M = Syntax + Prosody: a syntactic–prosodic labelling scheme for large spontaneous speech databases

Anton Batliner, Ralf Kompe, Andreas Kießling, Marion Mast, Heinrich Niemann, Elmar Nöth

Angaben zur Veröffentlichung / Publication details:

Batliner, Anton, Ralf Kompe, Andreas Kießling, Marion Mast, Heinrich Niemann, and Elmar Nöth. 1998. "M = Syntax + Prosody: a syntactic–prosodic labelling scheme for large spontaneous speech databases." *Speech Communication* 25 (4): 193–222. https://doi.org/10.1016/S0167-6393(98)00037-5.

M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases

A. Batliner [a,*], R. Kompe [a,b], A. Kießling [a,c], M. Mast [a,d], H. Niemann [a], E. Nöth [a]

[a] *Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung, Informatik 5, Martensstr. 3, 91058 Erlangen, Germany*
[b] *Sony Stuttgart Technology Center, Stuttgarterstr. 106, 70736 Fellbach, Germany*
[c] *Ericsson Eurolab Deutschland GmbH, Nordostpark 12, 90411 Nürnberg, Germany*
[d] *IBM Heidelberg Scientific Center, Vangerowstr. 18, 69115 Heidelberg, Germany*

Abstract

In automatic speech understanding, division of continuous running speech into syntactic chunks is a great problem. Syntactic boundaries are often marked by prosodic means. For the training of statistical models for prosodic boundaries large databases are necessary. For the German Verbmobil (VM) project (automatic speech-to-speech translation), we developed a syntactic–prosodic labelling scheme where different types of syntactic boundaries are labelled for a large spontaneous speech corpus. This labelling scheme is presented and compared with other labelling schemes for perceptual–prosodic, syntactic, and dialogue act boundaries. Interlabeller consistencies and estimation of effort needed are discussed. We compare the results of classifiers (multi-layer perceptrons (MLPs) and *n*-gram language models) trained on these syntactic–prosodic boundary labels with classifiers trained on perceptual–prosodic and pure syntactic labels. The main advantage of the rough syntactic–prosodic labels presented in this paper is that large amounts of data can be labelled with relatively little effort. The classifiers trained with these labels turned out to be superior with respect to purely prosodic or syntactic labelling schemes, yielding recognition rates of up to 96% for the two-class-problem 'boundary versus no boundary'. The use of boundary information leads to a marked improvement in the syntactic processing of the VM system.

Zusammenfassung

Die Segmentierung von kontinuierlich gesprochener Sprache in syntaktisch sinnvolle Einheiten ist für die automatische Sprachverarbeitung ein großes Problem. Syntaktische Grenzen sind oft prosodisch markiert. Um prosodische Grenzen mit statistischen Modellen bestimmen zu können, benötigt man allerdings große Trainingskorpora. Für das Forschungsprojekt Verbmobil zur automatischen Übersetzung spontaner Sprache wurde daher ein syntaktisch–prosodisches Annotationsschema entwickelt und auf ein großes Korpus angewendet. Dieses Schema wird mit anderen Annotationsschemata verglichen, mit denen prosodisch–perzeptive, rein syntaktische bzw. Dialogakt-Grenzen etikettiert wurden; Konsistenz der Annotation und benötigter Aufwand werden diskutiert. Das Ergebnis einer automatischen Klassifikation (multi-layer perceptrons bzw. Sprachmodelle) für diese neuen Grenzen wird mit den Erkennungsraten verglichen, die für die anderen Grenzen erzielt wurden. Der Hauptvorteil der groben syntaktisch–prosodischen Grenzen, die in diesem Aufsatz eingeführt werden, besteht darin, daß ein großes Trainingskorpus in

[*] Corresponding author. E-mail: batliner@informatik.uni-erlangen.de.

relativ kurzer Zeit erstellt werden kann. Die Klassifikatoren, die mit diesem Korpus trainiert wurden, erzielten bessere Ergebnisse als alle früher verwendeten; die beste Erkennungsrate lag bei 96% für das Zwei-Klassen-Problem 'Grenze vs. Nicht-Grenze'. Die Berücksichtigung der Grenzinformation in der syntaktischen Verarbeitung führt zu einer wesentlichen Verbesserung.

Résumé

En compréhension automatique de la parole, la segmentation de parole continue en composants syntaxiques pose un grand problème. Ces composants sont souvent délimitées par des indices prosodiques. Cependant l'entraînement de modèles d'étiquetage de frontières prosodiques statistiques nécessite de très grandes bases de données. Dans le cadre du projet allemand Verbmobil (traduction automatique de parole à parole), nous avons donc développé une méthode d'étiquetage prosodique–syntaxique de larges corpus de parole spontanée où deux principaux types de frontières (frontières syntaxiques majeures et frontières syntaxiques ambigües) et certaines autres frontières spéciales sont étiquetés. Cette méthode d'étiquetage est présentée et comparée à d'autres méthodes d'étiquetage de frontières basées sur des critères prosodiques perceptifs, syntaxiques, et de dynamique du dialogue. Plus précisement, nous comparons les résultats obtenus à partir de classificateurs (perceptrons multi-couches et modèles du language) entraînés sur les frontières établies par notre étiqueteur prosodique–syntaxique à ceux obtenus à partir d'étiqueteurs strictement syntaxiques ou prosodiques perceptifs. L'un des avantages principaux de la méthode d'étiquetage prosodique–syntaxique présentée dans cet article est la rapidité avec laquelle elle permet d'étiqueter de grandes bases de données. De plus, les classificateurs entraînés avec les étiquettes de frontière produites se révèlent être très performants, les taux de reconnaissance atteignant 96%.

Keywords: Syntax; Dialogue; Prosody; Phrase boundaries; Prosodic labelling; Automatic boundary classification; Spontaneous speech; Large databases; Neural networks; Stochastic language models

1. Introduction

The research presented in this paper has been conducted under the Verbmobil (VM) project, cf. (Wahlster, 1993; Bub and Schwinn, 1996; Wahlster et al., 1997), which aims at automatic speech-to-speech translation of appointment scheduling dialogues; details are given in Section 2. In VM, prosodic boundaries are used for disambiguation during syntactic parsing. In spontaneous speech, many elliptic sentences or nonsentential free elements occur. Without knowledge of the prosodic phrasing and/or the dialogue history, a correct syntactic phrasing that mirrors the intention of the speaker is often not possible for a parser in such cases. Consider the following turn – a typical example taken from the VM corpora:

ja | zur Not | geht's | auch | am Samstag|

The vertical bars indicate possible positions for clause boundaries. In written language, most of these bars can be replaced by either comma, period or question mark. In total, there exist at least 36 different syntactically correct alternatives for putting the punctuation marks. Examples 1 and 2 show two of these alternatives together with a translation into English.

Example. 1. *Ja? Zur Not geht's? Auch am Samstag?* (Really? It's possible if necessary? Even on Saturday?)

Example. 2. *Ja. Zur Not. Geht's auch am Samstag?* (Yes. If necessary. Would Saturday be possible as well?)

Without knowledge of context and/or dialogue history, for such syntactically ambiguous turns, the use of prosodic information might be the only way to find the correct interpretation, e.g., whether *ja* is a question or a confirmation. But even if knowledge of the context is available, we believe that it is much cheaper and easier to infer the same information from prosody than from the context. Furthermore, for syntactically nonambiguous word sequences, the search space during parsing can be enormous because locally, it is often not possible to decide for some word boundaries if there is a clause boundary or not. Therefore, the search effort can be reduced considerably during parsing if prosodic information about clause boundaries is available, cf. (Batliner et al., 1996a; Kompe et al., 1997).

Researchers have long noted that there is a strong (albeit not perfect) correspondence between prosodic phrasing and syntactic phrasing, cf. (Lea, 1980; Vaissière, 1988; Price et al., 1991; Cutler et al., 1997). Most of the pertinent research on this topic that has been conducted in the last two decades is basic research with controlled, elicited speech, a small database, and only a few speakers. This restriction holds even for most of the studies that definitely aim at the use of prosody in automatic speech processing, cf., e.g. (Bear and Price, 1990; Ostendorf and Veilleux, 1994; Hunt, 1994), who used pairs of ambiguous sentences read by professional radio news speakers, and (Wang and Hirschberg, 1992), which is one of the very few studies based on spontaneous speech (a subset of the ATIS database (Lamel, 1992)). For a review of the different statistical methods used in these studies, see (Kompe, 1997). In these studies with (American) English, the labelling is generally based on the ToBI approach (Beckman and Ayers, 1994), in which tone sequences are annotated on the tonal tier and a hierarchy of break indices on the break (phrasing) tier. There is no general agreement on the number of levels needed. The lowest possible number is of course two: 'boundary versus no boundary'. ToBI has four levels, and seven levels can be found in (Ostendorf et al., 1990; Price et al., 1990; Bear and Price, 1990). Usually, these detailed schemes are, however, in practice reduced to some two or three levels.

In this paper, we describe scenario and architecture of the VM project in Section 2. In Section 3 we discuss briefly the boundary annotations used up to now in VM: The prosodic–perceptual B boundaries, the purely syntactic S boundaries, and the dialogue act boundaries D. We then introduce in Section 4 a new labelling scheme, the syntactic–prosodic M boundaries, that constitutes a first step towards an integrated labelling system especially suited for the processing of large spontaneous speech databases used for automatic speech processing. With these labels, we were able to annotate large amounts of data with relatively little effort without much loss of information which is essential for the exploitation of prosody in the interpretation of utterances. The correspondence between these M labels and the other three labelling schemes (Section 5) corroborates this assumption; its discussion leads to a more detailed relabelling of the M boundaries described in Section 6. Correspondences of the new M boundaries with B and D are dealt with in Section 7, interlabeller consistency and an estimation of overall effort for the annotation of the M labels in comparison with other annotation schemes is given in Section 8. Finally, we discuss recognition rates obtained with a combination of acoustic classifiers and language models (Section 9) as well as the overall usefulness of the M labels for syntactic processing in VM (Section 10).

2. The Verbmobil project

VM is a speech-to-speech translation project (Wahlster, 1993; Wahlster et al., 1997; Block, 1997) in the domain of appointment scheduling dialogues, i.e., two persons try to fix a meeting date, time, and place. Currently, the emphasis lies on the translation of German utterances into English. VM deals thus with a special variety of spontaneous speech found in such human–machine–human communications which is, however, fairly representative for spontaneous speech in general, as far as typical spontaneous speech phenomena (word fragments, false starts, syntactic constructions not found in written language, etc.) are concerned. After the recording of the spontaneous utterance, a word hypotheses graph (WHG) is computed by a standard Hidden Markov Model (HMM) word recognizer and enriched with prosodic information. This information currently consists of probabilities for a major syntactic boundary being after the word hypotheses, a probability for accentuation and for three classes of sentence mood (Kießling, 1997; Kompe, 1997). The WHG is parsed by one of two alternative syntactic modules, and the best scored word chain together with its different possible parse trees (readings), cf. Section 10, is passed onto the semantic analysis. Governed by the syntax module and the dialogue module, the utterance is translated onto the semantic level (transfer module), and an English utterance is generated and synthesized. Parallel to the *deep* analysis performed by syntax and semantics, the dialogue module conducts a *shallow* processing, i.e., the important dialogue acts are detected in the utterance and translated roughly. A more detailed account of the architecture can be found in (Bub and Schwinn, 1996). For the time being, the following modules use prosodic information: syntactic analysis, semantics, transfer, dialogue processing, and speech synthesis (cf. Niemann et al., 1997); in this paper, we will only deal with the use of prosody in syntactic analysis and dialogue processing.

The architecture of VM is a so-called multi-agent architecture which means that several autonomous modules (developed by 29 different partners at different sites) process speech and language data and interact with each other. Note that VM uses a sequential, bottom–up approach; syntax and dialogue do not interact with each other, there is no dialogue history available for the syntax module and no deep syntactic analysis for the dialogue module. A system like VM can thus not mirror perception and understanding of a competent native speaker that is certainly a combination of bottom–up and top–down processes and much more parallel than the architecture of VM allows, or, for that matter, the architecture of any other successful system that can be imagined in the near future. Prosodic analysis has to be adapted to the demands of those modules (syntax and dialogue) that use its information. This means that we use prosodic information to predict events that can be useful for processing in these higher modules. We conceive prosody as a means to convey different functions, e.g., segmentation of speech into syntactic boundaries or dialogue act boundaries. Other functions of prosody (denoting rhythm, speaker idiosyncrasies, emotions, etc.) are treated in this approach either as intervening factors that have to be controlled, if their influence is strong and systematic, or as random noise that can be handled quite well with statistical classifiers, if their influence is weak and/or unsystematic.

For all VM dialogues, a so-called ‘basis transliteration’ is provided, i.e., the words in orthography together with non-words, as, e.g., pauses, breathing, undefined noise, etc., and with specific comments, if, e.g., the pronunciation of a word deviates considerably from the canonical pronunciation (slurring, dialectal variants). A phonetic transcription exists only for a small number of dialogues and is used for special tasks. Irregular phenomena, e.g., boundaries at speech repairs, are annotated as well.

In a second step, the basis transliteration is supplemented by other partners with annotations needed for special tasks, e.g., perceptual–prosodic labels for accentuation and phrasing, and dialogue act annotations including dialogue act boundaries, cf. Section 3. The resources for these time–consuming annotations are limited, and often, the databases available are too small for a robust training of statistical classifiers. This in turn prevents an improvement of automatic classification. This point will be taken up again in the discussion of the different labelling schemes in Section 3.

3. Prosodic, syntactic, and dialogue act boundaries: Bs, Ss and Ds

In written forms of languages such as German and English, syntactic phrasing is – on the surface – at least partly indicated by word order; for instance, a wh-word after an infinite verb form normally indicates a syntactic boundary before the wh-word: *Wir können gehen wer kommt mit* (We can go. Who will join us?). Above that, syntactic phrasing in written language can be disambiguated with the help of punctuation marks. In spontaneous speech, prosodic marking of boundaries can take over the role of punctuation. In order to use prosodic boundaries during syntactic analysis, automatic classifiers have to be trained; for this, prosodic reference labels are needed.

Reyelt and Batliner (1994) describe an inventory of prosodic labels for the speech data in VM along the lines of ToBI, an inventory that contains a boundary tier amongst other tiers. The following different types of perceptual boundaries were labelled by the VM partner University of Braunschweig, cf. (Reyelt, 1998).

B3: full intonational boundary with strong intonational marking with/without lengthening or change in speech tempo.

B2: minor (intermediate) phrase boundary with rather weak intonational marking.

B0: normal word boundary (default, not labelled explicitly).

B9: 'agrammatical' boundaries indicating hesitations or repairs.

There are, however, some drawbacks in this approach if one wants to use this information in parsing: First, prosodic labelling by hand is very time consuming; the labelled database up to now is therefore rather small. Second, perceptual labelling of prosodic boundaries is not an easy task and possibly not very robust. Effort needed and consistency of annotation will be discussed in Section 8. Finally and most important, prosodic boundaries do not only mirror syntactic boundaries but are influenced by other factors such as rhythmic constraints and speaker specific style. In the worst case, discrepancies between prosodic and syntactic phrasing might be lethal for a syntactic analysis if the parser goes on the wrong track and never returns.

Feldhaus and Kiss (1995) therefore argued for a labelling without listening which is based solely on linguistic definitions of syntactic phrase boundaries in German (cf. as well (Batliner et al., 1996a)). They proposed that only syntactic boundaries should be labelled, and they should be labelled whether they are marked prosodically or not. (The assumption behind is, of course, that most of the time, a correspondence between syntactic and prosodic boundaries really exists. Otherwise, prosodic classification of such boundaries would be rather useless for parsing.) 21 dialogues of the VM corpus that are labelled with Bs were labelled by the VM partner IBM Heidelberg using a very detailed and precise scheme with 59 different types of syntactic boundaries that are described in (Feldhaus and Kiss, 1995). The labels code the left context and the right context of every word boundary. The distribution of the 59 different types is of course very unequal and often, only a few tokens per type exist in the database. In this paper, we therefore use only three main labels, S3+ (syntactic clause boundary obligatory), S3− (syntactic clause boundary impossible), and S3? (syntactic clause boundary ambiguous). Note that S3? is sometimes used for constellations that could not be analyzed syntactically at all or where the labeller could not decide between two competing analyses. The correspondence between S and B labels showed a high agreement of 95%, cf. (Batliner et al., 1996a).

In VM, the dialogue as a whole is seen as a sequence of basic units, the dialogue acts. One task of the dialogue module is to follow the dialogue in order to provide predictions for the ongoing dialogue. Therefore each turn has to be segmented into units which correspond to one dialogue act, i.e. the boundaries – in the following called D3 in analogy to B3 and S3 – have to be detected and for each unit, the corresponding dialogue act(s) has (have) to be found. For the training of classifiers, a large subsample of the VM database has been labelled with D3 at dialogue act boundaries; every other word boundary is automatically labelled with D0. Dialogue acts are defined according to their *illocutionary force*, e.g., ACCEPT, SUGGEST, REQUEST, and can be subcategorised for their *functional role* in the dialogue or for their *propositional content*, e.g., DATE or LOCATION depending on the application; in our domain, 18 dialogue acts on the illocutionary level and 42 subcategories are defined at present (Jekat et al., 1995). Dialogue acts on the illocutionary (functional) level are e.g. REJECT and CLARIFICATION. REJECT is subcategorised w.r.t. the propositional context in the appointment scheduling domain in REJECT_LOCATION and REJECT_TIME. Clarification is subcategorised w.r.t. the function the dialogue act has in the dialogue in CLARIFICATION_ANSWER (when it follows a question) and CLARIFICATION_STATE (when there precedes no question, but the speaker wants to clarify something with a statement). With this multi-level definition of dialogue acts, the ones on the illocutionary level can be applied to other domains. In VM, dialogue acts are used in different modules for quite different purposes (e.g. tracking the dialogue history, robust translation with keywords) for which the more domain specific dialogue acts on the lower levels are needed.

Turns can be subdivided into smaller units where each unit corresponds to one or more dialogue acts. It is not easy to give a quantitative and qualitative definition of the term dialogue act in spontaneous speech. We defined criteria for the segmentation of turns based on the textual representation of the dialogue (Mast et al., 1995). Such criteria are mostly syntactic, e.g.: All ‘material’ that belongs to the verb frame of a finite verb belongs to the same dialogue act. That way it is guaranteed that both the obligatory and the optional elements of a verb are included in the same dialogue act, cf. Example 3. For dependent clauses the preceding rule is also applicable: Each dependent clause which contains a finite verb is seen as a unit of its own, cf. Example 4. Conventionalised expressions are seen as one unit even if they do not contain a verb. Typical examples are: *hello, good morning, thanks*. Prosody is not taken into account in order to be able to label dialogues without having to listen to them and thus to reduce the labelling effort (cf. Section 8): In (Carletta et al., 1997, p. 35) it is reported that the segmentation of dialogues changes only slightly when the annotators can listen to speech data.

Example 3.

and on the fourteenth I am leaving for my bobsledding vacation until the nineteenth	SUGGEST_EXCLUDE_DATE.

Example 4.

no Friday is not any good	REJECT_DATE
because I have got a seminar all day	GIVE_REASON

Large-scaled parallel annotations of B, S and D boundaries might be desirable; it is, however, more realistic to aim at smaller reference databases with annotations of different boundaries in combination with a large database annotated with an integrated labelling system. ‘Integrated’ in this context means that we want to favour a labelling system that basically is syntactic but takes expectations about prosodic and dialogue structure into consideration by subclassifying boundaries; by this, the subclasses can be clustered differently according either to the needs of syntax or to the needs of dialogue analysis. If such labels can be annotated in a relatively short amount of time, they can be annotated for a very large corpus for the training of automatic classifiers without too much effort. Other annotations, e.g., prosodic or dialogue act boundaries, can be used to evaluate such a labelling system. For that, however, smaller reference corpora can be used.

4. A rough syntactic–prosodic labelling scheme: the Ms

Prosodic boundaries do not mirror syntactic boundaries exactly, but prosodic marking can, on the other hand, be the only way to disambiguate the syntactic structure of speech. In the past, we successfully incorporated syntactic–prosodic boundary labels in a context-free grammar which was used to generate a large database with read speech; in this grammar, 36 sentence templates were used to generate automatically 10.000 unique sentences out of the time-table-inquiry-domain. We added boundary labels to the grammar and thus to the generation process. The sentences the speakers had to read contained punctuation marks but no prosodic labels. In listening experiments, boundaries were defined perceptually and used later as reference in automatic classification experiments. For these perceptually defined boundaries, recognition rates of above 90% could be obtained with acoustic–prosodic classifiers trained on automatically generated boundary labels; for details, cf. (Kompe et al., 1995; Batliner et al., 1995; Kompe, 1997).

For read, constructed speech, it is thus possible to label syntactic–prosodic boundaries automatically; for spontaneous speech, however, it is – at least for the time being – necessary to label such boundaries manually. These labels shall be used both for acoustic–prosodic classifiers and for stochastic language models (prosodic–syntactic classifiers). The requirements for such a labelling are described in the following.

First, it should allow for fast labelling. The labelling scheme should be rather rough and not based on a deep syntactic analysis because the more precise it is the more complicated and the more time consuming the labelling will be; rough in this context means considerably rougher than the annotation with S labels. A 'small' amount of labelling errors can be tolerated, since it shall be used to train statistical models, which should be robust to cope for these errors.

Second, prosodic tendencies and regularities should be taken into account. For our purposes, it is suboptimal to label a syntactic boundary that is most of the time not marked prosodically with the same label as an often prosodically marked boundary for the following reasons: Syntactic labels that – implicitly – model the words and/or parts-of-speech before and after the boundary can be used for language models. Prosodic–perceptual labels can be used for acoustic–prosodic classifiers irrespective of the syntactic context. Labels that are used both for language models and for acoustic–prosodic classifiers have to be subclassifed accordingly; examples for such subclassifications are given below. Since large quantities of data should be labelled within a short time, only expectations about prosodic regularities based on the textual representation of a turn can be considered. These expectations are either based on the experience of the labeller or rely on the basis transliteration of the VM dialogues where, e.g., pauses that are longer than half a second are annotated; examples will be given below. For the same reason, we will not use the whole dialogue history for interpretation and disambiguation, but only the immediate context, i.e., the whole turn or at the most the turns before and after. Pauses and/or breathing that are labelled in the transliteration will be taken as an indication of a prosodic boundary and used for a subclassification of our syntactic boundaries. Note that pauses – and hesitations, for that matter – are often an indication of agrammatical boundaries which, however, are already labelled in the basis transliteration.

Third, the specific characteristics of spontaneous speech (elliptic sentences, frequent use of discourse particles, etc.) have to be taken into account.

The labels will be used for statistical models (hence M for the first character), corresponding to B, D and S. The strength of the boundary is indicated by the second character: 3 at sentences/clauses/phrases, 2 at prosodically heavy constituents, 1 at prosodically weak constituents, 0 at any other word boundary. The third character tries to code the type of the adjacent clauses or phrases, as described below. In Table 1, the context of the boundaries is described shortly, and the label and the main class it is attached to is given, as well as one example for each boundary type; in addition, the frequency of occurrence in the whole database is given as well, not counting the end of turns. These, by default, are labelled implicitly with M3. So far a reliable detection of M3 had priority; therefore, for the time being, M2I is only labelled in three dialogues and mapped onto M0 for our classification experiments; M1I is currently not labelled at all.

Agrammatical phenomena such as hesitations, repairs and restarts, are labelled in the basis transliteration, cf. (Kohler et al., 1994), and were also used for disambiguating between alternative M labels. However, in very agrammatical passages, a reasonable labelling with M labels is almost impossible. In general, we follow the strategy that after short agrammatical passages, no label is given but after rather long passages, especially if the syntactic construction starts anew, either M3S or M3P is labelled; a more detailed discussion can be found in (Batliner et al., 1996b).

Syntactic main boundaries M3S ('S'entence) are annotated between main clause and main clause, between main clause and subordinate clause, and before coordinating particles between clauses. Boundaries at 'nonsentential free elements' functioning as elliptic sentences are labelled with M3P ('P'hrase), as well as left dislocations (cf. Section 6). Normally, these phrases do not contain a verb. They are idiomatic performative phrases with a sort of lexicalized meaning such as *Guten Tag* (hello), *Wiedersehen* (good bye) and vocatives, or they are 'normal, productive' elliptic sentences such as *um vierzehn Uhr* (at two p.m.). Boundaries between a sentence and a phrase to its right, which in written language normally would be inside the verbal brace, are labelled with M3E ('E' for ex-

Table 1
Description and examples for boundary labels and their main classes in parentheses, with frequency of occurence of the annotated labels in the whole database (326 dialogues, 7075 turns, 147 110 words)

Label	Main class	#	Description with example
M3S	(M3)	11 473	main/subordinate clause: *vielleicht stelle ich mich kurz vorher noch vor M3S mein Name ist Lerch* perhaps I should first introduce myself M3S my name is Lerch
M3P	(M3)	4535	nonsentential free element/phrase, elliptic sentence, left dislocation: *guten Tag M3P Herr Meier* hello M3P Mr. Meier
M3E	(M3)	1398	extraposition: *wie würde es Ihnen denn am Dienstag passen M3E den achten Juni* will Tuesday suit you M3E the eighth of June
M3I	(M3)	369	embedded sentence/phrase: *eventuell M3I wenn Sie noch mehr Zeit haben M3I ⟨Atmung⟩ 'n bißchen länger* possibly M3I if you've got even more time ⟨breathing⟩ M3I a bit longer
M3T	(M3)	325	pre-/ postsentential particle with ⟨pause⟩/⟨breathing⟩: *gut M3T ⟨Pause⟩ okay* fine ⟨pause⟩ M3T okay
M3D	(MU)	5052	pre-/ postsentential particle without ⟨pause⟩/⟨breathing⟩: *also M3D dienstags paßt es Ihnen M3D ja M3S ⟨Atmung⟩* then M3D Tuesday will suit you M3D won't it / after all M3S ⟨breathing⟩
M3A	(MU)	707	syntactically ambiguous: *würde ich vorschlagen M3A vielleicht M3A im Dezember M3A noch mal M3A dann* I'd propose M3A possibly M3A in December M3A again M3A then
M2I	(M0)	–	constituent, marked prosodically: *wie sähe es denn M2I bei Ihnen M2I Anfang November aus* will it be possible M2I for you M2I early in November
M1I	(M0)	–	constituent, not marked prosodically: *M3S hätten Sie da M1I 'ne Idee M3S* M3s have you've got M1I any idea M3S
M0I	(M0)		every other word (default)

traposition, or right dislocation with or without a pro element): In *Dann können wir es machen* M3E *das Treffen* (Then we can do that, the meeting), there is a pro element (*es* = it), whereas in *Wurde es Ihnen passen* M3E *am Dienstag* (Will it suit you on Tuesday), no pro element exists. In written language, the dislocated element would be inside the verbal brace: *Dann können wir das Treffen machen* and *Würde es Ihnen am Dienstag passen*. Note that the verbal brace ('Verbklammer') is a syntactic phenomenon that does not exist in English. M3E is also labelled at boundaries where there is no verbal brace (so-called 'open verbal brace') and thus no defining criterion, but where a pause etc. in the transliteration denotes a stronger separation from the clause to its left, e.g. in *Treffen wir uns* M3E *<pause> am Freitag* (Let's meet M3E <pause> on Friday). This difference can influence presuppositions and thus semantic interpretation because a clear pause that can be accompanied by a pronounced accent on the extrapolated element indicates an – implicit – contrast, if, e.g., the dialogue partner has proposed Saturday, and the speaker wants to reject this day not explicitely (*Nein, nicht am Samstag, sondern am Freitag* (No, not on Saturday, but on Friday)) but with the help of syntactic (and prosodic) means (extraposition with or without special accentuation).

Sentences or nonsentential free elements that are embedded in a sentence are labelled with M3I ('I'nternal). Typically, these are parenthetical asides or embedded relative clauses.

Very often in spontaneous speech, a turn or a sentence inside a turn begins with presentential particles ('Satzauftaktpartikeln') such as *ja* (well), *also* (well), *gut* (well), *okay* (okay), etc. The term 'presentential particle' is used here purely syntactically for a particle that is the first word in a turn or in a sentence. Such particles can have different functions. Often, it cannot be decided whether they are just discourse particles or whether they have a certain meaning: Their functions are neutralized. Often, but not always, prosody can help to disambiguate these different functions. Note that normally, only affirmative but not negative particles can be neutralized in presentential position, cf. the following four answers to the question *Kommst du morgen* (Will you come tomorrow): (1) confirmation, semantics of particle neutralized: *Ja das geht* (Yes/Well, that's possible), (2) rejection, presentential discourse particle without semantic function: *Ja das geht überhaupt nicht* (Well, that's not possible at all), (3) rejection: *Nein das geht nicht* (No, that's not possible), (4) ungrammatical and contradictory combination of negative particle and affirmation: *Nein das geht* (No, that's possible). The specific function can, however, be marked by prosodic means; presentential particles that are followed by a pause or by breathing denoted as such in the transliteration are therefore labelled with M3T and all other with M3D ('D'iscourse particle). In postsentential position, we label these particles analogously. Here, they normally function as tags: *Geht gut ja* (That's ok isn't it). Note that inside a clause, they are modal particles that normally cannot be translated into English: *Das ist ja gut* (That's great). (The mnemonic reason for M3T versus M3D is that M3T represents a stronger boundary than M3D because it is marked by a pause, and the phoneme /t/ is phonologically/phonetically stronger than /d/ as well, cf. (Grammont, 1923).) Note that for a correct interpretation and translation, a much finer classification than the one between M3D and M3T should distinguish between lexical categories and their possible syntax/dialogue specific roles.

Syntactically ambiguous boundaries M3A ('A'mbiguous) cannot be determined only based on syntactic criteria. Often there are two or more alternative word boundaries where the syntactic boundary could be placed. It is thus the job of prosody to disambiguate between two alternative readings. M3A and M3D labels are mapped onto the main class MU ('undefined'), all other labels mentioned so far are mapped onto the main class M3 ('strong boundary').

The labels M2I and M1I denote ('I'nternal) syntactic constituent boundaries within a sentence (typically NPs or PPs) and are mapped onto the main class M0, together with the default class M0I, which implicitly is labelled at each word boundary (except for turn–final ones) where none of the above labels is placed. An M1I constituent boundary is in the vicinity of the beginning or the end of a clause; it is normally not marked prosodically because of rhythmic constraints. An M2I constituent boundary is inside a clause or phrase, not in the vicinity of beginning or end of the turn; it is rather often marked prosodically, again because of rhythmic constraints. In the experiments conducted so far, we distinguish only between the three main classes M3, MU and M0 given in Table 1; these are for the time being most relevant for the linguistic analysis in VM. Besides, M3 and M0 are 'robust' in the sense that their assignment is less equivocal and thus less prone to different syntactic interpretations or misconceptions by different labellers than it might be the case for the other, more detailed labels, cf. Section 8. We believe, however, that these detailed labels might be relevant in the future. A more specific account of all labels can be found in (Batliner et al., 1996b).

With the labelling schemes described so far, different sub-corpora were labelled whose size is given in Table 2. The following corpora are used in this paper.

- TEST is usually used for the test of classifiers, cf. below. It was spoken by six different speakers

Table 2
The different subsets used

	# Dialogues	# Turns	Minutes	# Word tokens
TEST	3	64	11	1513
B-TRAIN	30	797	96	13 145
M-TRAIN	293	6214	869	132 452
S-TRAIN	21	583	66	8222

(three male, three female). In all other corpora about one third of the speakers are female.

- B-TRAIN contains all turns annotated by the VM partner University of Braunschweig with prosodic labels except the ones contained in TEST.
- M-TRAIN are all turns labelled with the M labels except the ones in TEST and B-TRAIN for which M labels are available as well.
- S-TRAIN is the subset of B-TRAIN for which syntactic labels were created by our colleagues from IBM.
- D-TRAIN is the sub-corpus for which D labels are available; this corpus is constantly growing, cf. Section 8.

The M labels of TEST were checked several times by the labeller; the labels of all other corpora were not checked thoroughly. To give an impression of the effort needed: The S labelling was done by one linguist at IBM in about two months, the M labelling by the first author in about four months, i.e., for the M labels, the effort is reduced almost by the factor 10 (twice the amount of time for a material that is 17 times larger). Note, however, that these figures are only rough estimates. We will come back to this topic in Section 8 below.

5. Correspondence between Ms and Ss, Ds and Bs

We will generally refer to S-TRAIN if we discuss correspondences between labels because for this sub-corpus, all four types of boundaries are available: B labels that denote prosodic boundaries as perceived by human listeners, S labels that denote (detailed) syntactic boundaries, D labels that denote dialogue act boundaries, and M labels that denote rough syntactic–prosodic boundaries.

Tables 3–11 give the figures of correspondence. In these tables, the second column shows the frequencies of the labels given in the first column. All other numbers show the percentage of the labels in the first column coinciding with the labels in the first row. For example, the number 84.3 in the third column, second row of Table 3 means that 84.3% of the word boundaries labelled with M3 are also labelled with S3+. Those numbers, where from the definition of the labels a high correspondence could have been expected a priori, are given in bold face: We expect high correspondences between the boundaries M3, B3, S3+ and D3, and also high correspondences between the nonboundaries M0, B0, S3− and D0. Note that turn final word boundaries are not considered in the tables, because these are in all cases labelled with S3, M3 and D3, and in most cases with B3. As an almost total correspondence between the labels at these turn final positions is obvious, their inclusion would yield very high but not realistic correspondences.

There are at least three different factors that can be responsible for missing correspondences between the different label types.

- *Labelling errors:* Simple labelling errors occur once in a while, in particular if the labelling has to be done rather fast. Such errors are typically omitting a label or shifting its position one word to the left or to the right of the correct position. To give an exact figure for these errors is not possible because this would imply a very exact and time consuming check of the labels – which is exactly what we want to avoid. Such errors should be randomly distributed and thus no problem for statistical classifiers if they do not occur very often.
- *Systematic factors:* Either the clustering is too rough and we had to use a more detailed scheme, or there is no systematic relationship between two types of labels.
- *Interlabeller differences:* These exist of course not only for labellers using the same scheme but for labellers who use different schemes; they are either randomly distributed or caused by systematic differences in analyzing the structure of the turn. While within the same scheme, interlabeller consistency can be investigated, cf. Section 8, across schemes, this is normally not possible.

5.1. A comparison of M *and* S *labels*

Of primary interest is the degree of correspondence between the M and the S labels, because the latter were based on a thorough syntactic analysis (deep linguistic analysis with syntactic categorization of left and right context for each label) while for the M labels, we used a rough scheme and mainly took into account the left context.

Tables 3–5 show that there is a very high correspondence between M0 and S3–: 96.6% of the M0 correspond to S3– and 98.2% of the S3– to M0. 84.4% of the M3 are also labelled as S3+. Only 67.9% of the S3+ labels correspond to M3; most of the remaining S3+ (26.2%) are labelled as MU. A closer look at the subclasses of the M labels shows that this is not due to labelling errors which can be found in both annotations, but that it has systematic reasons resulting from different syntactic–theoretic assumptions. Mainly responsible for this mismatch is that the majority of the M3D labels, a subclass of MU, is labelled with S3+. Examples 5 and 6 show further typical parts of turns, where S3+ systematically correspond to M0. Neither the one or the other labelling system is wrong, but they use different syntactic analyses resulting in different labels: In Example 5, the time of day expression can be considered to be just an object to the verb (no boundary), or a sort of elliptic subordinate clause (boundary); the conjunction in Example 6 can either be attributed to the second clause (no boundary) or neither to the first or to the second clause (boundary).

Table 3
Percentage of M labels corresponding to S labels

Label	#	S3+	S3?	S3–
M3	951	**84.3**	8.4	7.2
MU	391	79.3	9.2	11.5
M0	6297	1.1	2.3	**96.6**

Table 4
Percentage of S labels corresponding to M labels

Label	#	M3	MU	M0
S3+	1181	**67.9**	26.2	5.8
S3?	259	30.9	13.9	55.2
S3–	6199	1.1	0.7	**98.2**

Table 5
Percentage of detailed M labels corresponding to S labels

Label	#	S3+	S3?	S3–
M3S	502	**94.2**	0.6	5.2
M3P	288	**93.4**	3.8	2.8
M3E	148	**34.5**	43.2	22.3
M3I	6	**50.0**	33.3	16.7
M3T	7	**85.7**	0.0	14.3
M3D	301	90.0	3.6	6.3
M3A	90	43.3	27.8	28.9
M0	6297	1.1	2.3	**96.6**

Example 5. *sagen wir lieber* M0/S3+ *vierzehn Uhr fünfundzwanzig* (let's rather say M0/S3+ two twenty five p.m.).

Example 6. *aber* M0/S3+ *Donnerstag vormittag…wär' mir recht* (but M0/S3+ Thursday in the morning…would be fine)

Only 34.5% of the M3E labels, a subclass of M3, correspond to S3+. This might partly be due to the fact that we took into account pauses denoted in the transliterations while labelling this class: for these positions, a pause after the word boundary triggered the assignment of an M3E label. (Note that without listening to the turns, we cannot decide whether a pause is a 'regular' boundary marker or caused by planning processes; pauses due to deliberation are, however, very often adjacent to hesitations which are denoted in the transliterations. In such cases, the difference is sort of neutralized: Did the speaker pause only because of the interfering planning process or did boundary marking and indication of planning process simply coincide?) Alternatively, it might be due to the fact that extraposition is not a phenomenon everybody agrees on, cf. (Haftka, 1993). In any case, the M3E labels are surely candidates for a possible rearrangement; this was done in the next labelling phase, cf. Section 6. The subclasses M3S and M3P correspond to S3+ in over 90% of the cases. This meets our expectations, because these cases should be quite independent from the specific syntactic analysis. S3? is defined as 'syntactically ambiguous boundary' but at the same time, it is used for boundaries between elements that cannot yet be analyzed syntactically with certainty. M3A is only used for 'contextually' ambiguous boundaries; it is not used for all kinds of possible syntactically ambiguous boundaries but only for those that are not fully impossible in this context; this criterion is admittedly vague. Together with the fact that our rather fast labelling procedure did certainly not reveal all ambiguities these factors might explain the rather low correspondence between S3? and MU; cf. as well Section 8.

5.2. The prosodic marking of the M *labels*

The correspondence between M and B as well as between the different M subclasses and the B labels is given in Tables 6 and 7. The sentence or clause boundaries M3S are mostly (87.8%) marked with a B3 boundary. This corroborates the conventional wisdom that there is a high correspondence between syntactic and prosodic boundaries. However, to our knowledge this is the first investigation of a very large spontaneous speech corpus concerning this hypothesis. It is thus not the very fact but the amount of correlation that is interesting. The 8% of the M3S which correspond to B2 are often boundaries between main clause and subordinate clause, where the speaker has marked the boundary prosodically only by a slight continuation rise. Especially for subordinations, it might be at the discretion of the speaker to what extent prosodic marking is used. The overall speaking rate might play a role as well.

Table 6
Percentage of M labels corresponding to B labels

Label	#	B3	B2	B9	B0
M3	951	**78.7**	9.1	0.1	12.1
MU	391	27.1	29.1	0.5	43.2
M0	6297	2.8	4.6	3.7	**88.9**

Table 7
Percentage of detailed M labels corresponding to B labels

Label	#	B3	B2	B9	B0
M3S	502	**87.8**	8.0	0.0	4.2
M3P	288	**75.7**	10.4	0.3	13.5
M3E	148	**53.4**	11.5	0.0	35.1
M3I	6	**66.7**	0.0	0.0	33.3
M3T	7	**85.7**	0.0	0.0	14.3
M3D	301	24.6	32.9	0.3	42.2
M3A	90	35.5	16.7	1.1	46.7
M0	6297	2.8	4.6	3.7	**88.9**

In Example 7, a clause boundary has not been marked at all prosodically despite the fact that there is no subordinating particle on the syntactic surface. Nevertheless, from the syntax it is clear (in a left to right analysis already at the word *wir*) that there is such a boundary. The first sentence is rather short so that there is no need to separate it prosodically for the purpose of making the listeners understanding easier. Many of these B0/M3S correspondences occur after short main clauses such as *ich denke* (I think) or *meinen Sie* (do you think). These constellations will be taken into account in the relabelling, cf. Section 6.

Example 7. *<Atmung> ich denke* B0/M3S *wir sollten das Ganze dann doch auf die nächste Woche verschieben* (<breathing> I think B0/M3S we should move the whole thing to next week).

Also a high but lower number (75.7%) of the M3P boundaries are marked as B3. This is still within the range of agreements between different persons labelling the B boundaries (Reyelt, 1995). The lower correspondence of the M3P with respect to the M3S can be explained with the fact that M3P labels separate elliptic phrases or left dislocations. These are often quite short so that the same argumentation as above for the short main clauses holds here as well.

Some 35.1% of the M3E are not marked at all prosodically. This might on the one hand indicate that the definition and the labelling of M3E should be revised. On the other hand, we assume that for M3E positions, as well as for other M3 subclasses, it is left at the discretion of the speaker whether these positions are marked prosodically or not, cf. (de Pijper and Sanderman, 1994).

Two thirds of the M3I boundaries are marked prosodically. The M3D labels coincide with B3, B2 or B0, without any clear preference. This could be expected, because the M3D mark ambiguous boundaries at rather short phrases. On the other hand, the defining criterion of M3T, the presence of a pause, is responsible for the very high correspondence (85.7%) with B3.

At positions marked with M3A, the really ambiguous boundary positions between clauses, either a strong boundary marking (B3 in 35.5% of the cases) or no marking at all (B0, 46.7% of the cases) can be observed, which also meets our expectations.

In accordance with their definition, almost all B9 boundaries do not coincide with major syntactic boundaries (M3).

5.3. The difference between D *and* M *labels*

The creation of both M and D labels was rather rough and fast. Despite this, the numbers in Tables 8–10 are consistent with our expectations: Most of the D3 correspond to M3, and almost all of the M0 correspond to D0. Only about half of the M3 correspond to D3, that is, a turn segment corresponding to a dialogue act often consists of more than one clause or phrase – e.g., Example 8 can be segmented into four clauses but only into two dialogue acts. As for the MU labels, not surprisingly, only 3.3% of the M3D (no syntactic boundary and thus normally no D3 boundary) and 20% of the M3A (sometimes a syntactic boundary and as such, sometimes a D3 boundary) coincide with a D3 boundary. An M3P that coincides with a D3 boundary (as in Example 10) will usually be marked prosodically, whereas an M3P that does not coincide with D3 (as in Example 9) will be usually not marked prosodically at all or at least to a lesser extent than M3P in Example 10. M3P in *Guten Tag M3P Herr/Frau...*, e.g., can be assigned to a new subclass of M3P that is assumed not to be marked prosodically and thus to the main class M0 that corresponds to D0. Obviously, M3E does not mark a D3 boundary, cf. the low correspondence of 1.4%. Table 11 shows that 91.4% of the D3 boundaries are strongly marked prosodically, that is, they coincide with a B3 boundary. This number is even higher than that for the M3S boundaries. This confirms the results of other studies which showed that boundaries at discourse units are very strongly marked by prosodic means, cf. (Cahn, 1992; Swerts et al., 1992; Hirschberg and Grosz, 1994).

Table 8
Percentage of M labels corresponding to D labels

Label	#	D3	D0
M3	951	**51.5**	48.5
MU	391	7.2	92.8
M0	6297	0.2	**99.8**

Table 9
Percentage of D labels corresponding to M labels

Label	#	M3	MU	M0
D3	533	**91.9**	5.2	2.8
D0	7106	6.5	5.1	**88.4**

Table 10
Percentage of detailed M labels corresponding to D labels

Label	#	D3	D0
M3S	502	**75.5**	24.5
M3P	288	**37.1**	62.8
M3E	148	**1.4**	98.6
M3I	6	**0.0**	100.0
M3T	7	**28.6**	71.4
M3D	301	3.3	96.7
M3A	90	20.0	80.0
M0	6297	0.2	**99.8**

Table 11
Percentage of D labels corresponding to B labels

Label	#	B3	B2	B9	B0
D3	533	**91.4**	6.2	1.5	0.9
D0	7106	7.7	6.4	3.2	82.7

Example 8. *ich muß sagen* M3S *mir wär's dann lieber* M3S *wenn wir die ganze Sache auf Mai verschieben* D3/M3S <*Pause*> *geht es da bei Ihnen auch* (I would say M3S I then would prefer M3S if we moved the whole thing onto May D3/M3S <pause> does this suit you as well).

Example 9. *Guten Tag* M3P *Herr Meier* ... (Hello M3P Mr. Meier ...).

Example 10. *Guten Tag* D3/M3P *ich hätt 'ne Frage* ... (Hello D3/M3P I've got a question ...).

6. Relabelling of the M labels

The first version of the M labelling scheme was intended for use in the VM prototype which was due in October 1996. The labels were used for the training of automatic classifiers of boundaries as described in Section 9. The time schedule was rather tight, and elaboration and evaluation of the scheme were therefore not conducted for the first version. Even though the M3 boundaries turned out to be very successful, cf. Section 9, we revised and extended the labelling scheme for the following reasons.

- Correspondences to the other boundary types (prosodic–perceptual boundaries and dialogue act boundaries) was good, but in some cases, suboptimal, cf. the discussion in Section 5.
- It turned out that for the higher linguistic modules in VM, a subclassification into more specific classes is desirable.
- To reduce effort, the labelling of prosodic phrases (constituents) inside a sentence was not conducted for the first version. These boundaries are, however, very important for the modelling of accent positions, cf. Section 11.
- Different M3 boundaries were not labelled at the end of turn and adjacent to agrammatical boundaries. Although this is not necessary for our present purposes, such a labelling makes additional information available that might be useful in the future.

The new labels are listed in Tables 12 and 13, where the mapping onto the old labels, the context with one example for each label, the label itself, and the main class it is attached to are given. The names of the new labels consist of three characters each with the encoding given in Table 14. Type and hierarchy describe syntactic phenomena; with strength, we so to speak code our working hypothesis that prosodic (and thereby, to some extent, syntactic) marking of boundaries is scaled along these lines. Most of the revisions concern a subspecification of the former M labels that most of the time could not take into account hierarchical dependencies and left/right relationship. Of course, it will not be possible to model and train all new labels, especially if their frequency is low, but we will have ample possibilities to try and cluster these labels in different ways in order to get the optimal main classes for different demands: The dialogue module will most probably need a clustering that differs from that most useful for the syntax module, cf. Section 7.

The extensional definition of most of the labels did not change, but they will be subspecified. In a few cases, we decided in favour of a more plausible denotation of type with the first character, cf. Tables 12 and 13. The labelling is again introduced in the word chain immediately after the last word of the respective unit at the word boundary and before any 'nonverbal' such as <äh>, <pause>, <laughter>, etc. Turn-initially, no label is given. In contrast to the former strategy, turn-finally, the left context (last sentence/phrase etc.) is labelled with the appropriate label as well. By that, we will be able to model turn-final syntactic units, if this will be of any use for, e.g., dialogue act classification or dialogue act boundary classification. It would, however, be no problem to map these labels onto M3S or 'end of turn', if necessary. Up to now, we followed the strategy only to label with M adjacent to irregular boundaries if a sentence is not completed and another syntactic construction starts anew. At irregular boundaries, in contrast to this former strategy, a label is always given, if possible. If this information is not of any use, we can map these M labels that are adjacent to irregular boundaries onto 'null'; we can, however, have a closer look at these combinations as well.

Generally, we cannot subspecify beyond the levels encoded by our labels, i.e., we cannot specify two levels of subordination; other possible subspecifications are merged. For *sentences* (up to now M3S), we denote subordination, coordination, left/right relationship and prosodic marking. With these distinctions, we cannot denote *all* constellations. We only have one level for subordination, i.e., with SC2, we cannot denote which one of these clauses is subordinated w.r.t. the other one. After free phrases (elliptic sentences) followed by a subordinate clause, SM2 or SM1 is labelled as well: *Wunderbar* SM1 *daß Sie da Zeit haben* (Great SM1 that you'll have time by then); this constellation is very rare and because of that, it makes not much sense to model it in a special way. In analogy, phrasal coordination at subordinate clauses is labelled with SC3. For *free phrases* (up to now M3P), besides the 'main' label PM3, we annotate with PM1 free phrases that are prosodically integrated with the following adjacent sequence. Sequences inside free phrases are analogous to the

Table 12
Examples with context for new boundary labels and their main classes, part I (with reference to the old boundary labels)

Main class	Label	Context (between/at) with example
sentences, up to now: M3S		
M3	SM3	Main clause and main clause: *vielleicht stelle ich mich kurz vorher noch vor SM3 mein Name ist Lerch* perhaps I should first introduce myself SM3 my name is Lerch
M3	SM2	Main clause and subordinate clause: *ich weiß nicht SM2 ob es auch bei Ihnen dann paßt* I don't know SM2 whether it will suit you or not
M3	SS2	Subordinate clause and main clause: *da ich aus Kiel komme SS2 wird hier ja relativ wenig gefeiert* because I am from Kiel SS2 we don't celebrate that often
M3	SM1	Main clause and subordinate clause, prosodically integrated: *ich denke SM1 das können wir so machen* I think SM1 we can do it that way
M3	SS1	Subordinate clause and main clause, prosodically integrated: *das sieht sowieso ziemlich schlecht aus SS1 würd' ich sagen* anyway, that looks rather bad SS1 I'd say
M3	SC3	Coordination of main clauses and of subordinate clauses: *dann nehmen wir den Montag SC3 und treffen uns dann morgens* then we'll take Monday SC3 and meet in the morning
M3	SC2	Subordinate clause and subordinate clause: *da ich froh wäre SC2 diese Sache möglichst schnell hinter mich zu bringen* because I would be glad SC2 to get it over as soon as possible
free Phrases, up to now: M3P		
M3	PM3	free Phrase, stand alone: *sehr gerne PM3 ich liebe Ihre Stadt* with pleasure PM3 I love your town
M2	PC2	sequence in free Phrases: *um neun Uhr PC2 in 'nem Hotel PC2 in Stockholm* at nine o'clock PC2 in a hotel PC2 in Stockholm
M3	PM1	free Phrase, prosodically integrated, no dialogue act boundary: *guten Tag PM1 Herr Meier* hello PM1 Mr. Meier
Left dislocations, up to now: M3P		
M3	LS2	Left dislocation: *am fünften LS2 da hab' ich etwas* on the fifth LS2 I am busy
M2	LC2	sequence of Left dislocations: *aber zum Mittagessen LC2 am neunzehnten LS2 wenn Sie vielleicht da Zeit hätten* but for lunch LC2 on the 19th LS2 if you've got time then
Right dislocations, up to now: M3E		
M3	RS2	Right dislocation: *wie würde es Ihnen denn am Dienstag passen RS2 den achten Juni* will Tuesday suit you RS2 the eighth of June
M2	RC2	sequence of Right dislocations: *es wäre bei mir dann möglich RS2 ab Freitag RC2 dem fünfundzwanzigsten* it would be possible for me RS2 from Friday onwards RC2 the 25th
M2	RC1	Right 'dislocation' at open verbal brace: *treffen wir uns RC1 um eins* let's meet RC1 at one o'clock

Table 13
Examples with context for new boundary labels and their main classes, part II (with reference to the old boundary labels)

Main class	Label	Context (between/at) with example
Embedded strings, up to now: M31		
M3	EM3	embedded sentence/phrase: *eventuell EM3 wenn Sie noch mehr Zeit haben EM3 ⟨Atmung⟩ 'n bißchen länger* possibly EM3 if you've got even more time EM3 ⟨breathing⟩ a bit longer
Free particles, up to now: M3T		
M3	FM3	pre-/postsentential particle, with ⟨pause⟩ etc.: *gut FM3 ⟨Pause⟩ okay* fine FM3 ⟨pause⟩ okay
Discourse particles, up to now: M3D		
MU	DS3	pre-/postsentential particle, ambisentential: *dritter Februar DS3 ja DS3 ab vierzehn Uhr hätt' ich da Zeit* third February DS3 isn't it/well DS3 I have time then after two p.m.
MU	DS1	pre-/postsentential particle, no ⟨pause⟩ etc.: *also DS1 dienstags paßt es Ihnen DS1 ja M3S ⟨Atmung⟩* then DS1 Tuesday will suit you DS1 won't it / after all ⟨breathing⟩
Ambiguous boundaries, up to now: M3A		
MU	AM3	between sentences, Ambiguous: *würde ich vorschlagen AM3 vielleicht AM3 im Dezember AM3 noch mal AM3 dann* I'd propose AM3 possibly AM3 in December AM3 again AM3 then
MU	AM2	between free phrases, Ambiguous: *sicherlich AM2 sehr gerne* sure/-ely AM2 with pleasure
MU	AC1	between constituents, Ambiguous: *wollen wir dann AC1 noch AC1 'n Treffen machen* should we then (still) have a meeting / should we then have another meeting
Constituents, up to now: M2I		
M2	IC2	between Constituents: *ich wollte gerne mit Ihnen IC2 ein Frühstück vereinbaren* I'd like to arrange IC2 a breakfast with you
M2	IC1	asyndetic listing of Constituents (not labelled up to now): *wir haben bis jetzt eins IC1 zwei IC1 drei IC1 vier IC1 fünf IC1 sechs Termine* until now, we've got one IC1 two IC1 three IC1 four IC1 five IC1 six appointments
Default, no boundary, up to now: M0		
M0	IC0	every other word boundary: *da bin ich ganz Ihrer Meinung* I fully agree with you

constituent boundaries IC2 and labelled with PC2. *Left dislocations* (up to now M3P) are constituents to the left of the matrix sentence, typically but not necessarily with some sort of anaphoric reference in the matrix sentence. Sequences inside left dislocations are also analogous to the constituent boundaries IC2 and labelled with LC2. *Right dislocations* (up to now M3E) are subspecified further as well: Any constituent boundary appearing after RS2 has to be labelled with RC2 instead of IC2 because once a right dislocation is opened, all following constituents become additions to the dislocation. For right dislocations at open verbal brace, a new label RC1 is introduced. *Embedded sentences* (up to now M3I) are all sentences embedded in a matrix sentence that continues after the embedded sentence. In contrast to the former strategy, even very short parentheses (*glaub ich*)

Table 14
Encoding of syntactic type, syntactic hierarchy, and prosodic–syntactic strength of the new M labels

Label	Description
Type	**S**entence
	free **P**hrase
	Left dislocation
	Right dislocation
	Embedded sentence/phrase
	Free particle
	Discourse particle
	Ambiguous boundary
	Internal constituent boundary
Hierarchy	**M**ain, **S**ubordinate, **C**oordinate
Strength	Prosodic–syntactic strength: strong (3), intermediate (2), weak (1), very weak (0)

are annotated with E3; if necessary, these short parentheses (less or equal two words) can be relabelled automatically. *Free/discourse particles* (up to now M3T/M3D) are defined a bit differently than before: In contrast to the former strategy, we use PM3, if such a particle unequivocally can be classified as a confirmation, as in A: *Paßt Ihnen drei Uhr* SM3 – B: *Ja* PM3 *Dann zum zweiten Termin* ... (A: Is three o'clock ok with you? SM3 B: Yes. PM3 And now the second date ...). Much more common is, however, that the particle is followed by a sort of equivalent confirmation, e.g.: B: *Ja* DS1/FM3 *paßt ausgezeichnet* SM3 *Dann zum zweiten Termin* ... (B: Well/Yes DS1/FM3 that's fully ok with me. SM3 And now the second date ...). Here, we simply cannot tell apart the two functions 'confirmation' or 'discourse particle'. This is, however, not necessary because in these cases, the functional load on this particle is rather low. It might thus be the most appropriate solution *not* to decide in favour of the one or the other reading but to treat this distinction as underspecified or neutralized. This means for the higher linguistic modules that, in constellations like this, these particles might simply be treated as discourse particles without any pronounced semantic function; i.e., in the short run, they can be neglected.

There are three levels for *Ambiguous boundaries* (up to now M3A): AM3 and AM2 are ambiguous boundaries between clauses and phrases, respectively, and are discussed in more detail in (Batliner et al., 1996b); DS3 labels denote ambiguous sentence/phrase boundaries at pre-/postsentential particles. Particles that are very often surrounded by the new AC1 label are, e.g., *auch* (also/as well), *doch* (but/however/yet), *noch* (still). There are always at least two syntactic and semantic readings for clauses containing such particles; these readings can be described with different syntactic bracketing, but prosodically, accent structure is more important, i.e., whether the particle is accented or not. Consider the sentence: *Dann brauchen wir noch einen Termin.* If *noch* is accented, it has to be translated with Then we need another date, if *Termin* is accented and *noch* not, the translation is: Then we still need a date.

There are two types of *constituent boundaries* (up to now not labelled) between constituents (IC2) and between words/noun phrases in the case of asyndetic listing (IC1), i.e., a listing without any conjunction (*und*, etc.). The decision whether to put in an IC2 label or not is very often difficult to make. The criteria are basically prosodic: First, the boundary should be really inside the clause, i.e. far from left and right edges, and second, the constituent that precedes the boundary is 'prosodically heavy', i.e. normally a noun phrase that can be the carrier of a primary accent. Primarily, these boundaries will be used to trigger accent assignment, cf. Section 11 below and (Kompe, 1997). The criteria are discussed in more detail in (Batliner, 1997).

Basically, the old main classes are still valid; in addition, we introduce a fourth main class, M2, for those labels which denote boundaries at constituents (typically noun phrases) within larger syntactic units: PC2, LC2, RC2, RC1, IC2 and IC1.

The relabelling was conducted in the following steps: First, two linguists, C and N, discussed scheme and necessary editions with the first author who has annotated with the first version of the labelling scheme, cf. Table 1. Then C relabelled and thereby corrected or redefined the old labels, and afterwards, N checked and, if necessary, corrected C's labels. By that, we had two independent runs where errors could be detected and the labelling could be systematizised. A new subset, VM-CD 7 (1740 turns), was labelled independently by C and N and serves as database for the computation of the interlabeller consistency, cf. Section 8.

7. Correspondences of the new M labels with B and D labels

As far as the main classes are concerned, we did not expect that the new M labels (M_new) would show more correspondence with B and D than the old ones (M_old) but that our subcategorizations would lead to a better separation of subclasses. This is shown in Figs. 1–6 where for B-TRAIN, correspondences of M_new with B and with D are shown; note that due to technical reasons, only 737 out of the 797 turns of B-TRAIN could be processed for these correspondences. The mapping of M_old onto M_new is given in Tables 12 and 13. If inside M_old, the distribution of the new subclasses was equal, there should not be much of a difference between the new subclasses. For instance, the old label M3S for 'sentences' is now subspecified into seven new labels (SM3 to SC2), which are explained in Table 12. Their correlation with the B labels is shown in Fig. 3: There are considerable differences, ranging from more than 90% correlation with B3/B2 for SM3 to 0% correlation for SS1. The differences shown in Figs. 3 and 4 are almost always congruent with our working hypothesis that the coding of strength (third character of the labels) and the coding of hierarchy (second character of the labels) covaries with the percentage of items marked by prosodic–perceptual boundaries: Globally, from left to right, and within subgroups from left to right as well, correspondences with B3/B2 decrease and vice versa, those with B0 increase. This overall pattern shows up more clearly in figures than in tables; we decided therefore in favor of this presentation. This result is consistent with (de Pijper and Sanderman, 1994) who found that major syntactic boundaries tend to be marked prosodically to a greater extent than minor syntactic boundaries. A similar behaviour of the sublabels of M_new can be seen for the correspondence with D in Figs. 5 and 6. We can thus conclude that M_new meets prosodic regularities better than M_old.

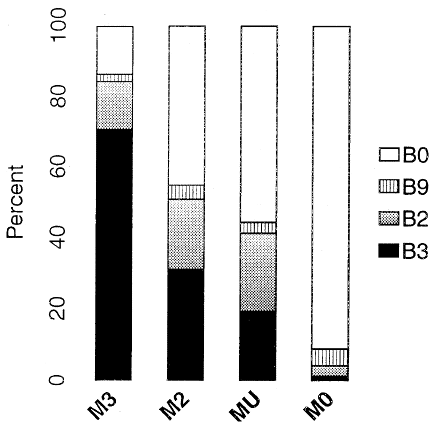

Fig. 1. Correspondences: M labels with B labels, main classes.

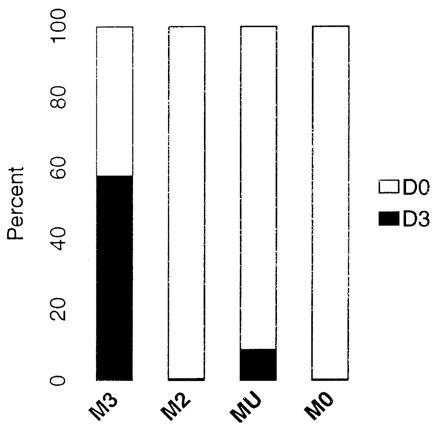

Fig. 2. Correspondences: M labels with D labels, main classes.

If we compare the correspondences of M_new with B0 and D0, it can be seen that for some few labels, there is a strong positive correlation between B0 and D0, in particular for SM3 and for IC0, but most of the other labels are somewhere 'in between'. That means that for a classification of D3 based on M_new, it would be suboptimal to use these labels only within a Multi-layer Perceptron (MLP) or only within a language model (LM). As D3 boundaries tend to be marked prosodically to a great extent, it might be better to use a two stage procedure instead: First, to classify these M boundaries with an MLP into prosodically marked or not, and then to use only the marked boundaries in a classification of D3. Our expectation is that D3 is *always* marked prosodically whereas D0 *might be* marked prosodically, because not every phrase boundary is a dialog act boundary. This is somehow confirmed with the figures, because when an M_new label highly correlates with B0 it also strongly correlates with D0; when a M_new label does not correlate with B0 it might still correlate with D0.

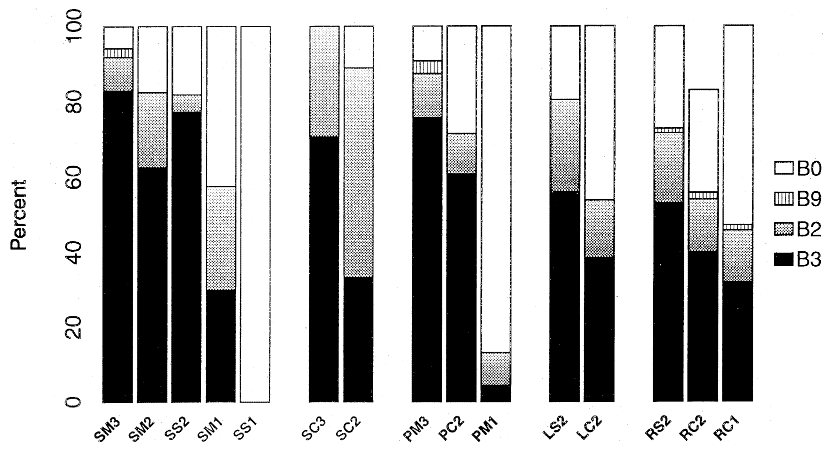

Fig. 3. Correspondences: M labels with B labels, sentences, free phrases, left dislocations, right dislocations.

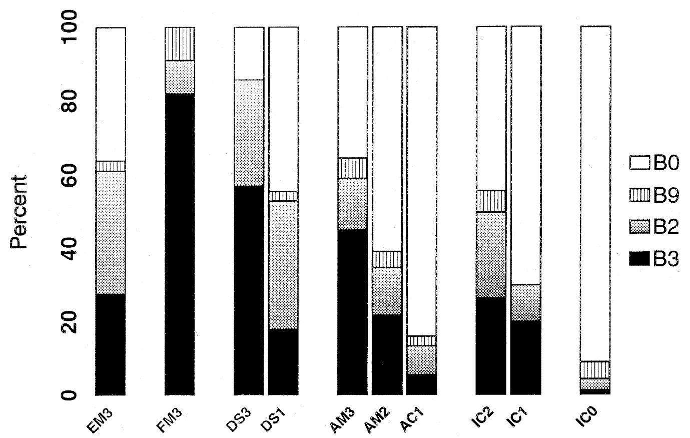

Fig. 4. Correspondences: M labels with B labels, embedding, free/discourse particles, ambiguous boundaries, constituents, no boundaries.

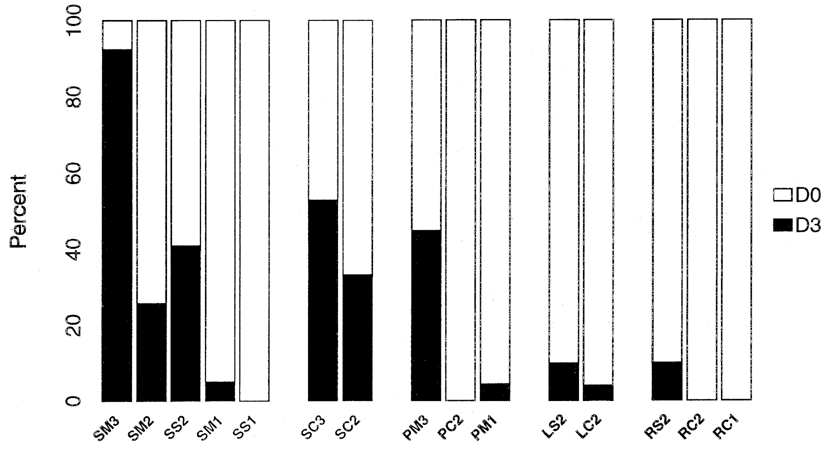

Fig. 5. Correspondences: M labels with D labels, sentences, free phrases, left dislocations, right dislocations.

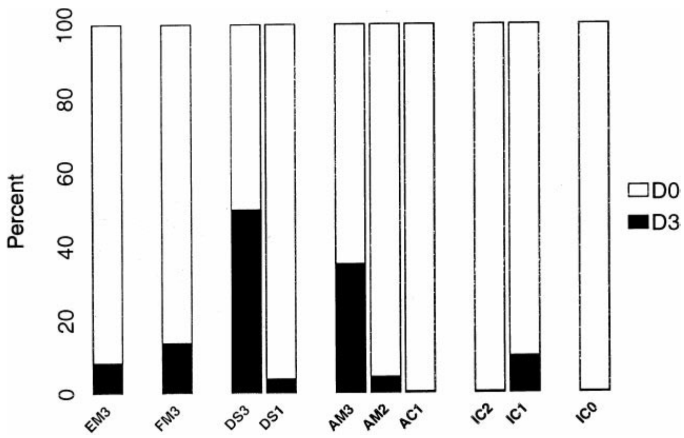

Fig. 6. Correspondences: M labels with D labels, embedding, free/discourse particles, ambiguous boundaries, constituents, no boundaries.

8. Further evaluation of the labelling scheme: reliability, effort

There exist different evaluation measures for labelling schemes. Common practice in basic research is to check the inter- or intralabeller consistency, cf. (Pitrelli et al., 1994; Grice et al., 1996; Reyelt, 1998), common practice in applied research is the quality of automatic classification, cf. Section 9: If an automatic classifier yields high recognition rates, this is great evidence for a consistent labelling. A common problem for all these measures is the fact that there is no 'objective' reference everybody agrees on as is the case for phoneme or word recognition: All these annotations are based on partly different and competing theoretical assumptions. The ultimate proof for such a scheme is, in our opinion, therefore only a sort of *external validity*, namely its usefulness for the intended purpose, in our case, the usefulness of the automatically predicted M labels for the higher linguistic modules in the VM system, cf. Section 10. In the present section, we will show how the M labels meet the criteria of different measures of *reliability*. We will distinguish *internal* and *external* reliability – internal or external to the specific annotation scheme, in our case, the M labels. Internal reliability can be measured *intra* labeller (same labelling scheme, same data, same labeller, different time) and *inter* labellers (same labelling scheme, same data, different labellers). External reliability can be measured *intra* labelling scheme (manual annotation versus automatic classification, same labelling scheme, same sort of data) or *inter* labelling schemes (correspondence of different labelling schemes, automatic classification of one scheme based on a sample trained with another scheme). *External reliability intra* will be dealt with in Section 9, *external reliability inter* has been dealt with above in Sections 5 and 7, as far as the correlation between the different schemes is concerned, and will be dealt with in Section 9, as far as the classification of the D labels based on the M labels is concerned.

As a metric for the internal reliability of our labelling scheme, we compute either correspondences (for single classes) or, as overall metric, the value κ because it has been computed for other German data and for prosodic–perceptual labels as well and because it is the most meaningful metric for such comparisons, cf. (Reyelt, 1998; Maier, 1997). κ cannot be used for the external reliability or for the correspondence between old and *all four* new M labels because there is no one-to-one relationship between the labels. For the comparison of old versus new labelling scheme, the new main labels can serve as reference, and the old ones, so to speak, as more or less correct 'approximation'. Table 15 thus displays frequency for each class and percent 'correct correspondence' for M_new. M2 has not been labelled in the first stage, so 82% correspondence with M0_old meets our expectation. The correspondence of M0_new with M0_old is almost perfect, the one of M3_new with

Table 15
Correspondence of new (ordinate) with old (abscissa) M labels in percent, four main classes, with number of occurrence

Label		M3	MU	M0
	#	19135	4706	99719
M3	20214	85	1	14
M2	6415	17	0	83
MU	8421	9	53	38
M0	88510	0	0	100

M3_old is 85%; most of the 14% correspondence with M0_old are certainly due to the modified labelling principles because in the first stage, M3 was not labelled at the end of turns or at irregular boundaries. For the three classes M3, MU and M0, κ is 0.88. Only the correlation of MU_new with MU_old (53%) is rather low; we will come back to this point in the discussion of Table 16, where we display interlabeller correspondence of the labellers C and N for the new labels. In this table, we do not display percent values but the frequencies of correspondences between the two labellers. Meaningful percent values are mean percent correspondence for the hits: M3: 90%, M2: 67%, MU: 65%, M0: 95%. Confusions of M3 with M0 occur very seldom, i.e., 'false alarms' of M3 as M0: 2%, 'false alarms' of M0 as M3: 0%. We see that there is a very good agreement for the most important classes M3 and M0, while it is not that good for M2 and MU. This is due to the fact that these labels cannot be defined as strictly as the two other ones: The labelling of M2 is triggered by the imagination of the labellers w.r.t. the possibility of a prosodic marking of the respective boundaries. As for the MU labels, they can be labelled rather easily at pre/postsentential particles with DS1 (83%) but rather less easily for the 'real' ambiguous boundaries AM3 and AM2 (27% and 20%, respectively). This can be explained by the fact that for these labels, the labellers have to decide quickly whether there is a syntactically ambiguous boundary and, at the same time, whether it is 'semantically/pragmatically reasonably ambiguous' as well – a task that obviously does not lead to a good interlabeller agreement. Each labeller might be biased towards MU labels for special constructions, but often, the decision whether to use MU or not might be random as well. In such a case, a random assignment of ambiguous boundaries to one of the four main classes should not disturb the behaviour of a statistical classifier trained on these labels too much.

Table 16
Interlabeller correspondence for new M labels, VM–CD 7, four main classes, with number of occurrence. κ: 0.79

Label		M3	M2	MU	M0
	#	6913	2607	3404	29865
M3	6749	6146	299	223	81
M2	2273	84	1635	56	498
MU	4369	532	259	2500	1078
M0	29398	151	414	625	28208

κ for the main classes is 0.79, for all 25 classes, it is 0.74. (Reyelt, 1998) obtained κ values between 0.5 and 0.8 for prosodic–perceptual boundary labels, depending on the experience of the labellers and on the specific tasks.

To check intralabeller consistency, labeller C relabelled eight dialogues from VM-CD 7 after several months. For the main classes, mean percent correspondence for the hits are M3: 96%, M2: 78%, MU: 78%, M0: 96%. κ is 0.86 for the main classes and 0.84 for all 25 classes. Maier (1997) reports a κ value of 0.93 for the segmentation of dialogue acts inter labellers and of 0.94 intra labeller. As is the case for regression coefficients, there is no threshold value that distinguishes 'good' from 'bad' values but κ values around 80% and above can be considered to represent a very good agreement.

The ToBI correspondence for boundaries (13 labellers, 4 categories) in (Grice et al., 1996) is 87%; this value could be compared with our mean percent values given above; all these coefficients cannot be compared in a strict sense, however, because number of labellers and/or categories differ. Moreover, the importance of the categories is not equal: as for the M labels, not overall correspondence might be the appropriate measure for the quality of the labelling scheme but only 'false alarms' for M3 if, e.g., only these are problematic for the processing in the higher linguistic modules.

It might thus be safe to conclude that interlabeller correspondence for the M labels is good enough, and that means, that – with a reasonable amount of training – the labelling can be conducted by different persons with a fairly good knowledge of German syntax. For many applications, only some main M boundaries might be relevant which presumably can be labelled by less experienced labellers as well.

Effort needed is, in practice, at least as important as reliability, and an explicit criterion in applied research. Here, procedures that are coarse but fast can and must be preferred because a certain time-out is always a delimiter: the end of a project, the financial means available, the maximum processing time allowed for an automatic speech processing system. In basic research, effort is not that often mentioned but is, in fact, implicitly equally important. Note, e.g., that practically all perception experiments violate the fundamental claim of inferential statistics that subjects have to be chosen random out of the population (Guttman, 1977). This is done simply because otherwise, effort needed for the selection of subjects would paralyse the experiments themselves.

We claim that the annotation of the M labels needs considerably less effort than any annotation where labellers have to listen to the speech material and particularly, less effort than the labelling of prosodic–perceptual boundaries. In order to get an impression of the time needed, labeller C annotated three different subsets of VM-CD 12 with 4 dialogues each, keeping track of the time needed, with the 8 old M labels (ca. 5 times real time), with the 25 new M labels (ca. 8 times real time), and with only the 3 old main classes M3, MU and M0 (ca. 5 times real time), respectively. Number of labels and time needed are obviously not correlated linearly with each other. Even if the annotation of only 3 main classes is intended, it seems that the labeller conducts a more thorough syntactic analysis in which he uses more than these 3 classes. Such a lower limit might correspond to the effort needed for the annotation of the 8 old M classes. Therefore, roughly the same time is needed for 8 and for 3 classes; only little more time is needed for a further splitting up into 25 detailed classes. Note that while discussing the new labelling scheme, these 25 classes seemed to be some sort of upper limit for a 'rough' labelling on which we could agree. It might be that a more detailed labelling scheme would mean to 'cross the Rubicon' towards a deep syntactic analysis for which much more time is needed.

The time needed for the annotation of a corresponding subsample, two dialogues of VM-CD 1, with ToBI (prosodic–perceptual annotation) is 23 times real time, and only with boundaries and accents, it is 13 times real time; the annotation of only the boundary labels takes not much less time. Note that these were 'easy' dialogues, i.e., on the average, it takes more time because the labeller has to listen to difficult turns much more often. Moreover, the only tools for the labelling of M needed are either paper and pencil and/or any ASCII-editor on any platform. For the labelling of prosodic–perceptual boundaries in ToBI, several preprocessing stages are necessary: word segmentation (by hand or automatically computed with correction by hand taking more than 10 times real time) and computation of F_0 curve. First, annotation tools have to be developed (or purchased), and there is always a workstation needed for each labeller. Besides, the training for the labellers of ToBI takes much more time than that needed for the labellers of the M labels. All these figures and details are provided by Reyelt (1997).

Up to the end of 1997, 700 German, 160 English and 100 Japanese dialogues were annotated at the DFKI, Saarbrücken, with dialogue act information including D3 boundaries without listening to the speech data. This took approximately 10 hours a week for two years. These figures are provided by Reithinger (1997). Note that for this task, the annotation of the dialogue acts takes much more time than that of the dialogue act boundaries which are a sort of 'by-product'.

9. Automatic classification of boundaries

First, we want to stress again that our ultimate goal is to prove the usefulness of our labelling scheme for syntactic and other higher processing as dialogue analysis in the higher modules of the VM project. A necessary but not necessarily sufficient precondition for such a successful incorporation in a whole system is a good automatic classification of boundaries with the help of the M labels: If classifications were good but if the classes were irrelevant for syntactic processing the M labels were of no use. On the other hand, if automatic classification was bad it would be very unlikely that the labels can be of any use. We therefore repeated on Ms and Ds some of the classification experiments we had done before on Bs and Ss, cf. (Kießling et al., 1994b; Kompe et al., 1994, 1995; Batliner et al., 1996a; Kießling, 1997; Kompe, 1997). According to the comparison of Ms and Ds with Bs, cf. above, we can expect these boundary types to be often marked prosodically. Therefore, MLPs were used for the *acoustic–prosodic* classification of M0 versus M3 and of D0 versus D3 boundaries, respectively. All experiments described in this section are based on the unrevised old main classes M3, MU and M0 (M_old). The revised classes will be used in the second stage of the VM project in 1998–2000. Note that the mapping of old onto new labels is very good, cf. Section 8, and, if changes were made, transparent. We therefore do not expect that overall classification results with the new labels will be very different but that the subspecification of the new labels makes even better classification results likely for some specific tasks.

We distinguish different categories of prosodic feature levels. The *acoustic prosodic features* are signal-based features that usually span over speech units that are larger than phonemes (syllables, words, turns, etc.). Normally, they are extracted from the specific speech signal interval that belongs to the prosodic unit, describing its specific prosodic properties, and can be fed directly into a classifier. Within this group we can further distinguish as follows.

- *Basic prosodic features* are extracted from the pure speech signal without any explicit segmentation into prosodic units. Examples are the frame-based extraction of fundamental frequency (F_0) and energy. Usually the basic prosodic features cannot be directly used for a prosodic classification.
- *Structured prosodic features* are computed over a larger speech unit (syllable, syllable nucleus, word, turn, etc.) partly from the prosodic basic features, e.g., features describing the shape of F_0 or energy contour, partly based on segmental information that can be taken from the output of a word recognizer, e.g., features describing durational properties of phonemes, syllable nuclei, syllables, pauses.

On the other hand, prosodic information is highly interrelated with ‘higher’ linguistic information, i.e., the underlying linguistic information strongly influences the actual realization and relevance of the measured acoustic prosodic features. In this sense, we speak of *linguistic prosodic features* that can be introduced from other knowledge sources, as lexicon, syntax or semantics; usually they have either an intensifying or an inhibitory effect on the acoustic prosodic features. The linguistic prosodic features can be further divided into two categories.

- *Lexical prosodic features* are categorical features that can be extracted from a lexicon that contains syllable boundaries in the phonetic transcription of the words. Examples for these features are flags marking if a syllable is word-final or not or denoting which syllable carries the lexical word accent. Other possibilities not considered here might be special flags marking, e.g., content and function words.
- *Syntactic/semantic prosodic features* encode the syntactic and/or semantic structure of an utterance. They can be obtained from syntax, e.g., from the syntactic tree, or they can be based on predictions of possibly important – and thus accented – words from the semantic or the dialogue module.

All these categories are dealt with in more detail in (Kießling, 1997). Here, we do not consider syntactic/semantic prosodic features; in the following, the cover term prosodic features means mostly structured prosodic features and some few lexical prosodic features. We only use the aligned spoken words thus simulating 100% word recognition – and by that, simulating the capability of a human listener. The time alignment is done by a standard HMM word recognizer. It is still an open question, which prosodic features are the most relevant for the different classification problems and how the different features are interrelated; cf. below and (Batliner et al., 1997). MLPs are generally good at handling features that are even highly correlated with each other; we therefore try to be as exhaustive as possible, and leave it to the statistical classifier to find out the relevant features and the optimal weighting of them. As many relevant prosodic features as possible are therefore extracted over a prosodic unit (here: the word-final syllable) and composed into a huge feature vector which represents the prosodic properties of this and of several surrounding units in a specific context.

In more detail the features used here are as follows:

- For each syllable and word in the specific context minimum and maximum of fundamental frequency (F_0) and their positions on the time axis relative to the position of the actual syllable as well as the F_0-mean.
- F_0-offset + position for actual and preceding word.
- F_0-onset + position for actual and succeeding word.
- Linear regression coefficients of F_0-contour and energy contour over different windows to the left and to the right of the actual syllable characterizing the slope – falling versus rising – of these contours.
- For each syllable and word in this context maximum energy (normalized as in Wightman, 1992) + positions and mean energy (also normalized).
- Duration (absolute and normalized) for each syllable/syllable nucleus/word.
- Length of the pause preceding/succeeding actual word.
- For an implicit normalization of the other features, measures for the speaking rate are computed over the whole utterance based on the absolute and the normalized syllable durations (as in (Wightman, 1992)).
- For each syllable: flags indicating whether the syllable carries the lexical word accent or whether it is in a word final position.

MLPs of varying topologies were investigated, using always M-TRAIN for the training of M3 versus M0 and TEST for testing M3, M0 and MU. In the following, all results reported are for the spoken word chain. The best recognition rate obtained so far is 86.0% for the classification of M3 versus M0; the confusion matrix is shown in Table 17. MU is of course not a class which can be identified by a specific prosodic marking but MU boundaries are either marked prosodically (denoting a syntactic boundary) or not. Therefore, in the table we give the percentage of MU mapped to M3 or M0. The fact that the decision is not clear in favour of one or the other proves that MU marks ambiguous boundaries.

Table 17
Best recognition results in percent so far for three M main classes using acoustic–prosodic features and MLP; 121 prosodic features computed in a context of ± 2 syllables were used

Reference		% Recognized	
Label	#	M3	M0
M3	177	87.6	12.4
MU	103	61.2	38.8
M0	1169	14.2	85.8

Tables 17 and 18 display frequency and percent recognized for each class; all information necessary to compute different metrics as, e.g., recall, precision, false alarms, errors, and chance level for each class are given in these figures. Chance level for M3, e.g., is 12% (177/(177 + 103 + 1169); for M0, it is 81%. We see, that both classes are recognized well above chance level, whereas an arbitrary mapping of all boundaries onto M0 would result in a class-wise computed recognition rate of 50%. Note that the class-wise results obtained are actually the most relevant figures for measuring the performance of the MLP, because the MLP got the same number of training patterns for M3 and M0. Therefore it does not "know" about a priori probabilities of the classes. We chose this approach because the MLP was combined with a stochastic language model (cf. below) which encodes the a priori probabilities of M3 or M0 given the words in the context of a boundary. The high recognition rates for M0 as well as for M3 and the fact that only 61% of the MU were recognized as

Table 18
Best recognition results in percent so far for three M main classes using *n*-grams

Reference		% Recognized		
Label	#	M3	MU	M0
M3	177	77	0	23
MU	103	8	52	40
M0	1169	2	0	98

M3 confirms the appropriateness of our labelling scheme: M3 and M0 can be told apart in most of the cases (the average of the class-wise computed recognition rates is 86.7%) which is in the same range as the result obtained with an MLP for B3 versus ¬B3 which was 86.8%, cf. (Kießling, 1997, p. 191). Remember that the MLP uses mostly acoustic features and could thus be imagined to work better for the perceptually evaluated B labels than for the M labels; of course, the larger training database for the M labels contributes as well. According to our expectations, the MU labels are neither assigned fully to M3 nor to M0 but distributed roughly equally.

Similar experiments with different feature subsets and MLP topologies were conducted for the D0 versus D3 classification as well. Trained with D0 versus D3, they yielded a recognition rate of 85% (class-wise recognition: 83%). We achieved a respectable recognition rate of 82% (class-wise recognition: 82%) with classifiers trained with M3 and M0. Here as well, the larger training database for the M labels might contribute, but this result proves at the same time that a mapping of M onto D is really possible; more details can be found in (Kompe, 1997, p. 202). The lower recognition rate for the Ds compared to the Ms is due to the fact that a lot of D0s correspond to M3 and thus can be expected to be marked prosodically.

In (Batliner et al., 1997), we investigated the predictive power of the different feature classes included in our feature vector. It turned out that practically all feature classes alone yield results above chance level, but that the best result can be achieved by simply using all features together (31.6% reduction of error rate w.r.t. all features taken together vs. best feature class alone). This result confirms our general approach not to look for important ('distinctive') features but to take into account as many features as possible.

Since the Ms and Ds were labelled by linguistic criteria, one should be able to reliably predict them by a stochastic language model provided they are labelled consistently. In (Kompe et al., 1995; Kompe, 1997), we introduced a classifier based on an *n*-gram language model (LM) and reported results for the prediction of boundaries. This classifier is based on estimates of probabilities for boundary labels given the words in the context. The same approach was used for the prediction of Ms and Ds, that is, the *n*-grams were trained on a large text corpus and then for the test corpus, each word boundary was labelled automatically with the most likely label. The results for the Ms given in Table 18 meet our expectations as well. Note especially, that the MUs really cannot be predicted from the text alone. In similar classification experiments, D0 versus D3 could be recognized correctly in 93% of the cases.

MLP and *n*-gram each model different properties; it thus makes sense to combine them. In Table 19, we compare the results for different combinations of classifiers (MLP for B versus ¬ B and LMs for S Labels: LM_S, and for M Labels: LM_M) for the two main classes boundary versus nonboundary for three different types of boundaries: B, S and M. Here, the undefined boundaries MU and S3? are not taken into account. The first number shows the overall recognition rate, the second is the average of the class-wise recognition rates. All recognition results were again measured on TEST. For the training of the MLP and the LM_S, all the available labelled data was used except for the test set (797 and 584 turns, respectively); for LM_M, 6297 turns were used. It can be noticed that roughly, the results get better from top left to bottom right. Best results can be achieved with a combination of the MLP with the LM_M no matter whether the perceptual B or the syntactic–prosodic M labels serve as reference. LM_M is better than the LM_S even for S3 versus S¬3 because of the greater amount of training data. The LMs alone are already very good; we have, however, to consider that they cannot be

Table 19
Recognition rates (total/class–wise average) in percent for different combinations of classifiers (first column) distinguishing between different types of boundaries (B, S and M)

Cases	B3 vs. ¬B3 165 vs. 1284	S3+ vs. S3- 210 vs. 1179	M3 vs. M0 190 vs. 1259
MLP	87/86	85/76	86/81
LM_S	86/80	92/86	92/83
MLP+LM_S	90/87	92/86	93/87
LM_M	92/85	95/87	95/86
MLP+LM_M	94/91	94/86	96/90

applied to the undefined classes MU and S3? which are of course very important for a correct syntactic/semantic processing. Particularly for these cases, we need a classifier trained with perceptual–prosodic labels. Due to the different a priori probabilities, the boundaries are recognized less accurately than the nonboundaries with the LMs; this causes the lower class-wise recognition rates (e.g., 80.8% for M3 versus 97.7% for M0 for MLP + LM_M). It is of course possible to adapt the classification to various demands, e.g., in order to get better recognition rates for the boundaries if more false alarms can be tolerated. More details are given in (Kompe, 1997).

Similar classification experiments with syntactic–prosodic boundaries are reported in (Wang and Hirschberg, 1992; Ostendorf et al., 1993), where HMMs or classification trees were used. The authors rely on perceptual–prosodic labels created on the basis of the ToBI system (Beckman and Ayers, 1994); for such labels, however, a much smaller amount of data can be obtained than in our case, cf. Section 8. Our recognition rates are higher maybe because of the larger amount of training data; note, however, that the studies cannot be compared in a strict sense because they differ considerably w.r.t. several factors: the labelling systems are different, the numbers of classes differ, Wang and Hirschberg (1992) included the end of turns as label, Ostendorf et al. (1993) used elicited, systematically ambiguous material that already because of that should be marked prosodically to a greater extent, and the languages differ.

10. The use of the M labels in the VM system

In this section, we will describe the usefulness of prosodic information based on an automatic classification of the M labels for syntactic processing in the VM system. More details can be found in (Niemann et al., 1997; Kompe et al., 1997; Kompe, 1997, p. 248ff). Input to the prosody module is the WHG and the speech signal. Output is a prosodically scored WHG (Kompe et al., 1995), i.e., to each of the word hypotheses, probabilities for prosodic accent, for prosodic boundaries, and for sentence mood are attached. Here, we will only deal with the boundary classification.

There are two reasons why syntactic processing heavily depends on prosody: First, to ensure that most of the spoken words are recognized, for spontaneous speech a large WHG has to be generated. Currently, WHGs of about 10 hypotheses per spoken word are generated. Finding the correct (or approximately correct) path through a WHG is thus an enormous search problem. A corpus analysis of VM data showed that about 70% of the utterances contain more than a single sentence (Tropf, 1994). About 25% of the utterances are longer than 10 s. This is worsened by the fact that spontaneous speech may contain elliptical constructions. Second, even if the spoken word sequence has been recovered by word recognition correctly without alternative word hypotheses, there still might be many different parses possible, due to the high number of ambiguities contained in spontaneous speech and due to the relatively long utterances occurring in the VM domain.

Both VM syntax modules use the boundary scores of the prosody module along with the acoustic score of the word hypotheses and n-gram stochastic LMs. Siemens uses a Trace Unification Grammar (TUG) (Block and Schachtl, 1992; Block, 1997), IBM a Head-driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1987; Kiss, 1995). These syntax modules are used alternatively. Both grammars contain boundary symbols. The basic difference is that the Siemens system has a word graph parser which searches in the word graph for the optimal word chain. It simultaneously explores different promising (i.e., well scored) paths. The search is guided by all scores including the prosodic score. In the IBM system, the word graph is first expanded to the n-best word chains, which include prosodic boundary labels. Here, ‘best’ refers to the scores including the prosodic score. Two alternative word chains might just differ in the boundary positions. These word chains are parsed one after the other until a parse could be completed successfully.

The following evaluations were conducted on 594 word graphs consisting of about 10 hypotheses per spoken word. The usefulness of prosodic information in the syntax module from Siemens can

Table 20
With the Siemens parser, 594 word graphs were parsed

	Without prosody	With prosody	Improvement (%)
# Successful analyses	368	359	−2
# Readings	137.7	5.6	96
Parse time (s)	38.6	3.1	92

Results are given for parse experiments using prosodic information and compared with results where no prosodic information was used. The last column shows the relative improvement achieved by using prosody.

be seen in Table 20 that shows the improvement of the Siemens WHG parser by using the prosodic boundary probabilities: The number of readings as well as the parse time are reduced drastically. The fact that 9 word graphs (i.e. 2%) could not be analyzed with the use of prosody is due to the fact that the search space is explored differently and that the fixed time limit has been reached before the analysis succeeded. However, this small number of nonanalyzable word graphs is neglectable considering the fact that without prosody, the average real-time factor is 6.1 for the parsing. With prosodic information the real-time factor drops to 0.5; the real-time factor for the computation of prosodic information is 1.0. Note that furthermore a high number of readings results in a larger computational load in the higher linguistic modules. For the IBM parser results are only available for speech recorded during tests with the VM system by non-naive users. With this material a speed-up of 46% was achieved by using the prosodic clause boundary information, cf. (Batliner et al., 1996a).

It is therefore safe to conclude that the use of syntactic–prosodic information, i.e., information coded in the M boundaries, is, at least for the time being, vital for the VM system because without this information, the overall processing time would be too long and thus acceptability too low. Note that we do not want to claim that it is only prosodic information that could do this job. This can be demonstrated easily because everybody who reads the transliterations without punctuation and keeps track of the dialogue history can understand practically all turns. The information needed for the M labels can, however, for the time being and for the near future be computed at much lower costs than the information w.r.t. dialogue history. This is partly due to the relatively advanced state of prosodic and syntactic analysis compared to dialogue analysis (Carletta et al., 1997) where there still is no agreement on even basic categories. Most important is, however, that the above mentioned human capacity cannot be transfered onto end-to-end-systems like VM which have to deal with spoken language, and that means, with WHGs. To extract the spoken word chain is normally no problem for human beings but a great problem for speech understanding systems. We do not want to claim either that we could not have come quite far with only a prosodic classification of the B boundaries. A look at Table 19 shows, however, that the reduction of error rate from top left to bottom right (B to M) is 69% for overall classification and 29% for class-wise computed classification results. (The overall error rate for B with an MLP is, e.g., $100 - 87 = 13\%$; for M with the combination of MLP with LM, it is $100 - 96 = 4\%$. The difference between these two error rates is 9%, which means 69% reduction of error rate: $9/13 = 0.69$.) Moreover, the M labels model more exactly than the B labels the entities we are interested in.

In Section 1, we claimed that prosodic information might be the only means to find the correct reading/parse for one and the same ambiguous sequence of words, cf. Examples 1 and 2. Several experiments on VM sub-corpora proved that prosody not only improves the parse time or the number of parse trees but also the average quality of the parse trees. This holds especially for the search through the WHG. The experiments are described in detail in (Kompe, 1997, pp. 263–269).

11. Concluding remarks and future work

For a robust training of stochastic language models like *n*-grams, huge amounts of appropriately labelled training data are necessary. Thus, the main advantage of the syntactic–prosodic M labels is the comparatively low effort of the labelling process and the comparatively large amount of data obtained. Above that, these labels are directly designed for use in the syntax module of the VM project but detailed enough so they can be used for automatic classification of dialogue act boundaries as well. This fact will become even more important when switching to different applications or languages. For the time being, we adapt the M system onto American English; this seems to be possible with only some slight modifications; e.g., the verbal brace does not exist in English, and thus the labels for right dislocations have to be redefined. As is the case with ToBI, the M labelling framework can thus be used at least for these two related languages; that means that labellers can use basically the same guidelines and principles for the annotation of German and English. For other applications in the same language, the training material can always be used for a bootstrap, i.e., the classifiers can be trained with the large M database and then later calibrated with a smaller database taken from the new application.

The large training database might be the reason why we could predict the S and the B boundaries with the M labels to such an extent. This does, however, not mean that we can do without the B labels: without these, we would be at a loss to disambiguate the semantically crucial MU boundaries. (Besides, there is 29% reduction of error rate for the class-wise computed recognition rate for the combination of LM and MLP w.r.t. the use of the LM alone, cf. Table 19.) For this task, however, we can do with a rather small database labelled with B.

We want to stress again that this sort of labelling is no substitute for a thorough syntactic analysis. We can define and label prototypical boundaries with a great reliability but for quite a few possible boundary locations, different interpretations are possible. A small percentage of the labels in the training database is certainly incorrect because the labelling was done rather fast. It is always possible that the labelling strategy is not fully consistent across the whole labelling procedure. On the other hand, the good classification results show that across this great database, such inconsistencies do not matter much. The *philosophy* behind this labelling scheme is a 'knowledge based' clustering of syntactic boundaries into subclasses that might be marked distinctively by prosodic means or that are prone not to be marked at all prosodically. The *purpose* of this labelling scheme is not to optimize a stand alone prosodic classification but to optimize its usefulness for syntactic analysis in particular and for the linguistic modules in VM in general. The whole concept is thus a compromise between a – manageable – generalization and a – conceivable – specification of the labels.

In the future, we want to use the new M labels for the automatic labelling of 'phrase' accent positions along the lines of (Kießling et al., 1994a). There, we designed rules for the assignment of phrase accent positions based on syntactic–prosodic boundary labels for a large read database achieving recognition rates of up to 88.3% for the two class problem 'accent' versus 'no accent'. The most important factors were part-of-speech and position in the accent-phrase, i.e., in the syntactic unit that is delimited by M boundaries; details can be found in (Kießling, 1997).

For the time being, the VM project is in a transition stage, from VM I to VM II, the latter being planned for 1997–2000. Database, task and all modules are redesigned. It is planned to use the M labels for similar purposes as in the past; above that, they will be annotated for English and used in the processing of some additional higher linguistic modules, as statistical grammar and statistical translation.

Acknowledgements

We discussed the overall usefulness of our labelling scheme as well as the labelling strategy and some problematic cases with S. Schachtl (Siemens, München), with A. Feldhaus, S. Geißler, T. Kiss (all of them from IBM, Heidelberg) and with M. Reyelt (TU, Braunschweig). N. Schumann and C. Schäferle (LMU, München) annotated the new M labels, V. Warnke and R. Huber (FAU, Erlangen) conducted some additional analyses. We want to thank all of them. Needless to say that none of them should be blamed for any inconsistencies or errors. We would like to thank two anonymous reviewers for helpful discussions of this work. This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grants 01 IV 102 F/4, 01 IV 102 H/0 and 01 IV 701 K/5. The responsibility for the contents lies with the authors.

References

Batliner, A., 1997. M specified: A revision of the syntactic–prosodic labelling system for large spontaneous speech databases. Verbmobil Memo 124.

Batliner, A., Kompe, R., Kießling, A., Nöth, E., Niemann, H., Kilian, U., 1995. The prosodic marking of phrase boundaries: Expectations and results. In: Rubio Ayuso, A., López Soler, J. (Eds.), Speech Recognition and Coding. New Advances and Trends, NATO ASI Series F, Vol. 147. Springer, Berlin, pp. 325–328.

Batliner, A., Kompe, R., Kießling, A., Mast, M., Nöth, E., 1996b. All about Ms and Is, not to forget As, and a comparison with Bs and Ss and Ds. Towards a syntactic–prosodic labeling system for large spontaneous speech databases. Verbmobil Memo 102.

Batliner, A., Feldhaus, A., Geissler, S., Kießling, A., Kiss, T., Kompe, R., Nöth, E., 1996a. Integrating syntactic and prosodic information for the efficient detection of empty categories. In: Proc. Internat. Conf. on Computational Linguistics, Copenhagen, Vol. 1, pp. 71–76.

Batliner, A., Kießling, A., Kompe, R., Niemann, H., Nöth, E., 1997. Can we tell apart intonation from prosody (if we look at accents and boundaries)?. In: Kouroupetroglou, G. (Ed.), Proc. ESCA Workshop on Intonation, Department of Informatics, University of Athens, Athens, pp. 39–42.

Bear, J., Price, P., 1990. Prosody, syntax, and parsing. In: Proc. 28th Conference of the Association for Computational Lingustics, Banff, pp. 17–22.

Beckman, M., Ayers, G., 1994. Guidelines for ToBI transcription, Version 2. Department of Linguistics, Ohio State University.

Block, H., 1997. The language components in Verbmobil. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, München, Vol. 79–82, p. 1.

Block, H., Schachtl, S., 1992. Trace & unification grammar. In: Proc. Internat. Conf. on Computational Linguistics, Nantes, Vol. 1, pp. 87–93.

Bub, T., Schwinn, J., 1996. Verbmobil: The evolution of a complex large speech-to-speech translation system. In: Internat. Conf. on Spoken Language Processing, Philadelphia, Vol. 4, pp. 1026–1029.

Cahn, J., 1992. An investigation into the correlation of cue phrases, unfilled pauses and the structuring of spoken discourse. In: Proc. IRCS Workshop on Prosody in Natural Speech, University of Pennsylvania, Pennsylvania, pp. 19–30.

Carletta, J., Dahlbäck, N., Reithinger, N., Walker, M., 1997. Standards for dialogue coding in natural language processing. Dagstuhl-Seminar-Report 167.

Cutler, A., Dahan, D., van Donselaar, W., 1997. Prosody in the comprehension of spoken language: A literature review. Language and Speech 40, 141–201.

de Pijper, J., Sanderman, A., 1994. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. Journal of the Acoustic Society of America 96, 2037–2047.

Feldhaus, A., Kiss, T., 1995. Kategoriale Etikettierung der Karlsruher Dialoge. Verbmobil Memo 94.

Grammont, M., 1923. Notes de phonétique générale. VIII. L'assimilation. In: Bulletin de la Société de Linguistique, Vol. 24, pp. 1–109.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J., Batliner, A., 1996. Consistency in transcription and labelling of German intonation with GToBI. In: Internat. Conf. on Spoken Language Processing, Philadelphia, Vol. 3, pp. 1716–1719.

Guttman, L., 1977. What is not what in statistics. The Statistician 26, 81–107.

Haftka, B., 1993. Topologische Felder und Versetzungsphänomene. In: Jacobs, J., van Stechow, A., Sternefeld, W., Vennemann, T. (Eds.), Syntax – Ein Internationales Handbuch zeitgenössischer Forschung – An International Handbook of Contemporary Research, Vol. 1. De Gruyter, Berlin, pp. 846–867.

Hirschberg, J., Grosz, B., 1994. Intonation and discourse structure in spontaneous and read direction-giving. In: Proc. Internat. Symp. on Prosody, Yokohama, Japan, pp. 103–109.

Hunt, A., 1994. A prosodic recognition module based on linear discriminant analysis. In: Internat. Conf. on Spoken Language Processing, Yokohama, Japan, Vol. 3, pp. 1119–1122.

Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M., Quantz, J., 1995. Dialogue acts in Verbmobil. Verbmobil Report 65.

Kießling, A., 1997. Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Berichte aus der Informatik. Shaker Verlag, Aachen.

Kießling, A., Kompe, R., Batliner, A., Niemann, H., Nöth, E., 1994a. Automatic labeling of phrase accents in German. In: Internat. Conf. on Spoken Language Processing, Yokohama, Japan, Vol. 1, pp. 115–118.

Kießling, A., Kompe, R., Niemann, H., Nöth, E., Batliner, A., 1994b. Detection of phrase boundaries and accents. In: Niemann, H., De Mori, R., Hanrieder, G. (Eds.), Progress and Prospects of Speech Research and Technology: Proc. CRIM / FORWISS Workshop. Infix, Sankt Augustin, PAI 1, pp. 266–269.

Kiss, T., 1995. Merkmale und Repräsentationen. Westdeutscher Verlag, Opladen.

Kohler, K., Lex, G., Pätzold, M., Scheffers, M., Simpson, A., Thon, W., 1994. Handbuch zur Datenaufnahme und Transliteration in TP14 von Verbmobil, V3.0. Verbmobil Technisches-Dokument 11, Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel, Kiel.

Kompe, R., 1997. Prosody in Speech Understanding Systems, Lecture Notes in Artificial Intelligence. Springer, Berlin.

Kompe, R., Batliner, A., Kießling, A., Kilian, U., Niemann, H., Nöth, E., Regel-Brietzmann, P., 1994. Automatic classification of prosodically marked phrase boundaries in German. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Adelaide, Vol. 2, pp. 173–176.

Kompe, R., Kießling, A., Niemann, H., Nöth, E., Schukat-Talamazzini, E., Zottmann, A., Batliner, A., 1995. Prosodic scoring of word hypotheses graphs. In: Proc. European Conf. on Speech Communication and Technology, Madrid, Vol. 2, pp. 1333–1336.

Kompe, R., Kießling, A., Niemann, H., Nöth, E., Batliner, A., Schachtl, S., Ruland, T., Block, H., 1997. Improving parsing of spontaneous speech with the help of prosodic boundaries. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, München, Vol. 2, pp. 811–814.

Lamel, L., 1992. Report on speech corpora development in the US. ESCA Newsletter 8, 7–10.

Lea, W., 1980. Prosodic aids to speech recognition. In: Lea, W. (Ed.), Trends in Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ, pp. 166–205.

Maier, E., 1997. Evaluating a scheme for dialogue annotation. Verbmobil Report 193.

Mast, M., Maier, E., Schmitz, B., 1995. Criteria for the segmentation of spoken input into individual utterances. Verbmobil Report 97.

Niemann, H., Nöth, E., Kießling, A., Kompe, R., Batliner, A., 1997. Prosodic processing and its use in Verbmobil. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, München, Vol. 1, pp. 75–78.

Ostendorf, M., Veilleux, N., 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. Computational Linguistics 20 (1), 27–53.

Ostendorf, M., Price, P., Bear, J., Wightman, C., 1990. The use of relative duration in syntactic disambiguation. In: Speech and Natural Language Workshop. Morgan Kaufmann, Hidden Valley, PA, pp. 26–31.

Ostendorf, M., Wightman, C., Veilleux, N., 1993. Parse scoring with prosodic information: An analysis/synthesis approach. Computer Speech & Language 7 (3), 193–210.

Pitrelli, J.F., Beckman, M.E., Hirschberg, J., 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In: Internat. Conf. on Spoken Language Processing, Yokohama, Japan, Vol. 1, pp. 123–126.

Pollard, C., Sag, I., 1987. Information-based Syntax and Semantics, Vol. 1, CSLI Lecture Notes, Vol. 13. CSLI, Stanford, CA.

Price, P., Wightman, C., Ostendorf, M., Bear, J., 1990. The use of relative duration in syntactic disambiguation. In: Internat. Conf. on Spoken Language Processing, Kobe, Japan, Vol. 1, pp. 13–18.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., Fong, C., 1991. The use of prosody in syntactic disambiguation. Journal of the Acoustic Society of America 90, 2956–2970.

Reithinger, N., 1997. Personal communication.

Reyelt, M., 1995. Consistency of prosodic transcriptions. Labelling experiments with trained and untrained transcribers. In: Proc. 13th Internat. Congress of Phonetic Sciences, Stockholm, Vol. 4, pp. 212–215.

Reyelt, M., 1997. Personal communication.

Reyelt, M., 1998. Experimentelle Untersuchungen zur Festlegung und Konsistenz suprasegmentaler Einheiten für die automatische Sprachverarbeitung. Berichte aus der Informatik. Shaker Verlag, Aachen.

Reyelt, M., Batliner, A., 1994. Ein Inventar prosodischer Etiketten für Verbmobil. Verbmobil Memo 33.

Swerts, M., Geluykens, R., Terken, J., 1992. Prosodic correlates of discourse units in spontaneous speech. In: Internat. Conf. on Spoken Language Processing, Banff, Vol. 1, pp. 421–424.

Tropf, H., 1994. Spontansprachliche syntaktische Phänomene: Analyse eines Korpus aus der Domäne "Terminabsprache". Technical report, Siemens AG, ZFE ST SN 54, München.

Vaissière, J., 1988. The use of prosodic parameters in automatic speech recognition. In: Niemann, H., Lang, M., Sagerer, G. (Eds.), Recent Advances in Speech Understanding and Dialog Systems, NATO ASI Series F, Vol. 46. Springer, Berlin, pp. 71–99.

Wahlster, W., 1993. Verbmobil – Translation of face–to–face dialogs. In: Proc. European Conf. on Speech Communication and Technology, Berlin, Opening and Plenary Sessions, pp. 29–38.

Wahlster, W., Bub, T., Waibel, A., 1997. Verbmobil: The combination of deep and shallow processing for spontaneous speech translation. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, München, Vol. 1, pp. 71–74.

Wang, M., Hirschberg, J., 1992. Automatic classification of intonational phrase boundaries. Computer Speech & Language 6 (2), 175–196.

Wightman, C., 1992. Automatic detection of prosodic constituents. Ph.D. Thesis, Boston University Graduate School.