BMBF

Verbmobil

Verbundvorhaben

Prosody, Focus, and Focal Structure: Some Remarks on Methodology

A. Batliner

L.M.-Universität München

Vm

Report 58

Dezember 1994

Dezember 1994

A. Batliner

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

Tel.: (089) 2180 - 2916
e-mail: `ue102ac@cd1.lrz-muenchen.de`

Gehört zum Antragsabschnitt: 3.11, 3.12, 6.4

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

Prosody, Focus, and Focal Structure: Some Remarks on Methodology

A. Batliner
L.M.-Universität München

Abstract

Prosody falls between several established fields as e.g. phonetics, phonology, syntax, and dialogue structure. It is therefore prone to misconceptions: often, its relevancy is overestimated, and often, it is underestimated. The traditional method in linguistics in general and in phonology in particular is the construction and evaluation of sometimes rather complex examples based on the intuition of the linguist. This intuition is replaced by more or less naive and thus non-expert subjects and inferential statistics in experimental phonetics but the examples, i.e. the experimental material, are often rather complex as well. It is a truism that in both cases, conclusions are made on an "as if" basis: as if a final proof had been found that the phenomenon A really exists regularily in the language B. In fact, it only can be proven that the phenomenon A sometimes can be detected in the production of some speakers of a variety of language B. This dilemma matters if prosody has to be put into practice, e.g. in automatic speech and language processing. In this field, large speech databases are already available for English and will be available for other languages as e.g. German in the near future. At least in the beginning, the problems that can – hopefully – be solved with the help of such databases might look trivial and thus not interesting – a step backwards and not forwards. "As if" statements (concerning, e.g., narrow vs. broad focus) and problems that are trivial at face value (concerning, e.g., the relationship between phrasing units and accentuation and the ontology of sentence accent) will be illustrated with own material. I will argue that such trivial problems have to be dealt with in the beginning, and that they can constitute the very basis for the proper treatment of more far reaching and complex problems.[1]

1 INTRODUCTION

For the study of the prosodic marking of focus, some phenomena must be accounted for:

[1]This contribution does not exactly live up to the normal standards of a paper. It is rather a summary of my oral contribution at the workshop; the transparencies are so to speak converted into text and explained in more detail. The bibliography is rather sketchy. This compromise was made in order to meet the deadline of the preproceedings.

- The place of the prosodic **phrase boundaries** in order to get the chunks of speech one has to analyse and thereby (hopefully) the focus domain
- The place of the prosodic **phrase accents** in order to get the most prominent part in the phrase (the focus exponent)
- The **manner** of the prosodic phrase accents in order to get more information on the focal structure and/or the special meaning

It is not always easy to get this information; at least the following **intervening factors** have to be controlled:

- There are "regular", **grammatical** factors as, e.g., sentence modality: e.g. in a question, the intonational shape of the focal accent can differ from that of a statement
- **Extra-grammatical** factors as, e.g., speaker idiosyncrasies, rhythm, tempo (isolating vs. integrating phrasing/accentuation) can heavily influence manner, number, and placement of boundaries and accents
- Spontaneous, "irregular", **a-grammatical** speech phenomena (hesitations, false starts, etc.) must be told apart from regular phenomena; this task is sometimes straightforward, sometimes not.

2 TWO DIFFERENT APPROACHES, AND A THIRD ONE

In this section, two rather traditional approaches will be characterized shortly: linguistics vs. phonetics. The third one that I want to contrast is applied research, i.e. automatic speech and language processing.

2.1 LINGUISTIC APPROACH

In this context, phonology is subsumed under linguistics. In its prototypical form, this approach goes like this: **the linguist** defines the intended focal structure, **the linguist** assumes / hypothezises place and form of the focal accent, **the linguist** produces the focal accent "in the right way" (either by speaking aloud or just by reasoning), and **the linguist** writes a rule / a theory that is consistent with these data. By that, it is possible that there are **no intervening factors** whatsoever, i.e. this approach can constitute a fully closed loop. It is characterized by Price/Hirschberg (1992) as follows: *"... linguistics, which has produced a volume of intuitive, anecdotal attributions of prosody's role in higher linguistic levels, such as pragmatics and discourse."*

This characterization might seem to be a bit unfair, because linguistics is of course not always anecdotal; moreover, it is sometimes based on "real" empirical data. The verification with a database along the lines of the two other approaches, cf. below, is, however, no integral part of this approach.

2.2 PHONETIC APPROACH

In the prototypical form of this approach, more stages are needed and more people are involved than in the linguistic approach: **the linguist** defines the intended focal structure and **the linguist** assumes / hypothezises place and form of the focal accent. A (more or less) naive **subject** comprehends the given focal structure (hopefully) and produces the focal accent "in the right way" (hopefully). **The phonetician** finds the produced focal accent e.g. via perception experiments and determines with the help of **the statistician** and on the basis of **instrumental measurements** the acoustic features that are relevant for the prosodic marking of boundaries and accents. There can be several **intervening factors** that are, however, kept constant as far as possible. This sort of approach is characterized by Price/Hirschberg (1992) as follows: *"[...] speech science, which has focused on the search for acoustic correlates of linguistic entities (such as stress and accent) in laboratory conditions."* (In this context, "speech science" is equivalent to "phonetics".)

2.3 THIRD APPROACH — AUTOMATIC SPEECH AND LANGUAGE PROCESSING

In the linguistic and in the phonetic approach, the database can most of the time be tailored by the researcher to suit his or her needs and that means to keep constant as many factors as possible. The database is very often rather small and consists in the prototypical case of **minimal pairs** because in this constellation, systematic differences show up more clearly. It is rather different (albeit not totally) in applied research in general and in automatic speech and language processing (henceforth ASLP) in particular: on the one hand, we are far from being able to process unrestricted speech data. On the other hand, applied research has sooner or later to be put into practice; that means that it has to cope with "real life" data. But "real life" data are "contaminated" with intervening factors and do unfortunately not consist of minimal pairs. I do not want to go into a detailed discussion of the two prominent antagonistic approaches – the knowledge based (close to linguistics) and the statistic one. (These approaches might hopefully converge in the future, cf. below the quotations of Price/Hirschberg 1992 and Ostendorf et al. 1993.) Nevertheless, it is a fact that for the time being, the statistic approach is more successful in ASLP.[2] As a consequence, large databases are needed in order to train and test the classifiers. Such databases exist for (American) English and will be available for other languages as e.g. German in the near future, cf. the VERBMOBIL–project (Wahlster 1993). Confronted with these databases, both the linguist and the phonetician face the same problem: they have to look for interesting data. The **pros** are the following: These databases contain productions of many speakers and many dialects; they contain more natural speech and are thus more representative for "real life" data. The **cons** result from the same fact: real life data contain a plethora of intervening factors; there might be no way to get the intended phenomena (as e.g. focal structure), there might be too much data mongering to get through, and there might be not enough "interesting" data (no minimal pairs!).

In basic research, data are chosen that can be used to solve problems that are asked by or inherent in the theory. The theory is prior and defines what is "interesting" or not. In

[2] But cf. Moore (1994:11): "In the end, statistics is just a sound mathematical approach for modelling uncertainty or *ignorance* [...]. When speech is fully understood, there may be very little residual uncertainty remaining to be modelled and the stochastic approach will have both served and lost its purpose."

a sense, it is exactly the other way round in applied research (ASLP). The final criterion is external to the theory: either a "dumb" measure of correct classification in percent or ultimately the judgment of the user – whether he or she is pleased or not (a sort of felicity criterion). No wonder that these two cultures do have some difficulty while communicating with each other. Sometimes, however, they are obliged to do that and I want to argue along the lines of Price/Hirschberg (1992) and Ostendorf et al. (1994)[3] that both can profit from that. I want to illustrate this statement below with some results taken from studies by myself and colleagues (Batliner 1994, Kießling et al. 1994a, and Kießling et al. 1994b).

3 A FLOW CHART FOR THE STUDY OF PROSODY

Table 1 tries to sum up the comparison of these three different approaches and to arrange them in a sort of diachronic/synchronic coordinate system. The **"diachronic aspect"** is given in the horizontal relation — from basic research in the eighties to applied research in the nineties.[4] The **"synchronic aspect"** is given in the vertical relation. In basic research, linguistics is virtually always prior to phonetics: the linguist formulates the question and the phonetician tries to find the answer.[5] In applied research, it can be both ways: one can again search for some acoustic correlates of e.g. focus, but actually, the procedure should be bidirectional: the input is a speech signal. It is analysed, prosodic features are extracted and serve as an input to a classification of e.g. boundary positions. This information is passed on to the higher linguistic levels (box 3 → box 4). But the higher levels can as well check ambiguities with the lower levels (box 4 → box 3).

[3] Price/Hirschberg (1992): "It is hoped that the existence of large labeled corpora can help bridge the traditional conflict between using data in which known sources of variability are strictly controlled (e.g. readings of isolated utterances in a laboratory environment) versus using naturally occurring data (which may form a sample too small and too variable to be used for anything other than impressionistic analysis)." Ostendorf et al. (1993): " [statistical techniques] are currently underutilized because of cultural differences among linguists, computer scientists and engineers. [...] they model variability (e.g. randomness due to incomplete knowledge of sources of variability) [...] automatic training methods exist [...] for [...] adapting the models to different speaking styles or domains. [...] they enable the use of large corpora which is important because human intuitions can under-represent the full range of prosodic structure. [...] we use "linguistics" to include both phonological models of abstract units (i.e. prosodic phrase constituents, prominence, and intonation markers) and phonetic hypotheses about their observed acoustic correlates (i.e. f0, duration, and energy). By "statistics", we mean both statistical data analysis and modeling techniques [...]"

[4] By that I do of course not mean that there has been no applied research in the eighties and vice versa no basic research in the nineties. The juxtaposition holds, however, for at least prosodic research in rough outline. Up to the end of the eighties, only a few studies were published on prosodic research in ASLP. Since then, things have changed almost dramatically.

[5] The phonetician can do the job of the linguist himself. But it is always a matter of the search for some acoustic correlates of linguistic entities.

Table 1: A flow chart for the study of prosody

basic research 1980-	$\Longrightarrow$	applied research 1990-
box 1 (linguistics)		box 4 (linguistics)
(very) small databases restricted vocabulary restricted syntax		same databases as in box 3
$\Vert$		$\Uparrow$
primacy of linguistics: search for acoustic correlates of focus $\Downarrow$		bidirectional search for syntactic/semantic correlates of accent/boundary $\Downarrow$
box 2 (phonetics)		box 3 (phonetics/speech science)
small databases minimal pairs		large databases less restricted vocabulary less restricted syntax

4 TWO PROTOTYPICAL STUDIES

In this section, I want to illustrate box 2 (traditional phonetic approach using minimal pairs) and box 3 ("applied" phonetic/speech science approach using large databases) from table 1 with data taken from some studies that are published elsewhere: Batliner et al. (1991), Batliner et al. (1994), Kießling et al. (1994a) and Kießling et al. (1994b).[6]

4.1 THE PROSODIC MARKING OF FOCAL STRUCTURE: WISHFUL THINKING OR HARD FACTS?

4.1.1 Material and procedure

We examined the prosodic form of four types of focus structure realized in two question (Q) types (declarative and inversion Q) and in two non-question (NQ) types (declarative and imperative sentence). The material consists of 3 different A.c.I.- constructions with a dependent transitive verb. Six untrained speakers (3 male, 3 female) produced a total of 360 sentences together with context sentences, which induced sentence modality, focal structure and thereby the focal accent (FA). The intended FA in the embedded sentence can be on the 2nd phrase (2PHR), the 3rd phrase (3PHR) or on both phrases (2/3PHR); cf. table 2. Here, we want to address the question whether the focal structures "double focus" and "broad focus" are really indicated by prosodic means or whether they have to be extracted out of the linguistic/situational context. For each utterance the following

[6]Studies by other scolars could of course do as well. The reasons for taking just these four are first that I am familiar with them and second that I think it is worth while to present them to a readership that might not be familiar with them because of the sometimes sparse communication between these two cultures — linguistics on the one hand and phonetics/speech science on the other hand.

Table 2: Examples of focal structure (focus underlined) and intended FA (capitalized); declarative sentence *She makes Nina weave the linen;* context sentences in English translation.

(Narrow) object focus FA on 2PHR:	What does the master make Nina weave? Sie läßt die Nina <u>das LEInen</u> weben.
(Broad) object/verb focus (focus projection),FA on 2PHR:	What does the master make Nina do? Sie läßt die Nina <u>das LEInen weben</u>.
Double focus FA on 2/3PHR:	What does the master make Nina do with which material? Sie läßt die Nina <u>das LEInen</u> <u>WEben</u>.
(Narrow) verb focus FA on 3PHR:	What does the master make Nina do with the linen? Sie läßt die Nina das Leinen <u>WEben</u>.

features for the 2PHR and the 3PHR were extracted and normalized; for details see Batliner (1989) and Batliner / Nöth (1989):

- The maximal and minimal fundamental frequency (Fo) values MAX and MIN, transformed into semitones and normalized with respect to voice register by subtracting the lowest Fo value of the speakers;
- The difference DIF of the position on the time axis of MAX and MIN in centisec.;
- The duration DUR in centisec. The normalization of the speaking rate took into consideration the average duration of that phrase for each speaker and the average duration of the syllables in the utterance;
- The maximal energy in the 0-5000 Hz band.

The parameter values were extracted "by hand" on mingograms and automatically from the digitized versions of the utterances. An average of 12 listeners participated in the following perception experiments. The test sentence was presented in isolation. The listeners had to decide which of the phrases carried the FA. If FAi is the number of listeners who perceived the ith phrase as most stressed then

$$FOK = (FA2\text{-}FA3)/(FA1+FA2+FA3)$$

takes on values between 1 (all listeners perceived the 2PHR as stressed) and -1 (all listeners perceived the 3PHR as stressed). FA on the 2PHR takes on values above 0.5 and FA on the 3PHR below -0.5. Double focus on the 2/3PHR is defined operationally as $|FOK| < 0.5$, i.e. those items - about 25% of the whole corpus - where the subjects are rather uncertain about the place of the FA. Note that this value is in a way arbitrary, and that it is not a strict definition, but rather an "in these cases it is likely that..."-way of defining the focal structure. The results of a statistical classification procedure (discriminant analysis) will be reported for two different learn and test constellations:

l=t: All utterances were used for learning and testing with learn=test. This is the "best possible" constellation, i.e. it provides an upper limit for the predictive power of the variables, but overadaption is likely.

Table 3: Crosstabulations

a)	Qs (n=172) PERCFA			NQs (n=188) PERCFA		
INTFA	2	2/3	3	2	2/3	3
2	49	31	1	69	3	0
2/3	21	26	3	57	16	2
3	4	9	28	6	9	26

b)	Qs (n=172) PREDFA			NQs (n=180) PREDFA		
INTFA	2	2/3	3	2	2/3	3
2	42	39	0	62	8	0
2/3	22	22	6	54	11	5
3	4	5	32	5	13	22

c)	Qs (n=172) PREDFA			NQs (n=180) PREDFA		
PERCFA	2	2/3	3	2	2/3	3
2	43	30	1	115	12	0
2/3	23	33	10	6	14	6
3	2	3	27	0	6	21

l5t1: As a training sample we used 5 speakers, and the remaining speaker as the test sample (leave one out). This simulates speaker independence and avoids overadaption.

Since a separate treatment of Qs and NQs yields better results than when analysed together, only these results will be discussed.

4.1.2 Results and discussion

In experiments like ours, the linguist defines the intended focal structure and thereby place and (possibly special) form of the FA. The subjects must comprehend the given focal structure and produce the FA 'in the right way" . The produced FA should be judged with perception experiments as described above, because only then can we be sure that misproductions are filtered out. The acoustic parameter values can be used to predict (PREDFA) the perceived FA (PERCFA) as well as the intended FA (INTFA). The mapping from one step to another is never optimal. Table 3[7] shows for l=t and separated into Qs and NQs, 3 different crosstabulations. All variables were used as predictors. To give an example, the first 3 numbers in table 3 b) read as follows: 81 Qs had the INTFA on the 2PHR; 42 out of the 81 had a PREDFA on the 2PHR, the rest on the 2/3PHR.

The following points shall be discussed briefly:

- Double focus (INTFA on 2/3PHR) is not marked very often prosodically. It follows that subjects do not necessarily indicate double focus by prosodic means. At least for the rhythmical structure and the linguistic and non-linguistic context of our test sentences, the two ways of expressing double focus might be free variants, in the case of the FA on the 2PHR a sort of pseudo projection. Of course, the subjects

[7]For b) and c) in table 3, the sum of the NQs is only 180, because 8 items could not be predicted for technical reasons.

Table 4: Classification errors in percent

	Qs		NQs	
	l=t	l5t1	l=t	l5t1
single foci	5	14	4	6
triple foci	40	46	17	27
clear foci	2	2	0	2

might simply not have understood the intended focal structure. However, this is not very likely because in other perception experiments, where listeners had to judge the naturalness of the items, double focus items with the FA on the 2PHR did not get worse scores than those with the FA on the 2/3PHR.

- There is a greater confusion between 2PHR and 2/3PHR for Qs than for NQs. The reason might be that in Qs, the Fo offset is mostly high, and that intensity covaries to a certain extent with rising Fo. On the one hand, intensity is not relevant for Qs, on the other hand, this covariation might puzzle the listeners that much that their judgments are more uncertain and fall below the limit of FOK=0.5.
- The mapping PERCFA-PREDFA is best as expected, because here, perception is directly related via the acoustic features with the classification.
- Separation of verb focus vs. the other foci is best, i.e. separation of "clear" single foci is very good, cf. table 4. In this table, percentages of errors are given for 3 different constellations:
  - Prediction of single foci in Batliner (1989) and in Batliner/Nöth (1989) with the border between FA on the 2PHR and FA on the 3PHR at FOK=0.0;
  - Prediction of "triple foci" (FA on 2PHR, 2/3PHR, 3PHR) as in table 3 c);
  - Prediction of "clear" single foci, i.e. the confusion rate between FA on the 2PHR and on the 3PHR in table 3 c).

As for the Qs, the above mentioned covariation of intensity and Fo might be the reason for the marked difference of 12% between "single foci" and "clear foci" for l5t1 in table 4. It follows from this table, that for automatic speech recognition, it might be suitable not only to predict the FA, but also to try to predict clear FA in order to eliminate wrong hypotheses with a high probability.

4.1.3 Narrow vs. broad focus

Here, we will only report results for the NQs, because in Qs, the simultaneous marking of sentence modality and FA renders the discussion of the (poor) classification rate even more difficult. It can be seen in table 5, that the "realistic" recognition rate, 63% for l5t1 with no over-adaption, is rather low. That does not necessarily mean that narrow vs. broad focus is not indicated at all by prosodic means: because the sample size is rather

Table 5: Recognition rates for narrow vs. broad focus

	all features, stepwise selection	three most relevant features
l=t	73	65
l5t1	63	63

small (n=72), a few misproductions can influence the result markedly.[8] Besides that, a close inspection of the individual speakers indicates a speaker specific use of the variables. Nevertheless, the mean difference of the 3 most relevant features DUR on the 2PHR and DIF on the 2PHR and the 3PHR can be interpreted: the values of MAX and MIN are almost identical. For narrow focus however, DIF is greater on the 2PHR and smaller on the 3PHR than for broad focus, i.e. the slope is less steep on the 2PHR and steeper on the 3PHR. If long inflections are judged to be of greater impact than short ones of similar rate, for narrow focus, the 2PHR is marked more clearly than the 3PHR. The same applies to DUR on the 2PHR (mean value 3.47 for narrow and 3.22 for broad focus). Note that in another perception experiment ((Batliner 1989, 30), the position of the FA was equally distributed on 2PHR (80%) and on 3PHR (20%) for narrow and for broad focus. If these two structures are marked differently at all, it may be by features (as DIF) that are rather irrelevant for the marking of the FA in NQs.

4.2 CONCLUDING REMARKS ON THE PHONETIC APPROACH (box 2)

In the study reported in the previous section, all the stages described in 2.2 can be found and are necessary in order to find the answer to the question: is the prosodic marking of focal structure wishful thinking or hard fact? The answer might be: it is most certainly neither – nor. Some less intricate focal structures are more likely to be marked with prosodic means. A clear prosodic marking of focal structure is obviously no must, but if it is clear, it is reliable. We doubt, however, that these results – especially those concerning the more intricate structures – can simply be "mapped" onto "real life" data because of the following three conditions that hold for laboratory speech:

- In laboratory speech, the functional load on prosody is very heavy compared with real life data (awareness of the subjects, prosodic minimal pairs).
- An estimation of intervening factors is not possible.
- Experimental databases and experimental verification is time consuming and expensive; the data are therefore often reused. Thereby, an overadaptation is possible.[9]

[8] Of course, it could be the other way round as well: if more speakers can be analysed this slight difference might no longer show up at all.

[9] That means that the statistical procedure so to speak "learns by heart" the data. Strictly speaking, normal inferential statistics as it is often used by phoneticians is then not allowed without modification of the error level. This modification is , however, almost never made.

4.3 AUTOMATIC LABELING OF ACCENTS AND BOUNDARIES AND THEIR PERCEPTUAL EVALUATION

4.3.1 Material

The material we investigated is the German speech database ERBA, "**Er**langer **B**ahn **A**nfragen" (Erlangen train inquiries) a large speech training database for word recognition in the domain of train table inquiries. A stochastic sentence generator was used based on a context free grammar and 38 sentence templates to create a large text corpus. At four different sites a subset of 10,000 unique sentences was recorded in quiet office environments (100 untrained speakers, 100 utterances each) resulting in a speech database of about 14 hours. The speakers were given the word sequences with punctuation marks; for more details concerning ERBA see Batliner et al. (1994).

The set of 100 speakers was partitioned into the following three subsets: 69 speakers (44 male, 25 female, 6,900 sentences) for training, 21 speakers (12 male, 9 female, 2,100 sentences) for testing, and the remaining 10 speakers for perception tests and also for testing.

4.3.2 Perception Experiments

The perception experiments were conducted in order to get reference labels for prosodically marked phrase boundaries and accentuated syllables. This information is used to improve the automatic generation of boundaries and accents in an iterative process of generation and control. Ten "naive" listeners were given 500 utterances[10] from 10 speakers (5 male, 5 female, 50 utterances each) in orthographic form without any punctuation marks. In a first experiment their task was to mark the space between two words if they felt it separated two different "chunks" of speech. In a second experiment another group of ten "naive" listeners was asked to mark each syllable they perceived as stressed. Thus, each possible accent position (= syllable) and each possible phrase boundary position (= word boundary) got a perception score from 0 (no mark) up to 10 (all 10 subjects in the test perceived an accent or a phrase boundary as marked). The listeners were instructed not to rely upon their knowledge of canonical forms or sentence structure, although influence of these factors can certainly not be ruled out altogether.

4.3.3 Automatic Generation of Phrase Boundaries and Accents

Phrase boundary labels as prerequisite The automatic generation of phrase accents is based on the automatically generated phrase boundary markers described in Batliner et al. (1994): Syntactic boundaries were marked in the grammar and included in the sentence generation process with some context-sensitive post-processing. The result is the orthographic word chain separated by boundary labels. We distinguish four types of phrase boundaries: boundary B3 is placed between elliptic clause and clause or between main and subordinate clause, B2 is positioned between constituents or at coordinating particles

[10] For the perception tests only sufficiently long and semantically meaningful sentences were used: When generating sentences with a context free grammar "nonsense" sentences like *"between ten and ten o'clock"* can not be avoided. The intonation of such sentences might be irregular, even hesitations may occur, which can be the reason for "miss"-classification. Since ERBA initially was intended to train word recognizers such "nonsense" sentences were not discarded.

between constituents, B1 belongs syntactically to the normal constituent boundary B2 but is most certainly not marked prosodically because it is close to a B3 boundary or to the beginning/end of the utterance, and B0 is any other word boundary that does not belong to B1, B2, B3; an example is given below..

For the assignment of accents, it has to be decided which words in an utterance are accentuated. In words with more than one syllable, normally one of these syllables bears the word accent; this syllable can be looked up in the lexicon. Factors that might influence whether or not a word is accentuated include the form class of a word (content word: CW vs. function word: FW), its position in a larger prosodic context, and tempo (isolating vs. integrating accentuation). Rhythmic constraints can influence the location of accent within a word. In order to take into account most of these factors the automatic generation of accent labels was iteratively controlled with and adapted to the results of the perception experiments.

Assigning the lexical word accent Before creating the accent labels, we first compared the word accents marked in the lexicon with the results of the perception experiments in order to derive rules for the position of the phrase accent. For the labeling of the accents in the lexicon we decided in favor of a rather broad labeling, i.e. we only distinguish accentuated from unaccentuated syllables. Secondary accentuation is not labeled because in a canonical citation pronunciation, these differences might be produced and perceived systematically but not in a more casual pronunciation as is the case in fluent speech. In the lexicon, the 75 FWs (articles, pronomina, auxiliary verbs, prepositions, conjunctions) were not marked as accentuated. They are normally clitic i.e. without accent and integrate with the following constituent into a greater prosodic phrase. In general, CWs are represented with just one accentuated syllable (word accent). If more than one accentuation is possible without change of the meaning as e.g. in some proper nouns and longer words (*"Erlangen"* and *"zweiundzwanzig"* respectively with accent on the first **or** on the penultimate syllable) both positions are marked in the lexicon.

Assigning the accent label to a word within a phrase The next step was to decide which words within a phrase are accentuated. Since for the moment we do not consider emphatic or contrastive accents we assume that in each prosodic phrase (bounded by B1, B2, or B3[11]) one and only one word is more prominent than the others. In German, the phrase accent is normally positioned on the rightmost CW in a NP (*"rightmost principle"*); in a PP and in a VP, by default the argument is the carrier of the phrase accent, i.e. not the preposition or the verb.

The examination of the perception scores showed in some cases additional tendencies not to put stress on "semantically weak" CWs or to put stress on "strong" FWs. In the following example the syllables to be expected as stressed are typed bold: *ich möchte* B1 *am nächsten* **Diens***tag* B2 *zwischen* **drei** B2 *und* **sechs** *Uhr* B2 *von* **Ham***burg* B2 *nach* **Ulm** B1 *fahren* (*I would like* B1 *next Tuesday* B2 *between three* B2 *and six o'clock* B2 *from Hamburg* B2 *to Ulm* B1 *to go*). This example contains the two most important exceptions: The word *"Uhr"* and other CWs like e.g. verbs such as *"fahren"* that are rather predictable in the domain of train table inquiries and therefore semantically weak or

[11]Note, that for the generation of the accent labels also the beginning and the end of an utterance is assumed to be a B3 boundary.

clitic, are usually not accentuated and thus got a rather low perception score. Therefore, in the last two phrases not the verb *fahren* but the city name *Ulm* is expected to be stressed. On the other hand, often FWs with a rather high perception score could be observed, e.g. interrogative pronouns such as *"was"*, *"wann"*, *"welche"* that obviously are semantically and pragmatically strong words in this domain. The semantic weakness of the verb coincides with the above mentioned rule that verbs by default are not accentuated. There are, however, exceptions, as, e.g. the so called particle verbs like *"ankommen" (arrive)* and *"abfahren" (leave)* that might be accentuated.

Based on these observations the following rules for our algorithm were formulated: For each phrase bounded on the right by symbol Bx ($x \in \{1,2,3\}$) look successively for the rightmost CW*[12], or (if not found) for the rightmost verb, or for the word "Uhr" , or for an interrogative pronoun, or for an auxiliary verb, or for any other word and mark the first instance by symbol `Ax` (where `x` corresponds to x in Bx). After applying this rule to a sentence, in each phrase one and only one word is marked by an accent label `Ax` ($x \in \{1,2,3\}$).
In order to take into account that there are semantically weak words occurring in short phrases before a B3 boundary, we have to add another rule: If the actual word is not a CW* and the phrase is bounded on the left by B1 and on the right by B3 and there is a CW* on the left hand side of the B1 boundary then exchange the accent labels of these two words. This rule e.g. changes *"...nach* `A1`*Ulm* B1 `A3`*fahren* B3*"* into *"...nach* `A3`*Ulm* B1 `A1`*fahren* B3*"*.
Special treatment is necessary for certain compound words that occur very frequently in our application (e.g. city names). In our lexicon these words are characterized by a linking hyphen or dash. Following the rightmost principle, we marked the rightmost word by `Ax`, and all other words of the compound word by `Axi`, denoting that there is an "implication" from left to right, i.e. if any word of the compound word is stressed, all its right hand neighbors are stressed as well. It has to be noted, that this rule is rather straightforward and does not take into account other possibly relevant factors as, e.g. rhythmic constraints.

Assigning the accent label to a syllable within a word After the accent labels are assigned to the words, we have to determine the syllables within the words bearing the accent. In our material, this assignment depends on several factors:

- If the word has only one syllable marked as the (lexical) word accent in the lexicon, this syllable inherits the symbol `Ax` from the word.
- If more than one syllable can be accentuated, all these syllables get the symbol `Axa`, denoting that they are real alternatives, and that it is at discretion of the speaker which of those alternatives actually is stressed.
- If there is no lexical accent at all for this word (which is usually the case for FWs) the first syllable in the word gets the symbol `Axn`, denoting that it is just a default (root) accent[13].

[12] CW* denotes in our context any word that is not a FW, verb, auxiliary verb, interrogative pronoun or the word "Uhr" .

[13] This simple rule can of course not be applied to all German FWs but it works reasonably well within our lexicon.

These rules apply in the same way to single words and to the parts of the compound word marked by "implication" labels. For example, the syllables of the greeting *Grüß_Gott* are labeled with [Axi Ax] and the city name *Riebnitz-Damgarten-West* is labeled with [Axi A0 Axi A0 A0 Ax].

To take into account that syllables positioned directly before a phrase boundary are usually produced differently from others due to phrase final lengthening, we introduced additional markers. Another reason for labeling these syllables in a special way is that at present we are also investigating the combined recognition of phrase boundaries and phrase accents based on syllables (cf. Kießling et al. 1994a) as well as the modeling of phrase structures by Hidden Markov Models. Therefore, if one of the already marked accentuated syllables is positioned directly before a phrase boundary marked by Bz ($z \in \{1,2,3\}$) it gets the additional label +Bz. All the remaining (unaccentuated) syllables in the sentence are labeled with Bz if they are positioned directly before a phrase boundary marked by Bz, otherwise they are marked as A0.

By applying all these rules to the whole ERBA database of 10,000 sentences in total 199,078 syllables were marked by 30 different symbols.

Comparison of the generated labels with the listeners judgments The perception data were compared with the automatically labeled places of phrase boundaries and phrase accents. Each possible position (word boundary position for phrase boundaries and syllable position for phrase accents) could get a score from 0 (no mark) up to 10 (all ten subjects marked the position). The 500 utterances contain 71 types of FWs with 3346 tokens and 588 types of CWs with 3396 tokens. FWs got an average score of 1.4 with a minimum of 0 and a maximum of 8; 10% were above 5 and 36% above the mean. CWs got an average score of 7.4 with a minimum of 0 and a maximum of 10; 14% were less than 5 and 47% were less than the mean.

In Figure 1, the frequencies of the perceptual scores for boundaries and accents are plotted. The curve for the accents is V-shaped with a turning point at 5, that is in the middle of the scale. It thus makes sense to define syllables with a score higher than 5 as accentuated. For the phrase boundaries, the curve is U-shaped with no clear turning point. We assume that our boundary labels fall not into two but into three distinct classes: B01, B2, B3 (cf. also Batliner et al. 1994). It thus makes sense to define two turning points: B01 below 3, B3 above 8, and B2 in between. (The assumed thresholds are marked in Figure 1 by vertical lines.) This last assumption is supported by the relationship between accent and boundary scores illustrated in Figure 2: the abscissa represents a threshold M, partitioning the perceived accent scores (pas) into two classes: if $\mathsf{pas} \geq \mathsf{M}$, the syllable is defined to be accentuated, otherwise it is not accentuated. Each of the curves (marked with $\mathsf{N} \in [0;10]$) represents a threshold, partitioning the perceived boundary scores (pbs) into two classes: if $\mathsf{pbs} \geq \mathsf{N}$, the word boundary is defined to be a phrase boundary. The cross plotted indicates M=6, N=3 and an ordinate value of about 1; i.e. the mean value of the number of accent scores higher than 5 within a phrase bounded by a boundary with a perceptual score higher than 2 is about 1.

Usually it is assumed that in each phrase there is one prominent syllable, represented, e.g., in the tone sequence approach by one starred tone. As illustrated here, by setting the M threshold for the accent scores to 6, the N threshold for the boundary scores to 3, this assumption is supported pretty well by our empirically obtained perception data: the mean value of the number of accented scores is roughly 1, i.e. for each phrase defined in that way there is on the average one prominent syllable that can be defined as the carrier of the phrase accent. As can be easily seen these phrases correspond to the constituents that are marked by B2 boundaries (cf. Batliner et al. 1994). (This is of course no "prove" but rather a sort of cumulative evidence.)

In figure 3, the results of the perception experiments are given for the four different boundary types. The distributions of the B0, B1, and B3 boundaries meet our expectation and cluster at the left end (very few scores for B0 and B1 boundaries) or at the right end (many scores for B3 boundaries). Most probably, clause boundaries e.g. can thus be successfully handled in ASLP. The B2 boundaries behave differently, only 63% are above a score of 4 subjects. It might be at the discretion of the speaker if he/she wants to mark these boundaries. In 92% of the cases where at least 5 listeners perceived a boundary there was an automatically generated reference boundary (B2,B3). Also in 92% of the cases where less than 5 listeners perceived a boundary there was no automatically generated reference boundary (B0,B1). This and the fact that three of the four boundary classes in figure 3 are clear-cut and meet our expectations leads us to the conclusion that the automatically generated reference boundaries are adequate and can thus be used to train and test classifiers.

For the comparison of the generated accent labels with the listeners judgments the critical cases (i.e. the "alternative" and the "implicated" accents) are not taken into consideration and the original 30 accent symbols are mapped onto five accent types: A1, A2, and A3 denote phrase accents corresponding to the phrases of type B1, B2, and B3; B denotes unaccentuated syllables immediately preceding a phrase boundary, A0 any other (unaccentuated) syllable. In figure 4, these five accent types are cross-classified with the listeners judgments. The scores for A0 and B, i.e. the unaccentuated syllables meet our expectation: 90% of the A0 and more than 91% of the B syllables were perceived as stressed by less than 2 listeners. The accent types A2 and A3 clearly cluster at the right end although the tendency is not as distinct as for the corresponding phrase boundaries B2 and B3 (cf. Figure 3). The accent type A1 (word accent syllable in a prosodically "weak" constituent) is obviously marked more often than A0 (unaccentuated syllable). Note that the A3 scores are not markedly higher than the A2 scores. It is often assumed that the sentence accent in German is by default the rightmost phrase accent in an utterance (A3 accent in our material) and more prominent than any other phrase accents (A2 accents in our material). Our result might be taken as an argument against a phonetic manifestation of sentence accent in German.

4.4 CONCLUDING REMARKS ON THE PHONETIC / SPEECH SCIENCE APPROACH (box 3)

In the studies reported in the previous section, different assumptions were investigated:

- An "old, established" assumption could be corroborated: each prosodic phrase contains (on the average!) one phrase accent.
- On the other hand, an assumption that possibly is just as old could be questioned: the assumption that in a sentence, the rightmost phrase accent is, being the sentence accent, marked more clearly by prosodic means than any other phrase accent.

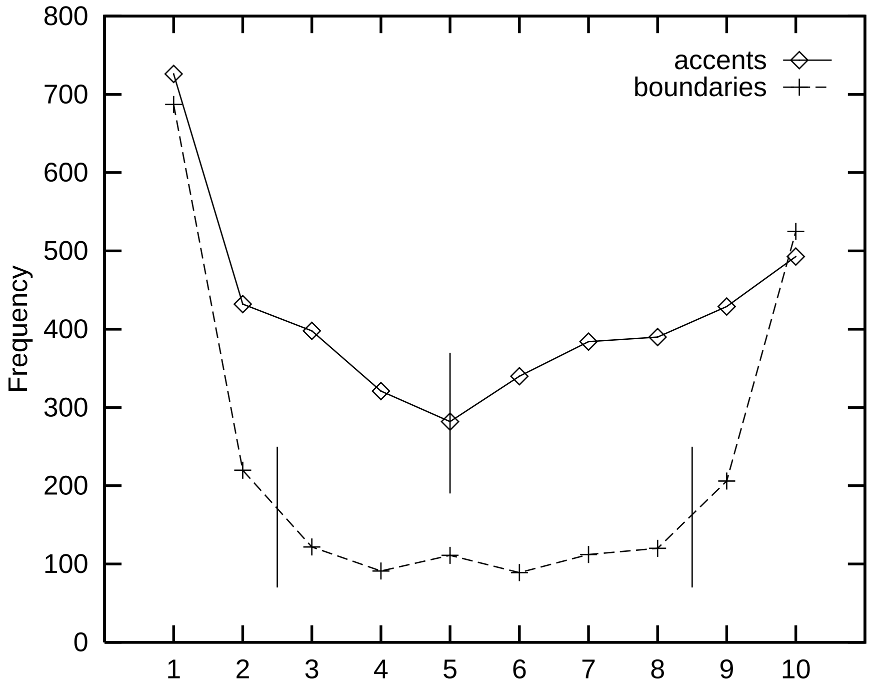

Figure 1: Frequency of accent and boundary types

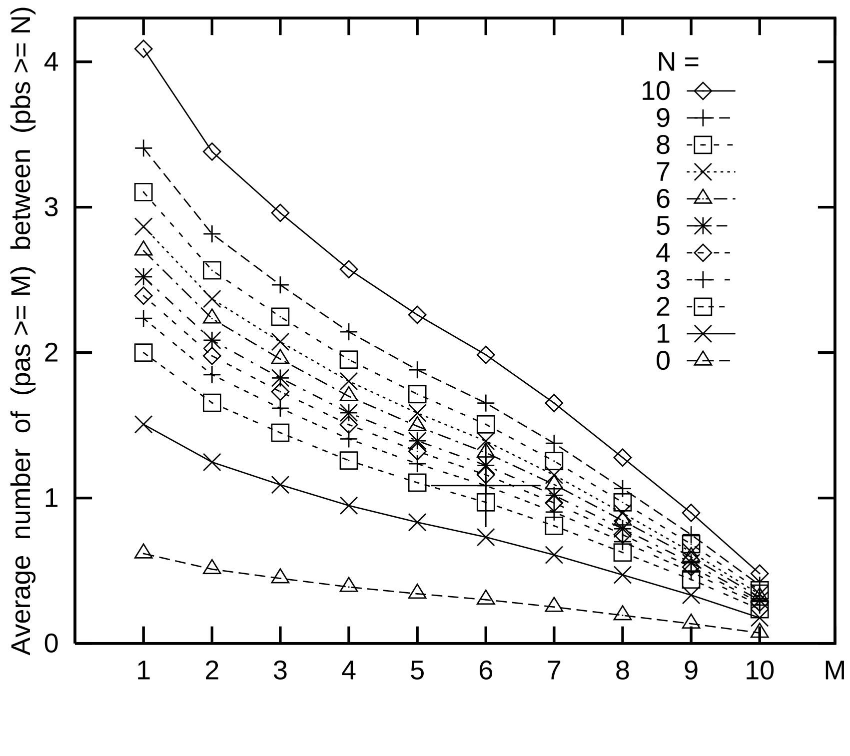

Figure 2: Relation between accent and boundary scores

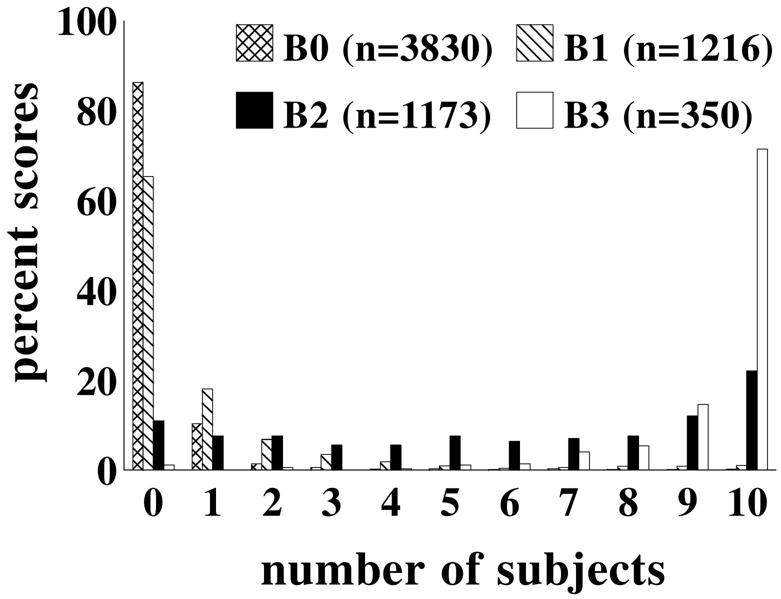

Figure 3: Frequency in % of scores for boundary types

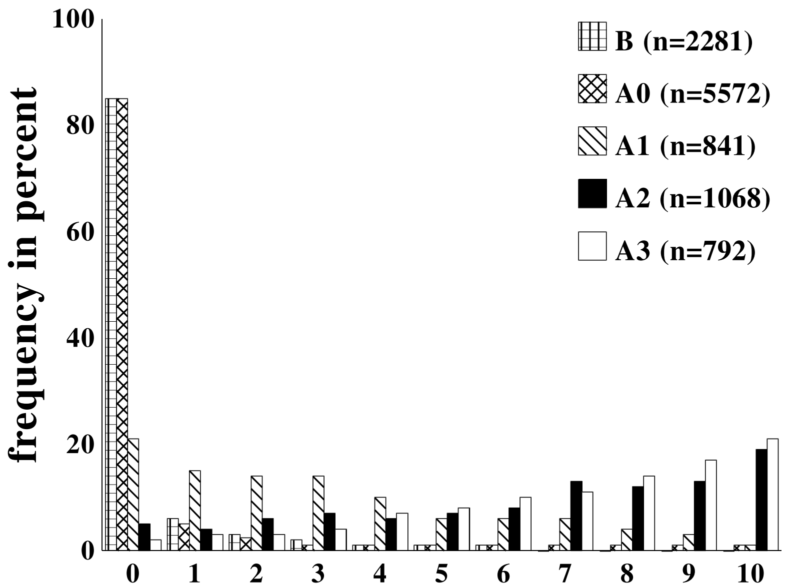

Figure 4: Frequency in % of scores for accent types

- An assumption that might be rather new could be validated: the assumption that "normal" constituent boundaries are not marked prosodically if they are in the vicinity of higher (clause) boundaries.

- The question whether "normal" constituent boundaries are marked prosodically if they are not in the vicinity of higher (clause) boundaries could not be answered unequivocally. That means most certainly, however, that there is simply no straightforward answer because of the influence of several intervening factors.

Note that the database contains read speech and not spontaneous speech. We are thus only half way towards "real life" data. The advantage of these controlled, read data is, however, that prosodic accents and boundaries could be labeled automatically. The 500 items in the perception experiments served so to speak as a handlabeled subsample for the evaluation of hypotheses. For the **automatic** classification of a very large subsample of the ERBA database, for distinguishing three different boundary classes a recognition rate of 75.7% and for distinguishing accentuated from unaccentuated syllables a recognition rate of 88.7% could be achieved so far. If we consider the fact that the automatic classification of prosodic phenomena in large databases has only been investigated for a rather short period of time in contrast to, e.g. phone or word recognition, these recognition rates can be taken as a strong argument in favor of the conclusions made on the basis of our perceptual evaluation.

5 CONCLUDING REMARKS

The development in ASLP tends to go towards large databases and towards spontaneous speech. Factors that can be kept constant in laboratory speech must first be accounted for in these "real life"data. Phenomena that are interesting from a theoretical point of view do not occur that often in "real life" speech or can simply not be found. In my opinion, the investigation of the prosodic marking of the following phenomena that traditionally are "object of investigation" is possible and useful for the time being:

- chunks of speech (prosodic boundaries)

- most prominent parts (accents)

- questions vs. non-questions

In addition, "new" phenomena must be accounted for, e.g. the influence of intervening factors (e.g. rhythmic constraints) on grammatical phenomena as well as irregular phenomena. Less possible, however, might be the investigation of "more interesting" (i.e. more complicated) phenomena as, e.g., PP-attachment and focal structure (broad vs. narrow, double focus).

We are faced with an increasing complexity of the material. If we are lucky, it might be accompanied with less complexity of the relevant linguistic phenomena: no complete deep analysis and disambiguation of all possible meanings, but in most cases, only a "flat" analysis step by step (left to right) might be necessary and only in case of conflict/ambiguity, a deeper (re-) analysis. For this enterprise it might, however, be necessary for linguistics and phonetics to redefine the notion of "interesting problem".[14]

REFERENCES

Batliner, A. (1989), Fokus, Modus und die große Zahl. Zur intonatorischen Indizierung des Fokus im Deutschen, in Altmann, H. / Batliner, A. / Oppenrieder, W. (eds.) (1989): *Zur Intonation von Modus und Fokus im Deutschen*, Tübingen: Niemeyer, 21-70.

Batliner, A. / Nöth, E. (1989), The Prediction of Focus, in *Proc. European Conf. on Speech Communication and Technology*, volume 1, Paris, 210-213.

Batliner, A. / Oppenrieder, W. / Nöth, E. / Stallwitz, G.(1991), The Intonational Marking of Focal Structure: Wishful Thinking or Hard Fact? in *Proc. XIIth Int. Cong. of Phonetic Sciences*, volume 3, 278-281.

Batliner, A. / Kompe, R. / Kießling, A. / Nöth, E. / Niemann, H. / Kilian, U. (1994), The prosodic marking of phrase boundaries: Expectations and results, in A. Rubio (ed.), *New Advances and Trends in Speech Recognition and coding*, NATO ASI Series F, (to appear), Springer Verlag, Berlin, Heidelberg, New York.

Kießling, A. / Kompe, R. / Niemann, H. / Nöth, E. / Batliner, A. (1994a), Detection of Phrase Boundaries and Accents, in H. Niemann / R. de Mori / G. Hanrieder (eds.), *Progress and Prospects of Speech Research and Technology*, infix, Sankt Augustin, 266-269.

Kießling, A. / Kompe, R. / Batliner, A. / Niemann, H. / Nöth, E. / (1994b), Automatic Labeling of Phrase Accents in German. To appear in *Proc. ICSLP-94*.

Moore, R. (1994), Twenty Things We Still Don't Know about Speech, in H. Niemann / R. de Mori / G. Hanrieder (eds.), *Progress and Prospects of Speech Research and Technology*, infix, Sankt Augustin, 9-17.

Ostendorf, M. / Price, P. / Shattuck-Hufnagel, S. (1993), Combining Statistical and Linguistic Methods for Modeling Prosody, in D. House / P. Touati, *Proceedings of an ESCA Workshop on Prosody*, Lund, Sweden, 272-275.

Price, P. / Hirschberg, J. (1992), SESSION 13: PROSODY – Introduction, in *Speech and Natural Language Workshop*, Morgan Kaufmann, 4 pages.

Wahlster, W. (1993), Verbmobil – Translation of Face–to–Face Dialogs, in *Proc. European Conf. on Speech Communication and Technology*, volume "Opening and Plenary Sessions", Berlin, September 1993, 29-38.

Anton Batliner
Universität München
Institut für Deutsche Philologie
Schellingstr. 3
80799 München
email: ue102ac@cd1.lrz-muenchen.de

[14]This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research projects ASL and VERBMOBIL and by the *Deutsche Forschungsgemeinschaft* (*DFG*). Only the author is responsible for the contents of this paper.