Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

Integration of Prosodic and Grammatical Information in the Analysis of Dialogs

Walter Kasper
Hans-Ulrich Krieger

DFKI

Vm
Report 141
July 1996

July 1996

Walter Kasper
Hans-Ulrich Krieger

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Stuhlsatzenhausweg 3
66123 Saarbrücken

Tel.: (0681) 302 - 5282
Fax: (0681) 302 - 5341
e-mail: {kasper,krieger}@dfki.uni-sb.de

Gehört zum Antragsabschnitt: 8.7 Dialogsemantik
15.8 Nichtsyntaktische Information für die semantische Auswertung

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 K/1 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

Integration of Prosodic and Grammatical Information in the Analysis of Dialogs*

Walter Kasper & Hans-Ulrich Krieger

Abstract

The analysis of spoken dialogs requires the analysis of complete multi-sentence turns. Especially, the segmentation of turns in sentential or phrasal segments is a problem. In this paper we present a system for turn analysis. It is based on an extension of HPSG grammar for turns and takes into account extra-linguistic prosodic information. We show how this information can be integrated and represented in the grammar, and how it is used to reduce the search space in parsing.

1 Introduction

A fundamental concept of Head-Driven Phrase Structure Grammar (HPSG; cf. [6, 7]) is the notion of a SIGN. A SIGN is a structure integrating information from all levels of linguistic analysis such as phonology, syntax and semantics. This structure also specifies interactions between these levels by means of coreferences which indicate the sharing of information and how the levels constrain each other mutually. Such a concept of linguistic description is attractive for several reasons:

- it supports the use of common formalisms and data structures on all levels of linguistics
- it provides declarative and reversible interface specifications between the levels
- all information is available simultaneously
- no procedural interaction between linguistic modules needs to be set up

*This report will also be published in: *Proc. of the 20th German Annual Conference on Artificial Intelligence, KI-96*, September, 17-19, 1996, Dresden.

Though the concept of SIGN is very general grammars developed in this framework usually only deal with morphological, syntactic and perhaps semantic information. Also, they are confined to the description of phrases not beyond the level of single sentences.

On the other hand, the VERBMOBIL project (cf. [13, 3]) deals with the translation of spoken dialogs. The basic unit of natural language dialogs is not a sentence but a *turn* representing the complete contribution of a participant. *Turns* usually consist of more than one segment as they would be described in a sentence-based grammar. One fundamental problem in analyzing dialog turns is to segment them correctly into such smaller units as described in a grammar. Correct segmentation is not only crucial for the correct semantic and pragmatic interpretation but also for the efficiency of the parsing process in order to reduce the search space. We will call this the *segmentation problem*. In the case of written text punctuation marks help in segmentation. In spoken language there are no punctuation marks, and audible breaks in spoken utterances often do not correspond to sensible linguistic phrase boundaries. Such breaks can be coughs or they are due, e.g., to breathing, hesitation or corrections. On the other hand there are prosodic, intonational clues to support the linguistic segmentation task. But this requires that the linguistic analysis process is sensitive to such grammar-external information. In this paper we describe the integration of grammatical and prosodic information from the representational as well as the computational point of view. The approach can serve as a model for integrating linguistic and other types of non-linguistic information as well.

In the following we first give a survey of the underlying system architecture and the types of prosodic information used (*focus* and utterance *mood*). After that, an extension of a HPSG grammar of German suitable for the analysis of dialog turns is discussed. Then the two kinds of interaction between grammatical analysis and prosody are described: on the one hand, the integration of prosodic information constrains the parsing process, on the other hand the representation of prosodic events constrains the possible distribution of such events. But prosodic information or its detection is not always reliable: it might be missing at expected positions or come in at wrong positions. Therefore a recovery procedure for prosodic errors is presented to make the parsing process robust.

2 Architecture

The architecture of the linguistic components of the dialog translation system is shown in Figure 1. The parsing process is distributed on two parsers running in tandem. The first one is the parser for the word lattices coming from speech recognition, the other one is the constraint solver which also builds the semantic representation.[1] The word-lattice parser [14] uses the full HPSG grammar offline

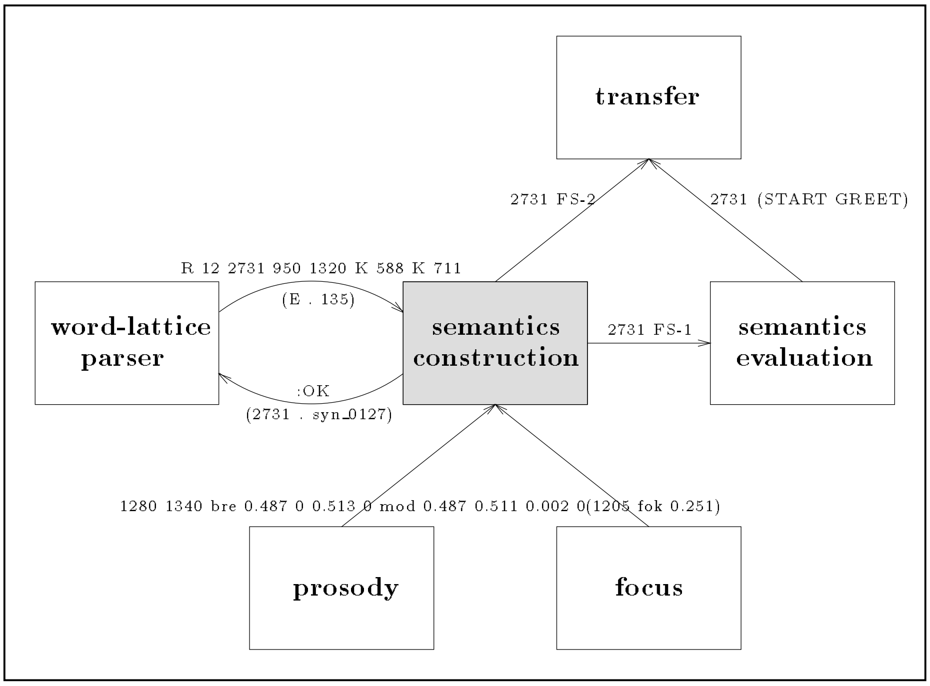

Figure 1: *The overall architecture of our experiments. The semantics construction (*SEM*-parser) integrates different sources of information: (i) word/rule application hypotheses from the word-lattice parser, (ii) prosodic boundaries, and (iii) focus. Both prosodic boundaries and focus information are mapped onto word lattice hypotheses inside the* SEM*-parser as described in Section 5. Legal readings are further sent to semantics evaluation and transfer. The annotations at the arrows depict the different protocols between the components.*

(viz., for training). At run time, only the context-free skeleton of the grammar is used. Because this set of rules overgenerates w.r.t. the original grammar, certain rule application are in fact not valid. Such applications are ruled out by the second parser.

Semantics construction (the so-called SEM-parser) is fed with hypotheses from the word-lattice parser and uses them to reconstruct deterministically the chart on the basis of the full grammar. This is possible by associating every lexicon entry and every rule with an identifier added at compile time and shared by both parsers. Because the search space is massively reduced by the word-lattice parser

[1] The approach to distributed parsing with HPSG grammars is described in [1]. A full description of the system is given in [2].

(approx. one order of magnitude less hypotheses), unification inside the SEM-parser is time-synchronous with the corresponding rule application inside the first parser.

This special architecture allows for efficient filtering of word hypotheses without giving up correctness of the analysis results which is guaranteed by the SEM-parser. Since lexicon entries and rules are identified by unique indexes, expensive communication via feature structures is avoided. In case that a rule application hypothesis fails under feature structure unification, a message is sent back to the word-lattice parser. This allows it to narrow down the set of emitted hypotheses.

The SEM-parser additionally receives hypotheses from two prosodic components. The one simply termed PROSODY is the detector for phrase boundaries and sentence mood as described in [12]. The other one is a detector for focus ([5]).

In VERBMOBIL three types of phrase boundaries are distinguished ([8]):

B2 "weak" phrase borders within intonational phrases

B3 "strong" intonational phrase border

B9 irregular phrase boundaries

Of these, the *B3* borders are the ones related to utterance mood and turn segmentation. Three types of moods are distinguished prosodically:

- progredient
- interrogative
- declarative (or rather "non-interrogative")

The PROSODY component transmits confidence values about the existence of a *B3* boundary together with confidence values about the mood. Similarly, the FOCUS components sends confidence values about focus events. Their use will be described in the following sections after a survey of the underlying HPSG grammar for turn analysis.

3 Codescriptive Grammars for Dialog Turns

The basic units of dialogs are not sentences but *turns*:

> tut mir leid. am neunundzwanzigsten um drei habe ich schon eine Besprechung. am Dienstag den dreißigsten um drei, das ginge bei mir. (I am sorry. On the 29th at 3 I already have a meeting. Thursday the 30th at 3, that would be fine)

Turns usually consist of more than one of what we call a *turn segment*. In the example the most likely segmentation is indicated by punctuation marks. Turn segments need not be complete sentences but can be sequences of nearly any kind of phrase:

> also. am Montag. um wieviel Uhr denn dann?
> (OK. On monday. At what time then?)

In spoken turns the punctuation marks of course are missing, and the fact that any kind of linguistic category can also be a turn segment, that is, a "complete" utterance in itself, makes segmentation on purely linguistic grounds a highly ambiguous task. A grammar provides only weak constraints on utterances such as

- subcategorization: verbs or prepositions require the presence of certain complements
- verb-end constructions in German (e.g., subordinate clauses) mark the end of a turn segment

On the other hand, a turn like *am montag kommt er* without any further clues can be understood as consisting of one declarative sentence but also as consisting of an elliptical prepositional phrase followed by an interrogative sentence.

In the VERBMOBIL project we use an HPSG grammar described in the typed feature formalism $\mathcal{TDL}$ (cf. [4]) for the analysis of dialog turns. It is a codescriptive grammar specifying simultaneously syntax and semantics. As an example, Figure 2 shows a lexicon entry of a verb with its syntactic subcategorization frame and predicate-argument-structure. In order to deal with turns consisting of several segments the HPSG approach had to be extended especially in order to deal with the semantic composition of the turn segments. Also non-linguistic events in a dialog turn, e.g. pauses and coughs required a treatment in the grammar. The additional rules for turns do not simply concatenate turn segments but impose an intermediate structure on turns between phrasal turn segments and complete turns in order to deal among other things with special properties of the *uptake* phase at the beginning of a turn, with interruptions, linguistic "garbage" and echo phrases. Semantically, turns are represented by a linear conjunction of the semantical representations of the turn segments which is passed to semantic evaluation for further processing, such as reference resolution and dialog act identification with respect to the dialog model.

Other extensions of the approach were required to capture information from prosody, especially information about the *mood* of an utterance and about focussed phrases.

$$
\begin{bmatrix}
\textit{intrans-fin-verb-lex} \\
\text{CAT-SYM } \textit{vfin-cat} \\
\text{SC } \left\langle \begin{bmatrix}
\textit{np-nom-type} \\
\text{SEM } \begin{bmatrix}
\textit{cont2quant} \\
\text{LAST } \langle\,\rangle \\
\text{QUANT } \begin{bmatrix}
\textit{*diff-list*} \\
\text{LIST } \left\langle \boxed{8} \begin{bmatrix} \text{VAR } \boxed{5} \begin{bmatrix} \textit{var-type} \\ \text{AGR } \boxed{7} \begin{bmatrix} \textit{sg-1-type} \\ \text{PERS } \textit{first} \\ \text{NUM } \textit{sg} \end{bmatrix} \\ \text{SORTE } \textit{moveable} \end{bmatrix} \end{bmatrix} \ldots \right\rangle \\
\text{LAST } \boxed{9}[\,]
\end{bmatrix} \\
\text{CONTENT } \boxed{8}
\end{bmatrix} \\
\text{CAT-SYM } \textit{np-cat} \\
\text{SL } \langle\,\rangle \\
\text{SC } \langle\,\rangle \\
\text{SYN } \begin{bmatrix}
\text{LOCAL } \begin{bmatrix} \text{DET } \textit{yes} \end{bmatrix} \\
\text{HEAD } \begin{bmatrix} \text{AGR } \boxed{7} \\ \text{CASE } \boxed{1}\, \textit{nom-val} \end{bmatrix}
\end{bmatrix}
\end{bmatrix} \right\rangle \\
\text{SEM } \begin{bmatrix}
\textit{semantics-type} \\
\text{LAST } \boxed{2}[\,] \\
\text{QUANT } \begin{bmatrix} \textit{*diff-list*} \\ \text{LIST } \boxed{6}[\,] \\ \text{LAST } \boxed{6} \end{bmatrix} \\
\text{CONTENT } \begin{bmatrix}
\textit{rp-type}\wedge\textit{subwff-inst-share-var} \\
\text{QFORCE } [\,] \\
\text{VAR } \boxed{4} \begin{bmatrix}
\textit{var-type} \\
\text{AGR } \textit{no} \\
\text{CLASS } \boxed{3}\, \textit{kommen} \\
\text{OBL } \boxed{5} \\
\text{SORTE } \textit{extended} \\
\text{TEMP } \begin{bmatrix} \text{TENSE } \textit{pres} \\ \text{MOOD } \textit{ind} \end{bmatrix}
\end{bmatrix} \\
\text{COND } \begin{bmatrix}
\textit{conj-wff-type} \\
\text{CONN } \texttt{and} \\
\text{SUB-WFFS } \left\langle \begin{bmatrix} \textit{atomic-wff-type} \\ \text{INST } \boxed{4} \\ \text{AGENT } \boxed{5} \\ \text{PRED } \boxed{3} \end{bmatrix} \ldots \right\rangle
\end{bmatrix}
\end{bmatrix}
\end{bmatrix} \\
\text{SYN } \begin{bmatrix} \text{HEAD } \begin{bmatrix} \text{CASE } \boxed{1} \\ \text{AGR } \boxed{7} \\ \text{PERF } \texttt{"haben"} \end{bmatrix} \end{bmatrix} \\
\text{PHON } \texttt{"komme"}
\end{bmatrix}
$$

Figure 2: *Lexical entry for the verb* komme *containing its syntactic and semantic properties simultaneously.*

The main problem for such a turn-based grammar is the problem of segmenting a turn into the correct turn segments. As indicated above, linguistic constraints are very weak and not sufficient. On the other hand, spoken language contains clues about segmentation. There is a significant prosodic difference when *on monday he will come* is uttered as one statement or a sequence of two segments each expressing a different speech act.

Taking into account such prosodic clues for turn segmentation is not only important for a correct grammatical analysis but also for the efficiency of the analysis process itself. Experiments showed that correct segmentation reduces the search space for the parser up to 70% by eliminating the segmentation ambiguities which give rise to different readings.

4 Integration of Grammar and Prosody

4.1 Typed Interfaces for Mood and Focus

Since the prosodic information especially about mood and focus is relevant for semantic interpretation it must be incorporated into the semantic representation built up in the parsing process. The use of a typed feature formalism allows an elegant and flexible solution to this task. We associate a type as shown below with each kind of prosodic event:

```
;;; types for representing prosodic mood

prosodic-decl-type := prosodic-b3-mood-type &
                      [ PRAG.PMOOD deklarativ-s ].

prosodic-frage-type := prosodic-b3-mood-type &
                       [ PRAG.PMOOD frage-s ].

;;; type for marking the focussed word

phon-focus-mark-type := [ PHON #focus,
                          SEM.CONTENT.VAR.FOCAL #focus ].
```

When a prosodic event occurs the associated type is unified with the feature structures for the word hypotheses which include the time of the prosodic event. This approach yields a typed interface between grammar and prosody with the following advantages:

- *Flexibility:* it is easy to modify the representation of mood since it is sufficient to change the type definition

- *Constraint Interaction:* it allows the grammar to use the prosodic information in a straightforward way as additional constraints which might interact with other constraints such as syntactic sentence mood. Constraints on prosodic events are discussed below (sections 4.2 and 4.3)
- *Reversibility:* the grammar can constrain the kind and loci of prosodic events. This is important in generating spoken language.

4.2 Constraining Prosodic Mood

The information about prosodic mood is first incorporated at the lexical level into the feature structure by unification. This means it fills the value of PMOOD (for *prosodic mood*) in the PRAGmatics substructure of the linguistic SIGN. Since the mood is not a property of the word but rather of the turn segment which is terminated by that word it is projected to the turn segment level along the right edge of the derivation tree for the segment. We call this the *prosodic-mood-principle* which is inherited by each rule schema:

```
prosodic-mood-principle := [ PRAG.PMOOD #pmood ]
                              -->
                            < ..., [ PRAG.PMOOD #pmood ] >.
```

The principle ensures that each turn segment can be marked for prosodic mood at most once. This predicts that between two *B3* hypotheses there must be a segment boundary.[2] In this way, the prosodic-mood-principle supports turn segmentation independently of the parser. It is important to note that the principle is not in conflict with the delay strategies the parser employs for reducing its search space on prosodic events (see section 5), especially not with its recover strategy: since the mood information is projected *only* along the right edge of the tree, the recognition of a prosodic boundary at a position which syntactically cannot be a segment boundary (e.g. between a preposition and its object) will do no harm as the prosodic information will be kept local in the word but will not be projected to higher phrasal levels.

4.3 Focus

Focus in spoken language serves to mark phrases by stress. Focus is not only important semantically as being associated with certain semantic operators ([10, 9]) but has also important discourse functions by highlighting on important parts

[2] Of course, this presupposes the reliability of the detector.

of the utterance. Also, for translation from German to English the focus position can make all the difference as in the following example:

> Lassen Sie uns *noch* einen Termin ausmachen
> (= Let's arrange another appointment)
>
> Lassen Sie uns noch einen *Termin* ausmachen
> (= Let us arrange an appointment, too)

The hypotheses from focus detection are incorporated into the lexical structure of the word by unification with the `phon-focus-mark-type`. This type marks the focussed word on the semantic index (the `SEM.CONTENT.VAR` structure) of the complete phrase. This expresses a strong constraint on possible focus distribution because this semantic index is global for the maximal projection line. The constraint allows that a turn segment carries more than one focus (as actually happens). But, within a maximal phrase, though any word can be the focussed one there cannot be more than one, as illustrated in the following example:

- *this* small man
- this small *man*
- * *this* small *man*

The data from VERBMOBIL dialogs show no violation of these principles.

Figure 3 shows the results of integrating the prosodic information in the first segment of the turn

> das ist *schlecht* . wie wär's am Dienstag.
> (= That's bad. How about tuesday?)

where the word *schlecht* was focussed and also correctly recognized as segment boundary. This is marked in the representation of the verbal head's `INST`ance structure in the features `FOCAL` and `PRAG.PMOOD`. If the focus had been on *that* the focus marker would have occurred on the index-`VAR` of the first element on the `QUANT`ifier list being the representation of *that*.

5 Rule Selection in the Parser

Information about utterance boundaries and focus is not only represented in the semantic analysis but also employed within the SEM-parser to reduce the space of possible rule applications. Mapping parsing hypotheses onto prosodic information is achieved by means of the signal time which is available to all modules.

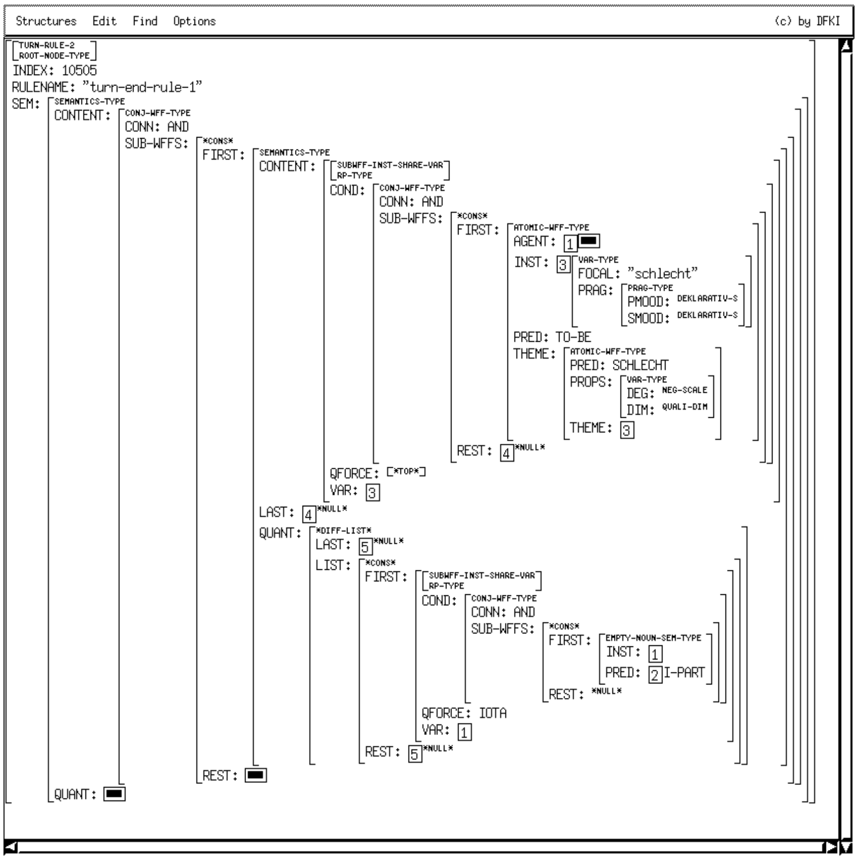

Figure 3: *Semantic turn representation of* das ist schlecht *(that is bad) with focus on* schlecht *and mood information.*

This is depicted in Figure 4. Information about *B3* boundaries is used inside the parser to rule out certain rule application. This is achieved by telling the parser which rules are segment-connecting and which are only segment-internal. Clearly, segment-connecting rules enforce a *B3* boundary between segments—actually between the last lexical chart edge of one segment and the first lexical edge of the following segment. The opposite case holds for the segment-internal rules: here, no *B3* boundary is allowed inside a chart edge which originates from such a rule application. As already mentioned, such constraints heavily reduce the number of possible readings of an utterance up to 70%. This increases parsing efficiency considerably (smaller chart, less unification/copying), but also non-valid readings are ruled out thereby.

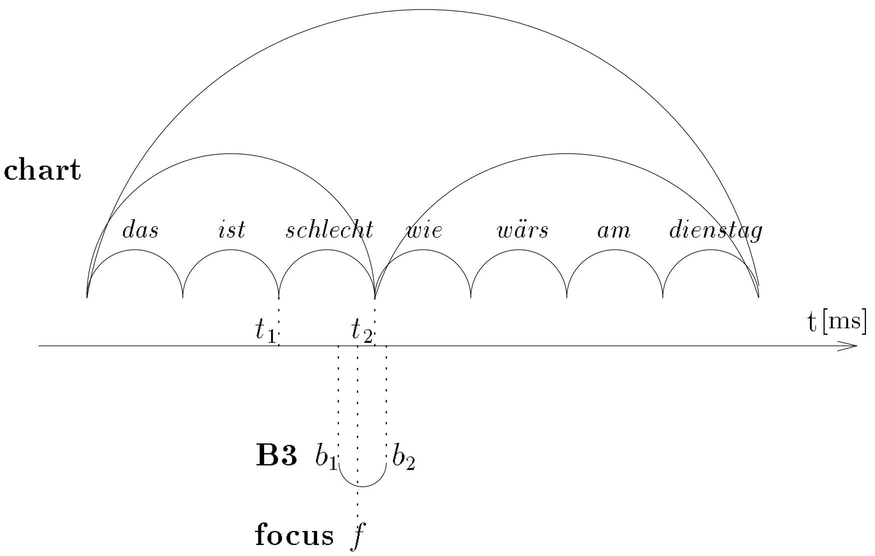

Figure 4: *Mapping* B3 *and focus information onto lexical chart edges.* B3 *boundaries are given in terms of time intervals which must satisfy the constraint* $t_1 \leq b_1 \leq t_2$ *with respect to the lexical item's time interval* $[t_1, t_2]$. *The focus is specified via a time point for which the constraint* $t_1 \leq f \leq t_2$ *must hold. Both* B3 *and focus are associated with confidence values which must be above a threshold. In our case,* schlecht *both bears the segment boundary as well as the focus.*

6 Robustness

Unfortunately, the prosodic data are not as reliable as the selection mechanism described above presupposes. The recoginition rate for the boundaries is about 86% ([11]). So, often segment boundaries are not detected and therefore missing and sometimes the position is excluded on syntactic grounds. Therefore, the system must be robust enough to handle such situations of missing or misplaced boundary information.

One mechanism for protecting the parser against wrong prosodic information is the use of thresholds. *B3* and focus information carry a confidence value between 0 and 1, expressing the (un)certainty of the data. There are different values for the mood information, that is, for progredient, interrogative, and declarative mood. The SEM-parser accepts a *B3* boundary value only if the confidence is above a certain threshold (and not merely above 0.5, i.e., greater than non-*B3*). One can further refine the thresholds additionally for the separate moods. Only if both thresholds are surpassed, the mood information is included into the corresponding lexical edges. The same holds for the focus information.

But this mechanism alone cannot prevent that prosody postulates an utterance boundary at a wrong position. In such a situation, the selection mechanism for rules would not allow that segment-internal rules are applied here. Clearly, this might lead to unwanted readings or even make an analysis impossible.

In order to deal with such situations we designed a dynamic recovery mechanism which allows to reuse previously excluded hypotheses. Instead of removing them completely from the agenda, their application is only delayed. Delayed hypotheses originate in different situations:

1. *Lexical hypotheses.* For lexical items for which prosodic information exists a copy of the original edge is added to the set of delayed hypotheses.

2. *Rule hypotheses.* Segment-connecting/segment-internal rule hypotheses which are excluded on prosodic grounds are also added to that set of delayed hypotheses.

3. *Missing hypotheses.* Rule hypotheses which depend on delayed/excluded edges must also be delayed.

Delayed hypotheses are applied only when the agenda of "legal" edges is empty and has not led to an analysis.

7 Conclusion

The VERBMOBIL project deals with translation of spoken dialogs. This requires the linguistic analysis components to be capable to deal with complete turns in a dialog because there are no obvious sentence boundaries in spoken language as in written language. Also, they must be capable to take into account extra-linguistic information from the speech signal. In this paper we presented an approach to solve these problems. We suggested an extension of sentence grammars to the turn-level. We showed how in a typed feature formalism such as $\mathcal{TDL}$ extra-linguistic information from prosody can be integrated elegantly with linguistic information embodied in the HPSG grammar. We also showed how the grammar can exploit this information as additional constraints. Additionally, we described how the parser itself can exploit the extra-linguistic information to reduce its search space in a robust manner, and so improve the parsing efficiency. The methods developed have the potential for application in other areas as well.

References

[1] Walter Kasper, Hans-Ulrich Krieger, and Abdel Kader Diagne. Distributed parsing with HPSG grammars. In *Proceedings of the 4th International Workshop on Parsing Technologies, IWPT-95*, pages 79–86, Prag, 1995.

[2] Walter Kasper, Hans-Ulrich Krieger, Jörg Spilker, and Hans Weber. From word hypotheses to logical form: An efficient interleaved approach. In *Proceedings of KONVENS '96*, Bielefeld, 1996.

[3] Martin Kay, Jean Mark Gawron, and Peter Norvig. *Verbmobil. A Translation System for Face-to-Face Dialog*, volume 33 of *CSLI Lecture Notes*. Chicago University Press, 1994.

[4] Hans-Ulrich Krieger and Ulrich Schäfer. $\mathcal{TDL}$—a type description language for constraint-based grammars. In *Proceedings of the 15th International Conference on Computational Linguistics, COLING-94, Kyoto, Japan*, pages 893–899, 1994.

[5] Anja Petzold. Strategies for focal accent detection in spontanous speech. In *Proc. of the 13th ICPhS*, pages 672–675, 1995.

[6] Carl Pollard and Ivan A. Sag. *Information-Based Syntax and Semantics.* Vol. 1: Fundamentals, volume 13 of *CSLI Lecture Notes*. Stanford: CSLI, 1987.

[7] Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar.* Chicago: University of Chicago Press, 1994.

[8] Matthias Reyelt. Ein System zur prosodischen Etikettierung von Spontansprache. Verbmobil-Report 86, TU Braunschweig, 7 1995.

[9] Mats Rooth. A theory of focus interpretations. *Natural Language Semantics*, 1:75–116, 1992.

[10] Mats E. Rooth. *Association with Focus.* PhD thesis, University of Massachusetts, 1985.

[11] Volker Strom. Die Prosodiekomponente in NTARC I.2: Satzmodusbestimmung aus der F0. Verbmobil:Technisches Dokument 6, IKP Universität Bonn, 1994.

[12] Volker Strom. Detection of accents, phrase boundaries and sentence modality in German with prosodic features. In *Eurospeech 1995*, pages 2039–2041, 1995.

[13] Wolfgang Wahlster. Verbmobil: Übersetzung von Verhandlungsdialogen. Verbmobil-Report 1, DFKI, Saarbrücken, 1993.

[14] Hans Weber. *LR-inkrementelles, probabilistisches Chartparsing von Worthypothesenmengen mit Unifikationsgrammatiken: Eine enge Kopplung von Suche und Analyse.* PhD thesis, Universität Hamburg, Department of Computer Science, 1995.