BMBF

Verbmobil

Verbundvorhaben

SYNTACTIC–PROSODIC LABELING OF LARGE SPONTANEOUS SPEECH DATA–BASES

A. Batliner, A. Kießling, R. Kompe, H. Niemann, E. Nöth

L.M.-Universität München

F.–A.–Universität Erlangen–Nürnberg

Vm

Report 131

Mai 1996

Mai 1996

A. Batliner, A. Kießling,
R. Kompe, H. Niemann,
E. Nöth

Institut für Deutsche Philologie
Ludwig–Maximilian Universität München
Schellingstr. 3
D–80799 München

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Tel.: (089) 2180 - 2916
e-mail: `Anton.Batliner@phonetik.uni-muenchen.d400.de`

Gehört zum Antragsabschnitt: 3.11, 3.12, 6.4

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Bildung, Wissenschaft, Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 F/4 und 01 IV 102 H/0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

SYNTACTIC–PROSODIC LABELING OF LARGE SPONTANEOUS SPEECH DATA–BASES

A. Batliner[1], *R. Kompe*[2], *A. Kießling*[2], *H. Niemann*[2], *E. Nöth*[2]

[1]Ludwig-Maximilians-Universität, Institut für Deutsche Philologie,
Schellingstr. 3, D-80799 München, F.R. of Germany
[2]Friedrich-Alexander-Universität Erlangen-Nürnberg,
Lehrstuhl für Mustererkennung (Informatik 5),
Martensstr. 3, D-91058 Erlangen, F.R. of Germany

ABSTRACT

In automatic speech understanding, the division of continuously running speech into syntactic chunks is a great problem. Syntactic boundaries are often marked by prosodic means. For the training of statistic models for prosodic boundaries large data-bases are necessary. For the German VERBMOBIL project (automatic speech–to–speech translation), we developed a syntactic-prosodic labeling scheme where two main types of boundaries (major syntactic boundaries and syntactically ambiguous boundaries) and some other special boundaries are labeled for a large VERBMOBIL spontaneous speech corpus. We compare the results of classifiers (multi-layer perceptrons and 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 could be labeled within a short time. Therefore, the classifiers trained with these labels turned out to be superior (recognition rates of up to 96%).

1. INTRODUCTION

The research presented in this paper has been conducted under the VERBMOBIL project (cf. [10]), which aims at automatic speech–to–speech translation in appointment scheduling dialogs. Syntactic boundaries are used for disambiguation during parsing. In spontaneous speech, many elliptic sentences or nonsentential free elements occur. Without knowledge of the prosodic phrasing and/or the dialog 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 VERBMOBIL 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 substituted 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.

1 *Ja? Zur Not geht's? Auch am Samstag?*
(Really? It's possible if necessary? Even on Saturday?)
2 *Ja. Zur Not. Geht's auch am Samstag?*
(Yes. If necessary. Would Saturday be possible as well?)

For such ambiguous turns, the use of prosodic information might be the only way to find the correct interpretation. But even for syntactically non–ambiguous utterances, the search space during parsing can be enormous, because locally it might not be decidable 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. [1].

2. PROSODIC OR SYNTACTIC LABELS

In written language, syntactic phrasing is indicated by word order; it 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. The following different types of perceptual–prosodic boundaries were labeled for 33 dialogs by the University of Braunschweig, cf. [8]:

- B3: full intonational boundary with strong intonational marking, often with lengthening
- B2: intermediate phrase boundary with weak marking
- B0: normal word boundary (not labeled explicitly)
- B9: "agrammatical" boundary (e.g., hesitation, repair)

There are some drawbacks in these boundary labels if one wants to use prosodic information in parsing: First, prosodic labeling by hand is very time consuming, the labeled data-base up to now is therefore rather small. Second, a perceptual labeling of prosodic boundaries is not an easy task and not very robust. Finally, prosodic boundaries do not only mirror syntactic boundaries but are influenced by other

*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 and 01 IV 102 H/0. The responsibility for the contents lies with the authors.

factors as rhythmic constraints and speaker specific style. In the worst case, clashes between prosody and syntax might be lethal for a syntactic analysis if the parser goes the wrong track and never returns.

Earlier experiments on a large corpus with read speech showed that syntactic–prosodic labels can be successfully used for the training of prosodic classifiers (cf. [6]). This result and the above mentioned problems motivated our colleagues from IBM (Heidelberg) to label pure syntactic boundaries only on the basis of syntactic criteria [3]. 25 dialogs were labeled, which are a subset of the turns labeled with the perceptual boundary labels. The developed labeling scheme distinguishes between 59 labels. Only syntactic boundaries ought to be labeled notwithstanding whether they are marked prosodically or not. The labels were assigned to word boundaries. Here, we only want to distinguish between the following main classes:

- S3+: for sure a syntactic boundary,
- S3–: for sure no syntactic boundary,
- S3?: Ambiguous boundary, i.e, based on the word chain it cannot be decided if there is a syntactic boundary*.

Acoustic–prosodic classifiers trained on the B or the S labels showed comparable recognition results, cf. [1].

3. SYNTACTIC–PROSODIC LABELS

These results and the urgent need for a larger training data-base for acoustic–prosodic classifiers and especially for syntactic–prosodic models encouraged us to develop a new labeling scheme with the following requirements:

- It should allow for fast labeling. Therefore the labeling scheme should be rather rough, because the more precise it is the more complicated and the more time consuming the labeling will be. A "small" amount of labeling errors can be tolerated, since it will be used to train statistical models, which should be robust to cope for these errors.
- Prosodic tendencies and regularities should be taken into account. In this context, 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. Since large quantities of data should be labeled within a short time, only expectations about prosodic regularities based on the textual representation of a turn (transliteration) can be considered.
- The specific characteristics of spontaneous speech have to be incorporated in the scheme.
- It should be independent of particular syntactic theories but at the same time, it should be compatible with syntactic theory in general.

According to these requirements, 7286 VERBMOBIL turns (17 hours of speech, 149514 word tokens counting word fragments but not non–verbals) were labeled by one person in about four months. An overview about the so called M labels is given in Table 1 where the context of the boundaries is described shortly, and the label and the main class it is attached to is given. Examples follow in Table 2 in the same order. Table 2 also shows the frequency of occurrence of the labels not counting the end of turns which by default are labeled with M3S.

*As for ambiguous boundaries cf. the M3A labels below.

In the experiments conducted so far, we distinguish only between the three main classes given in Table 1 that are for the time being robust enough and most relevant for the linguistic analysis in VERBMOBIL. Nevertheless, the distinction of the nine classes was considered to be useful, because their automatic discrimination might become important in the future. Furthermore, these boundary classes might be marked prosodically in a different way; cf. the short discussion below.

context	label	class
main/subordinate clause	M3S	M3
non–sentential free element/phrase, elliptic sentence	M3P	M3
extraposition	M3E	M3
embedded sentence/phrase	M3I	M3
pre–/ post–sentential particle with <pause>/<breathing>	M3T	M3
pre–/ post–sentential particle without <pause>/<breathing>	M3D	MU
syntactically ambiguous	M3A	MU
constituent, marked prosodically	M2I	M0
constituent, not marked prosodically	M1I	M0
every other word (default)	M0I	M0

Table 1: Overview over the M labels.

Syntactic main boundaries M3S are found between main clause and main clause, main clause and subordinate clause, and before coordinating particles between clauses. Boundaries at non–sentential free elements functioning as elliptic sentences are labeled with M3P. Normally, these phrases do not contain a verb. They might be idiomatic performative phrases with a sort of fixed meaning as *guten Tag* (*hello*) and vocatives, or they are "normal, productive" elliptic sentences as, e.g., *um vierzehn Uhr* (*at two p.m.*). With M3E we label boundaries between a sentence and a phrase to its right, which in written language normally would be inside the verbal brace. This phenomenon can be called extraposition or right dislocation with or without a pro element. M3E is also labeled at boundaries where for pure syntactic reasons, it should not be labeled, but where a pause etc. in the transliteration denotes a stronger separation from the clause to the left, e.g. in *Let's meet on Friday* M3E *<pause>* *the 9th*. Sentences or non–sentential free elements that are embedded in a sentence are labeled with M3I. Very often in spontaneous speech, a turn begins with pre–sentential particles, for example, with *ja*, *also*, *gut*, *okay*. These are either discourse particles with no specific meaning but having an important function as e.g. turn taking signals like *well* in English [4] or they are elliptic utterances functioning as, for example, a confirmation. The function is often marked by

label	example
M3S 11717	*vielleicht stelle ich mich kurz vorher noch vor* M3S *<Atmung> mein Name ist Lerch* *(perhaps I should first introduce myself* M3S *<breathing> my name is Lerch)*
M3P 4554	*<Atmung> guten Tag* M3P *Herr Meier* *(<breathing> hello* M3P *Mr. Meier)*
M3E 1409	*da hab' ich ein Seminar* M3E *den ganzen Tag* *(there I have a seminar* M3E *the entire day)*
M3I 369	*eventuell* M3I *wenn Sie noch mehr Zeit haben* M3I *'n bißchen länger* *(possibly* M3I *if you've got even more time* M3I *a bit longer)*
M3T 325	*gut* M3T *<Pause> okay (fine <pause>* M3T *okay)*
M3D 5150	*<Atmung> also* M3D *dienstags paßt es Ihnen* M3D *ja* M3S *(<breathing> then* M3D *Tuesday will suit you* M3D *isn't it / after all* M3S*)*
M3A 734	*würde ich vorschlagen* M3A *vielleicht* M3A *im Dezember* M3A *noch mal* M3A *dann* *(I would propose* M3A *possibly* M3A *in December* M3A *again* M3A *then)*
M2I	*wie sähe es denn* M2I *bei Ihnen* M2I *Anfang November aus* *(will it be possible* M2I *for you* M2I *early in November)*
M1I	M3S *hätten Sie da* M1I *'ne Idee* M3S (M3s *have you got* M1I *any idea* M3S)

Table 2: Parts of VERBMOBIL turns showing examples for the M labels and their frequency in the 7286 turns.

prosodic boundaries: pre–sentential particles that are followed by a pause or by breathing denoted in the transliteration are therefore labeled with M3T, all others with M3D. In post–sentential position, we label these words analogously, but not inside a clause or phrase. Syntactically ambiguous boundaries M3A cannot be determined solely based on syntactic criteria. Often there are two or more alternative word boundaries, where the syntactic boundary could be placed. It is therefore the job of prosody to disambiguate between two alternative readings. M3A and M3D labels are mapped onto the cover class MU ('undefined'), all other mentioned so far onto the cover class M3 ('strong boundary').
M2I and M1I denote constituent boundaries and are mapped onto the cover class M0, together with the default class M0I (any other word boundary). An M1I constituent boundary is in the vicinity of the beginning or the end of a clause and 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, and it is rather often marked prosodically, again because of rhythmic constraints. So far a reliable detection of M3 had priority, therefore, for the time being, M2I is only labeled in three dialogs, and M1I is not labeled at all.

4. CLASSIFICATION EXPERIMENTS: RESULTS AND DISCUSSION

We will now compare classification results obtained with a Multi-Layer Perceptron (MLP) and a Language Model (LM). The computation of the acoustic–prosodic features is based on an automatic time alignment of the phoneme sequence corresponding to the spoken or recognized words. In this paper, we only use the aligned spoken words thus simulating 100% word recognition. For each wordfinal syllable to be classified a vector of prosodic features is computed automatically from the speech signal. For the syllable itself and different syllables in the context the following features are considered (a total of 276): duration (+/– normalized); for F0, minimum, maximum, onset, and offset values, and their resp. relative positions on the time axis; for energy, minimum and maximum values, and their resp. relative positions on the time axis; linear regression coefficients for F0 and energy contours; length of the pause at boundary position; flags indicating whether the syllable carries a lexical word accent or whether it is in a word final position. The feature set is described in more detail in [5]. One Multi-layer perceptron (MLP) was trained to recognize the B labels based on the features and data as described above. In order to balance for the a priori probabilities of the different classes, during training the MLP was presented with an equal number of feature vectors from each class. For the experiments, MLPs with 40/20 nodes in the first/second hidden layer showed best results.

Trigram language models (LM) were additionally used for the classification of boundaries. They model word chains where the M3 boundaries have been inserted. This method as well as the combination of LM and MLP scores is described in more detail in [6].

In Table 3, we compare the results for different combinations of classifiers (MLP, LM for S-Labels: LM_S, and LM for M-Labels: LM_M) for the two main classes boundary vs. not–boundary 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 measured on the same test set comprising 3 dialogs (64 turns of 3 male and 3 female speakers, 12 minutes in total). For the training of the MLP and the LM_S all the available labeled data was used except for the test set (797 and 584 turns respectively) and 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 even for S3 vs. ¬S3 better than the LM_S because of the greater amount of training data. The LM alone are already very good; we have, however, to consider that they cannot be applied to the 'undefined' classes MU and S3? which are of course very important for a correct syntactic/semantic processing. Especially for these cases, we need a classifier trained with perceptual–prosodic labels. Due to the different a priori probabilities, the boundaries are recognized worse than the not–boundaries with the LMs; this causes the lower class–wise recognition rates (e.g., 80.8% for M3 vs. 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.

5. CONCLUDING REMARKS

A detailed analysis of correspondences and mismatches between the three types of boundaries is beyond the scope of this paper. In the following, we want to illustrate possible strategies for a more refined labeling and classification with one very simple example. Let us take the initiation of a dialog that often is done with greeting, as in example 3. For the Moment, we label M3P after *guten Tag* (*hello*), because the greeting need not necessarily be followed by the name of the dialog partner, cf. example 4, and because *guten Tag* (*hello*) is a typical free phrase. However, a M3P boundary as in example 3 is almost always not marked prosodically with a strong (B3) boundary. A sequence like *guten Tag, Herr Meier* occurs very often; it constitutes a dialog act and for the classification of dialog act boundaries, — another application of the M labels — it is here better not to have a boundary after *guten Tag* (*hello*). If we take contexts like these into account, we will achieve a better modeling of prosodic phrasing, and by that, a better classification of syntactic and dialog act boundaries.

3 *Guten Tag* M3P *Herr Meier.*
Hello M3P *Mr. Meier.*

4 *Guten Tag* M3P *Ich habe eine Frage.*
Hello M3P *I've got a question.*

Similar classification experiments of syntactic–prosodic boundaries are reported in [11, 7], where HMMs and classification trees were used. Our recognition rates are higher probably because of the large amount of training data. [11, 7] rely on perceptual–prosodic labels created on the basis of the ToBi system [9]. For such labels much less amounts of data can be obtained than in our case.

In the near future, we will further optimize the feature set and the classifiers. The boundary information achieved with our classifiers is already used in the VERBMOBIL project by the higher modules syntax [2], semantics, transfer, and dialog. The feedback based on results obtained with these modules and a parallel detailed error analysis will hopefully result in a further improvement of our labeling system and, in turn, an even more adequate use of prosodic information in the VERBMOBIL system.

cases	B3 vs. ¬B3 165 vs. 1284	S3+ vs. S3– 210 vs. 1179	M3 vs. M0 190 vs. 1259
MLP	87/87	85/78	87/83
LM_S	86/80	92/86	92/83
MLP+LM_S	89/85	93/87	93/86
LM_M	92/85	95/87	95/86
MLP+LM_M	94/89	94/86	96/89

Table 3: Percentage of correct classified labels for different combinations of classifiers: total vs. class–wise average

6. REFERENCES

1. A. Batliner, A. Feldhaus, S. Geissler, A. Kießling, T. Kiss, R. Kompe, and E. Nöth. Integrating Syntactic and Prosodic Information for the Efficient Detection of Empty Categories. In *Proc. of the Int. Conf. on Computational Linguistics*, Kopenhaven, 1996.
2. A. Batliner, A. Feldhaus, S. Geißler, T. Kiss, R. Kompe, and E. Nöth. Prosody, Empty Categories and Parsing — A Success Story. In *Int. Conf. on Spoken Language Processing*, Philadelphia, 1996.
3. A. Feldhaus and T. Kiss. Kategoriale Etikettierung der Karlsruher Dialoge. Verbmobil Memo 94, 1995.
4. J. Hirschberg and D. Litman. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3):501–529, 1993.
5. A. Kießling. Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Dissertation. Technische Fakultät der Universität Erlangen-Nürnberg, 1996.
6. R. Kompe, A. Kießling, H. Niemann, E. Nöth, E.G. Schukat-Talamazzini, A. Zottmann, and A. Batliner. Prosodic scoring of word hypotheses graphs. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 1333–1336, Madrid, 1995.
7. M. Ostendorf, C.W. Wightman, and N.M. Veilleux. Parse Scoring with Prosodic Information: an Analysis/Synthesis approach. *Computer Speech & Language*, 7(3):193–210, 1993.
8. M. Reyelt and A. Batliner. Ein Inventar prosodischer Etiketten für VERBMOBIL. Verbmobil Memo 33, 1994.
9. K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. TOBI: A standard for labeling English prosody. In *Int. Conf. on Spoken Language Processing*, volume 2, pages 867–870, Banff, 1992.
10. W. Wahlster. Verbmobil — Translation of Face–To–Face Dialogs. In *Proc. European Conf. on Speech Communication and Technology*, volume "Opening and Plenary Sessions", pages 29–38, Berlin, 1993.
11. M.Q. Wang and J. Hirschberg. Automatic Classification of Intonational Phrase Boundaries. *Computer Speech & Language*, 6(2):175–196, 1992.