that prosody automatically generated from deep syntactic structures can be used for the task of disambiguation in CTS and TTS was confirmed.

2.2.1 Modifier attachment

Best disambiguation results were obtained with contours H1 and L2, given in Figures 2 and 5. The results for these contours are summarised in Figure 9, where a value of 1 corresponds to perceived high attachment, and 0 to low. Interpretations assigned are provided for each of the 4 test sentences, averaged over all 58 subjects. Test sentences differed as to their inherent semantic attachment preferences (bias calculated from textually presented stimuli): while (a) does not display any clear preference, (b) and (c) have a strong preference for low attachment, while (d) is mainly attached high.

The most important result is that a clear disambiguation effect could be found not only for ambiguous sentences without any clear semantic attachment preference, such as (a), but that automatically generated prosody could effectively override even strong preferences for low (b,c) or high attachment (d).

With sentences showing a strong bias for low attachment, we observed that the speech stimuli designed to suggest low attachment do not quite reach the level of the bias. We tentatively attribute this difference to a mismatch between expected and actual prosodic contours in synthetic speech, which can hopefully be overcome on the basis of better prosody planning to be obtained from future experimental studies.

As a measure of the disambiguation effect we take the span between perceived high and low attachment. For H1 and L2 it ranges from 0.23 (=0.83-0.60; utterance d) to 0.47 (=0.69-0.22; utterance c), with an average around 0.37 (=0.76-0.39). The value for H2 (average: 0.63) shows a far lower disambiguation potential than H1, while the value reached by L1 (average: 0.50) proves this contour unsuitable for the task. This latter result confirms for German the findings made in Schafer (1997) for English that insertion of a pre-object boundary enhances perception of low attachment. The results for the high attachment contours (H1 vs. H2), however, suggest that choice of tones is almost as important as break insertion, in order to maximise the disambiguation effect.

2.2.2 Object Fronting

Results confirm previous findings on prosody-induced early effects, as well as our own claim concerning the disambiguation potential of prosody in speech synthesis.

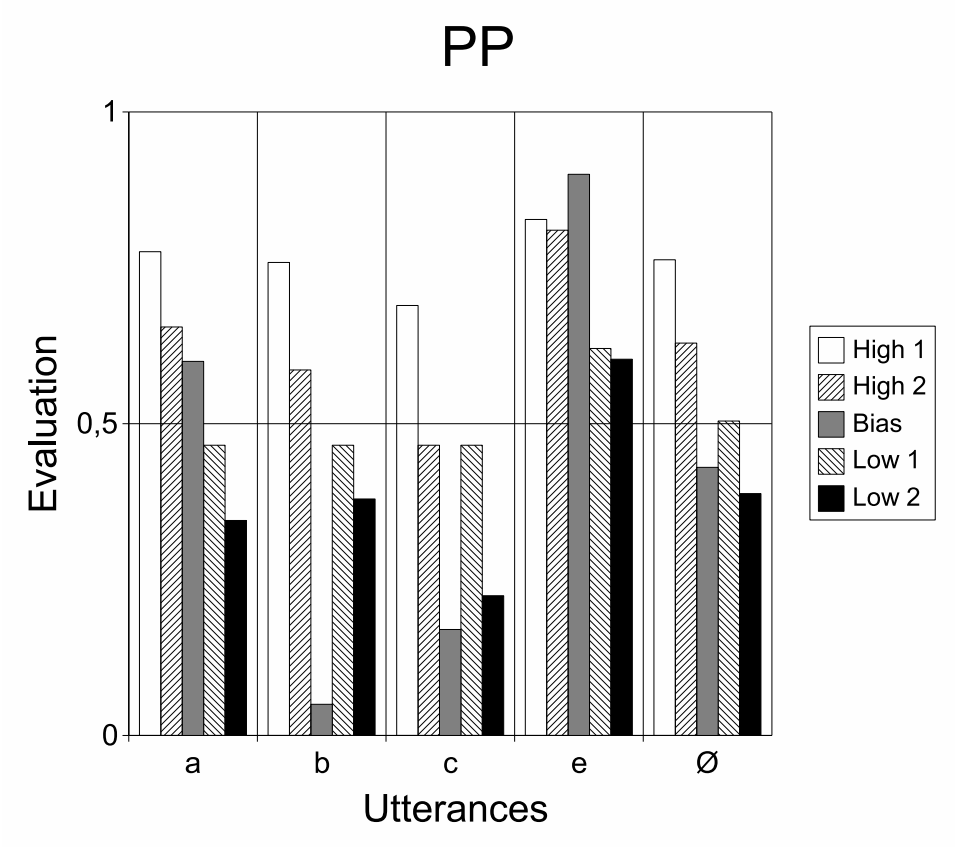

Figure 9: Interpretation of disambiguating contours for modifier attachment: high (H1,H2), textual bias, low (L1,L2), for each 4 test sentences.

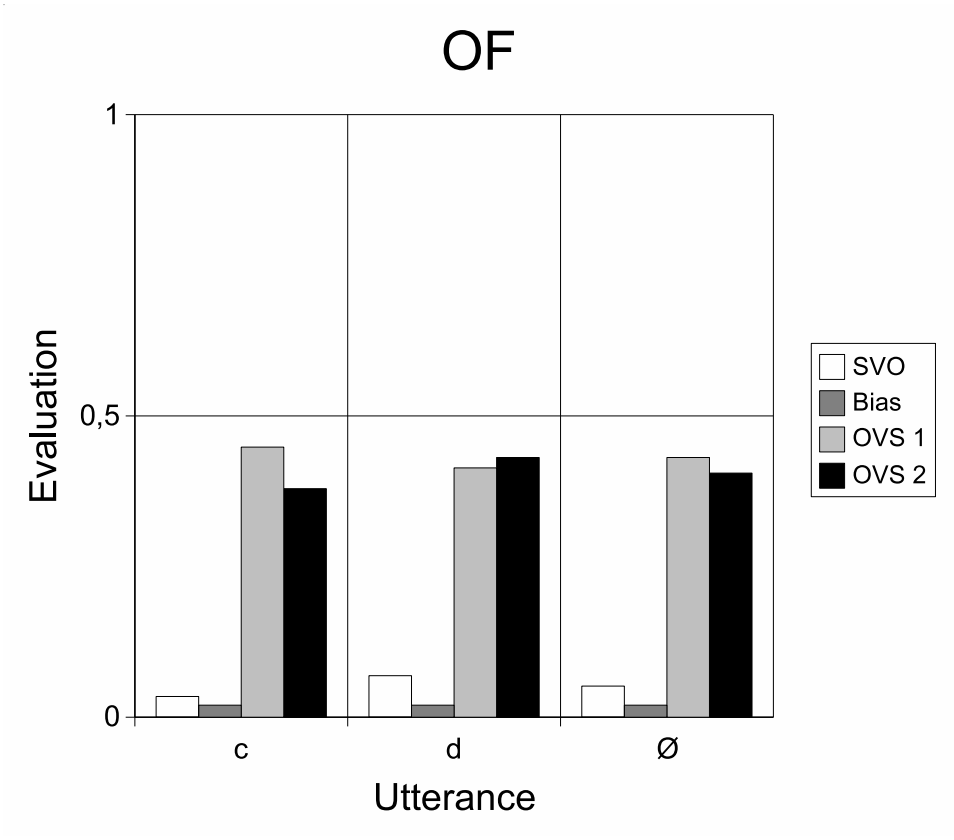

Figure 10: Interpretation of disambiguating contours for object fronting: SVO, textual bias, OVS1

Details are provided in Figure 10, where, again, a value of 1 corresponds to perceived object fronting, and 0 to subject fronting. In contrast to modifier attachment, however, the bias for SVO was extremely high (=0.02 for OVS). Still, by means of carefully designed disambiguating prosody, it was possible, with an average value of 0.43 (OVS1), to make available, for interpretation, a reading that was practically inaccessible with textual stimuli.

Although the values for unmarked SVO (e: 0.03; f: 0.07; average: 0.05) do not fully reach the bias determined with textual stimuli, we believe that these differences are negligible. The strength of the disambiguation effect, that is, how well prosody distinguishes SVO and OVS interpretation averages at 0.38 for OVS1 and at 0.36 for OVS2.

An observation that deserves discussion here, is that the disambiguation effect obtained with OVS1 and OVS2 is almost identical, despite the difference in tonal realisation, namely in the choice of the boundary tone. This is somewhat surprising at first, since with high modifier attachment, we observed a clear impact of the choice of tones. However, difference in tonal realisation is far less salient in the case at hand, compared to the test contours for high attachment: first, tonal difference only involves the choice of boundary tone here, whereas in the case of high attachment, it extends to the nuclear pitch accent. Furthermore, in the case of OVS, realisation of the boundary tone falls on a reduced, and short vowel (schwa). We hypothesise that the combination of these effects makes the tonal differences between these contours hard to perceive.

The contours for subject fronting and for object fronting (OVS1) are given in figures 6, 7, and 8.

3 Conclusion

In this paper we have presented experimental evidence showing how prosody automatically generated from deep syntactic trees can be used successfully to disambiguate structural ambiguities in German. The results we obtain using prosodically enhanced diphone synthesis compare well to disambiguation rates previously achieved for English modifier attachment ambiguities using human speech stimuli: Schafer (1997) reports a value of 0.651 in response to high attachment stimuli similar to our H1, and a value of 0.472 with low attachment stimuli similar to our L2, yielding an overall disambiguation effect of 0.18, compared to our 0.37. The result we obtained with object fronting further suggest that the disambiguating effect (0.38) of our automatically generated prosody is very robust, even against a very strong bias for subject fronting.

The disambiguation effects we obtain with synthesised speech also underline the potential of prosody derived from deep syntactic structures for the improvement of intelligibility in TTS applications. Finally, the fact that automatically generated stimuli can achieve disambiguation rates comparable to human speech makes our system a valuable test bed for studies at the syntax-prosody interface.

References

Callmeier, Ulrich. 2000. PET — a platform for experimentation with efficient HPSG processing techniques. *Journal of Natural Language Engineering* 6(1):99–108.

Copestake, Ann. 2001. *Implementing Typed Feature Structure Grammars*. Stanford: CSLI Publications.

Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan Sag. 2005. Minimal recursion semantics: an introduction. *Research on Language and Computation* 3(4):281–332.

Crysmann, Berthold. 2003. On the efficient implementation of German verb placement in HPSG. In *Proceedings of RANLP 2003*, pages 112–116. Borovets, Bulgaria.

Crysmann, Berthold. 2005. Relative clause extraposition in German: An efficient and portable implementation. *Research on Language and Computation* 3(1):61–82.

Grice, Martine and Stefan Baumann. 2002. Deutsche Intonation und GToBI. *Linguistische Berichte* 191:267–298.

McKeown, K. R. and S. Pan. 2000. Prosody modelling in concept-to-speech generation: methodological issues. *Philosophical Transactions of the Royal Society* 358:1419–1431.

Müller, Stefan. 2004. Continuous or discontinuous constituents? A comparison between syntactic analyses for constituent order and their processing systems. *Research on Language and Computation* 2(2):209–257.

Müller, Stefan and Walter Kasper. 2000. HPSG analysis of German. In W. Wahlster, ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 238–253. Berlin: Springer.

Oepen, Stephan. 2002. *Competence and Performance Profiling for Constraint-based Grammars: A New Methodology, Toolkit, and Applications*. Ph.D. thesis, Saarland University.

Oepen, Stephan, Dan Flickinger, Kristina Toutanova, and Christoper D. Manning. 2002. Lingo redwoods: A rich and dynamic treebank for HPSG. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT2002)*. Sozopol, Bulgaria.

Schafer, A. 1997. *Prosodic parsing: the role of prosody in sentence comprehension*. Ph.D. thesis, University of Massachusetts.

Schröder, Marc and Jürgen Trouvain. 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology* 6:365–377.

Toutanova, Kristina and Christopher D. Manning. 2002. Feature selection for a rich hpsg grammar using decision trees. In *Proceedings of CoNLL 2002*. Taipei, Taiwan.

Velldal, Erik and Stephan Oepen. 2005. Maximum entropy models for realization ranking. In *Proceedings of the 10th MT-Summit (X), Phuket, Thailand*.

Weber, Andrea, Martine Grice, and Matthew Crocker. 2006. The role of prosody in the interpretation of structural ambiguities: a study of anticipatory eye movements. *Cognition* 99(2):B63–B72.