Eurospeech 89

European Conference on Speech Communication and Technology

Paris - September 1989

Volume One

Editors: J P Tubach, J J Mariani

Scientific Co-Sponsors:

Association Belge des Acousticiens (ABAV), Belgium
Association pour la Recherche Cognitive (ARC), France
Commission of the European Communities (CEC), Directorate General for Telecommunication, Information Industries and Innovation
Centro di Studio per le Ricerche di Fonetica del CNR (CSRF-CNR), Italy
Centre National de la Recherche Scientifique (CNRS), France
Deutsche Arbeitsgemeinschaft fuer Akustik (DAGA), FRG
Deutsche Arbeitsgemeinschaft fuer Mustererkennung (DAGM), FRG
European Association for Signal Processing (EURASIP)
Gesellschaft fuer Angewandte Linguistik (GAL), FRG
Gesellschaft fuer Informatik (GI), FRG
Gesellschaft fuer Linguistische Datenverarbeitung (GLDV), FRG
Gruppo di Fonetica Sperimentale - Associazione Italiana di Acustica (GFS-AIA), Italy
GRECO PRC "Communication Homme-Machine" (CNRS), France
IEEE French Section, France
Institute of Acoustics (IOA) (Speech Group), UK
Ministère de la Recherche et de la Technologie (France)
Nederlands Akoestisch Genootschap (NAG) (Acoustical Society of the Netherlands)
Société Française d'Acoustique (SFA), France
Société des Electroniciens et Electro Techniciens (SEE), France

THE PREDICTION OF FOCUS

Anton Batliner*

*Institut für Deutsche Philologie, Ludwig-Maximilians-Universität, München, F.R.G.

Elmar Nöth**

**Lehrstuhl für Informatik 5 (Musterekennung), Friedrich-Alexander-Universität,Erlangen,F.R.G.

Abstract: *We present results on how focus is marked intonationally in German. Several speakers produced a large corpus of sentences. The corpus was constructed in a way that sentence modality and place of focus could only be differentiated by intonational means. Acoustic features representing the intonational parameters pitch, duration, and intensity, were extracted manually or automatically. The relevance of these features and the effect of several transformations were tested with statistical methods. Perceptual experiments where the listeners had to judge the naturalness and categories of the utterances were performed as well. By calculating average values for the (appropriately transformed) relevant features we found "normal", prototypical cases. We will show that by looking at utterances where all listeners agreed on the naturalness and (intended) categories we arrived at coinciding results. At the same time we found "unusual" but regular productions.*

MATERIAL AND PROCEDURES

This paper is concerned with the prediction of **focus**; focus is the part of an utterance which is *semantically most important*. On the phonetic surface focus is marked by the **focal accent** (**FA**). To be more exact, we will try to predict the **phrase** that carries the FA.

Our material consists of 360 German utterances, spoken by 6 untrained speakers (3 male, 3 female). Three different sentences with a similar syntactic structure were each put in different contexts that determined *sentence modality* as well as *place and manner of focus* (simple focus, focus projection, or multiple focus). For a detailed description of the corpus and the intended focal structures see the relevant contributions in /3/. In each of the sentences the last two phrases could be stressed, depending on the surrounding context. Based on the *sentence modality system* according to Altmann /1/, the sentences formed minimal pairs that could only be differentiated by their intonational form: *focus in final* vs. *focus in prefinal position* on the one hand, and *questions* vs. *non-questions* on the other hand. Table 1 shows an example of a context sentence, the pertinent test sentence, and the induced sentence modality and place of focus. Table 2 shows the three test sentences, a word-by-word translation into English, an appropriate translation, and a finer description of the induced sentence modalities *question/non-question* (**Q/NQ**).

The only instruction given to the speakers was to produce the context and the test sentence. We did not instruct the speakers to produce the FA in a certain way. By instructing the speakers, one can eliminate certain variabilities and facilitate the analysis. On the other hand one loses the chance to find regular and interesting deviations and merely receives several realizations of *representative cases* where representativeness is based on the intuition of the researcher. By evaluating a relatively large number of cases we expected to find both *representative cases* (which we will call **central types**) and *rarer but acceptable cases* (which we will call **marginal types**). We evaluated our data in two ways that proved to be converging:

- **Strategy 1**: We extracted acoustic feature values that represent the prosodic parameters pitch, duration, and intensity. Using a statistical classifier we tested the relevance of the features with respect to the place of the FA. By calculating average values for the relevant features we found the *central type* of each Q/NQ-FA constellation.

Table 1: Example of context and test sentence, induced sentence modality, and place of focus

Constellation of sentence modality and focus: Assertion, focus on "linen"	
Context:	Mother: *"What does the master make Nina weave at the moment?"*
Sentence:	Employee: *"She makes Nina weave the linen."*

Table 2: Test sentences, translation, and induced sentence modalities

Sie läßt die Nina das Leinen weben ?/.
She makes the Nina the linen weave
She makes Nina weave the linen
assertive question vs. assertion
Lassen Sie den Manni die Bohnen schneiden ?/!
Make the Manni the beans cut
Make Manni cut the beans
polar question vs. imperative
Lassen wir den Leo die Blumen düngen ?/!
let us make the Leo the flowers fertilize
let us make Leo fertilize the flowers
polar question vs. adhortative

- **Strategy 2**: We presented the utterances to a forum of listeners who judged the naturalness, category, and place of FA. Category roughly means sentence modality. As for the differences cf. /3/. By selecting the utterances that were judged to be the "best" ones and by comparing the feature values of those utterances with the average values from strategy 1 we found the *central type* as well as *marginal types*.

EXTRACTION OF FEATURES

For each utterance we calculated the following features:

For the whole utterance

- The *fundamental frequency* (F_0) at the end of the utterance (**offset**).
- The *all point regression line* of the F_0 values (**reg**).
- The *duration* in centiseconds (**dur**).

For the 2nd and 3rd phrase

- The *maximal* and *minimal* F_0 *value* (**max2, min2, max3, min3**).
- The *difference of the position on the time axis* of the extreme values in centiseconds (**pos2, pos3**).
- The *duration* in centiseconds (**dur2, dur3**).
- The *average* and *maximal logarithmic energy* (**aint2, mint2, aint3, mint3**).

The parameter values were extracted "by hand" on mingograms and automatically from the digitized versions of the utterances. (See /7/ for details on the F_0-algorithm and the computation of the energy values.) In /5/ we showed that automatically extracted F_0 values produced recognition rates comparable to those from mingogram values. An automatic extraction of the durational values however would pose a problem (see below).

PERCEPTION EXPERIMENTS

An average of 12 listeners participated in 3 different perception experiments:

- Context and test sentence were presented by ear phone and at the same time in a written version. On a rating scale from 1 (≙ test sentence matches very well with context) to 5 (≙ does not match at all) the listeners had to judge the *naturalness of the production*. We will name the average rating of the listeners **NAT**.
- The test sentence was presented in isolation. The listeners had to classify the sentence as *question, assertion, imperative, exclamation,* or *optative*). We will name the percentage of classifications as question **MOD**.
- The test sentence was again presented in isolation. The listeners had to decide which of the phrases carried the FA. If fa_i is the number of listeners who perceived the ith phrase as most stressed then

$$\mathbf{FOK} = (fa_2\text{-}fa_3)/(fa_1+fa_2+fa_3) \qquad (1)$$

takes on values between 1 (all listeners perceived the 2nd phrase as stressed) and -1 (all listeners perceived the 3rd phrase as stressed).

STATISTICAL EVALUATION

Each of the intonational features was used as a predictor variable in the **discriminant analysis** to predict Q/NQ and FA. Because of the combinatorial explosion the optimal feature combination had to be determined heuristically: The predictors entered the analysis separately and (if the feature was calculated for the 2nd and 3rd phrase) together with the corresponding variable for the other phrase. Several transformations for each variable were tested. In order to reduce the necessary amount of computation all cases were used both for learning and testing with learn=test. The relevant variables under the best transformation were put into multivariate discriminant analyses. We can only present the most important results; for a more detailed discussion see /4/.

- $\mathbf{F_0}$: The transformation of the *Hz* values into *semitones* (**st**) did not improve the classification results. A possible explanation could be that the semitone transformation "over" normalizes the different **voice ranges** of male and female speakers /5/. A normalization of the **voice register** by subtracting a reference value for either the speaker or the utterance resulted in significant improvements in the prediction. In the final analyses we used semitone values and subtracted the basic value of the speaker ($\mathbf{st_{bas}}$), i.e. the lowest F_0 value produced by the speaker.
- **Duration**: Best prediction was achieved after a normalization of the speaking rate that took into consideration average duration of that phrase for each speaker ($\mathbf{avdur_i}$) and the average duration of the syllables in the utterance (**dur / number of syllables**):

$$\frac{dur_i}{avdur_i} * \frac{dur_i}{dur \;/\; number\ of\ syllables} \qquad (2)$$

- **Intensity**: The best results were achieved with the *maximal energy in the 0-5000 Hz band.* Average values, "sonorant" energy subbands, and normalizations with respect to the average energy level of the utterance or with respect to the different intrinsic energy values of the vowels produced worse results.

Table 3 shows the predictor variables that proved to be the most relevant and that were used in the final analyses. For each variable (**var**) its transformation (**trans**), the recognition rate for Q/NQ (**modal**), and the recognition rate for FA for all utterances (**foc**), all questions (**focq**), and all non-questions (**focn**) is displayed. In the columns *focq* and *focn* only questions or non-questions were used for learning and testing. If the variable was only relevant in the prediction of either Q/NQ or FA, the other category is marked with a dash. Column **un** shows the recognition rate when the variable entered the analysis separately and column **bi** when the 2 corresponding variables were used.

Table 3: Relevant predictors, best transformations, and recognition rates for Q/NQ and FA

var	trans	modal un bi	foc un bi	focq un bi	focn un bi
offset	st_{bas}	93	-	-	-
reg	st	85	-	-	-
max2	st_{bas}	$73_{>94}$ $94_{>94}$	$60_{>78}$ $60_{>78}$	$80_{>81}$ $63_{>81}$	$54_{>92}$ $88_{>92}$
max3	st_{bas}				
min2	st_{bas}	$65_{>84}$ $84_{>84}$	$59_{>62}$ $47_{>62}$	$53_{>71}$ $70_{>71}$	$62_{>70}$ $47_{>70}$
min3	st_{bas}				
pos2	-	$76_{>85}$ $78_{>85}$	$51_{>69}$ $51_{>69}$	$78_{>82}$ $62_{>82}$	$69_{>52}$ $48_{>52}$
pos3	-				
dur2	(2)	-	$66_{>74}$ $72_{>74}$	$60_{>71}$ $70_{>71}$	$70_{>82}$ $82_{>82}$
dur3	(2)	-			
mint2	-	-	$58_{>66}$ $55_{>66}$	$52_{>56}$ $51_{>56}$	$62_{>70}$ $53_{>70}$
mint3	-	-			
multiv-l=t		97	93	95	96
multiv-l5t1		92	84	86	94
multiv-l1t5		92	78	76	82

For comparison, row **multiv-l=t** shows the recognition rates when all of the relevant variables were used in the learn=test constellation. Analyses with two further learn and test constellations were conducted:

- Training sample: *one speaker*, test sample: *5 speakers* (generalization from a single speaker to the other speakers, **multiv-l1t5**).
- Training sample: *5 speakers*, test sample: *one speaker* (simulation of speaker independence, **multiv-l5t1**).

By using the leave-one-out method, all cases could be used for learning and testing with learn≠test.

The results indicate that in questions other intonational parameters are used to mark the FA or the same parameters are used in a different way than in non-questions. The prediction is worse if questions and non-questions are analyzed together, than if they are treated separately.

The results under *focq* and *focn* were achieved with a grouping into questions and non-questions "by hand". For learn=test the grouping of the Q/NQ-classifier was used as an input to the *focq-* and *focn-classifier* as well. The classification errors of the first step even improved the results (as for a detailed error analysis cf. /4/).

CENTRAL AND MARGINAL TYPES

We will now show the two converging strategies (cf. above) how to find the central types:

1) Each of the 4 Q/NQ-FA constellations has one central type that is characterized by the average values of the predictors.
2) We inspected those cases where a strong agreement among the listeners could be observed: practically all the listeners agreed upon the intended Q/NQ grouping, the place of the FA, and the naturalness of the production (MOD≥80 for questions and MOD ≤20 for non-questions, |FOK| = 1, NAT ≤ 2). 24 out of the 360 cases passed these strict criteria. 19 cases could be identified as representatives of the central types.

For the 4 central types, Fig.1-Fig.4 show the average feature values as well as the F_0 contour of a typical production (4 out of the 19 cases): The dashed vertical line marks the border between the 2nd and the 3rd phrase of the actual production. For the 2nd and the 3rd phrase, each of the filled squares shows averages for the extreme values. The x-coordinate corresponds to the average position on the time axis in centiseconds starting from the beginning of the utterance, the y-coordinate corresponds to the average F_0 values (st_{bas}). On the top of each figure average

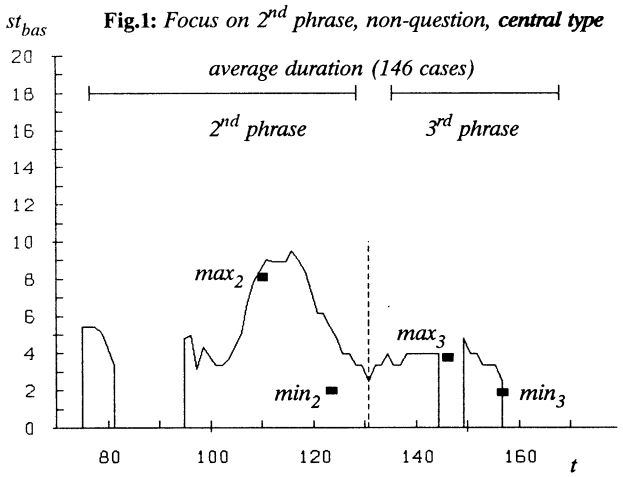

Fig.1: *Focus on 2^{nd} phrase, non-question,* ***central type***

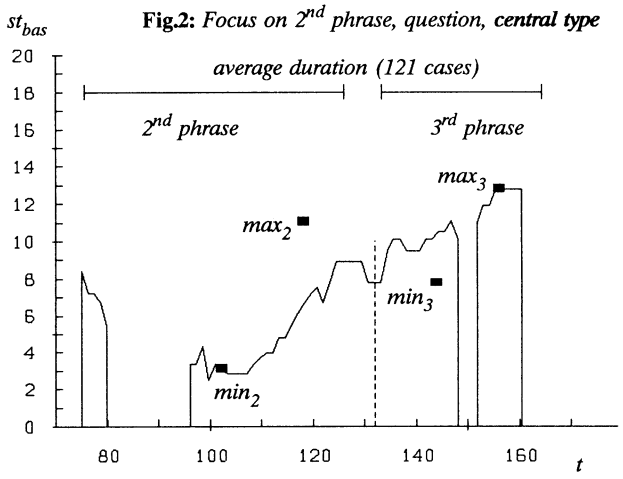

Fig.2: *Focus on 2^{nd} phrase, question,* ***central type***

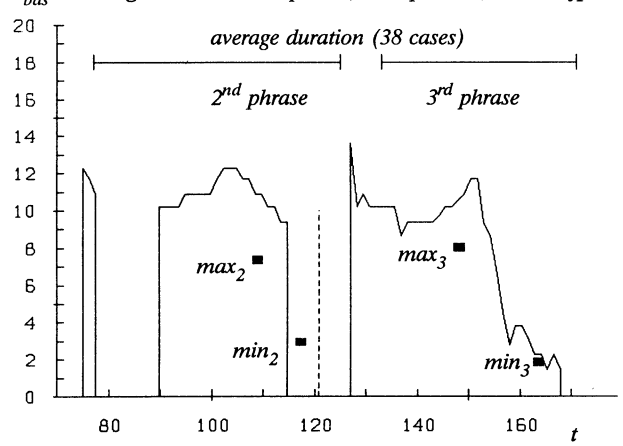

Fig.3: *Focus on 3^{rd} phrase, non-question,* ***central type***

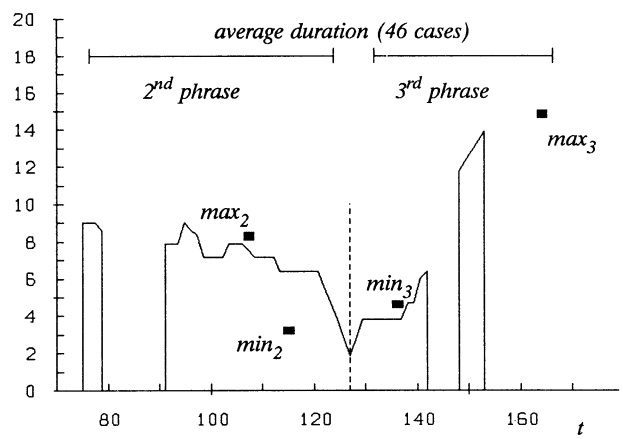

Fig.4: *Focus on 3^{rd} phrase, question,* ***central type***

beginning point and duration of the 2nd and 3rd phrases is displayed. In the following characterization of the central types, the terms *High, Low,* and *boundary tone* /8/ are used interchangeably with the terms *rising / falling contour.*

1) Focus on 2nd phrase, non-question (Fig.1): The contour is falling in both phrases (High Low). Max_2 is markedly higher than max_3; min_2 and min_3 do not differ.

2) Focus on 3rd phrase, non-question (Fig.3): The contour is again falling in both phrases (High Low). Max_3 is about as high as max_2, min_2 and min_3 do not differ.

Comparing the two types, we can say that the absolute values for the features of the 2nd phrase in Fig.1 and Fig.3 do not differ remarkably. It is rather the relative values of the features in comparison with the respective values of the 3rd phrase that mark the FA.

3) Focus on 2nd phrase, question (Fig.2): The contour is rising in both phrases (Low High).

4) Focus on 3rd phrase, question (Fig.4): In the 2nd phrase, this type has a falling contour comparable to the non-questions, whereas in the 3rd phrase, the contour is rising (Low High).

Comparing these two types, we can say that the F_0 range of the phrase with the FA is markedly greater than that of the other phrase. In the final phrase, a rising contour (high boundary tone) is used for both types to mark sentence modality.

The remaining five cases can be grouped into three marginal types which are displayed in Fig.5-Fig.7. To demonstrate the deviations from the central types, the respective average values are projected into the contours of the marginal types.

1) One speaker typically marked FA in prefinal position with a falling contour (High Low), even in questions. If one looks at the average feature values for all speakers and for this specific speaker, one could say that this marginal type across speakers is a central type for this speaker (Fig.5).

2) Another speaker typically marked questions only in the phrase with the FA, i.e. with FA in prefinal position, the final phrase showed a falling contour comparable to non-questions (Fig.6).

3) The last marginal type, a non-question with FA on 3rd phrase, could approximately be described as a "hat-contour" /6/, i.e. a concatenation of the two F_0-peaks on the 2nd and 3rd phrase and a low F_0-value at the end of the utterance (Fig.7).

RELEVANCE FOR AUTOMATIC SPEECH UNDERSTANDING

The material of this investigation is part of a larger corpus that was designed for a basic research project /3/. Hence the results are not directly transferable to an application field in an automatic speech understanding (**ASU**) project:

1) The phrase boundaries were extracted by hand. These boundaries are not only crucial for the extraction of durational features. They must also be known for the extraction of the extreme and average values of the F_0- and energy contours within each phrase.

2) The material was very consistent. In an unrestricted material, other factors such as the number of syllables in the phrase, the exact position of the syllable carrying the FA within the phrase, etc. would vary to a much greater extent.

On the other hand it is often necessary to prosodically verify competing sentence hypotheses. A sentence hypothesis is a syntactic and semantic interpretation of the speech signal. The interpretation normally includes information about word boundaries and thus phrase boundaries. Generally several hypotheses are generated of which one has to be selected by the

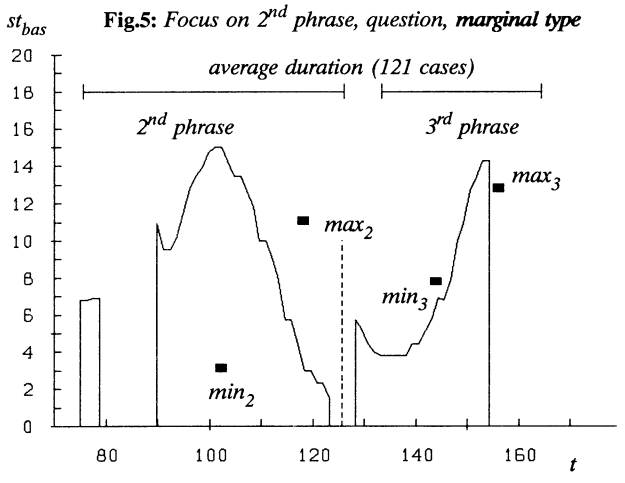

Fig.5: *Focus on 2nd phrase, question,* ***marginal type***

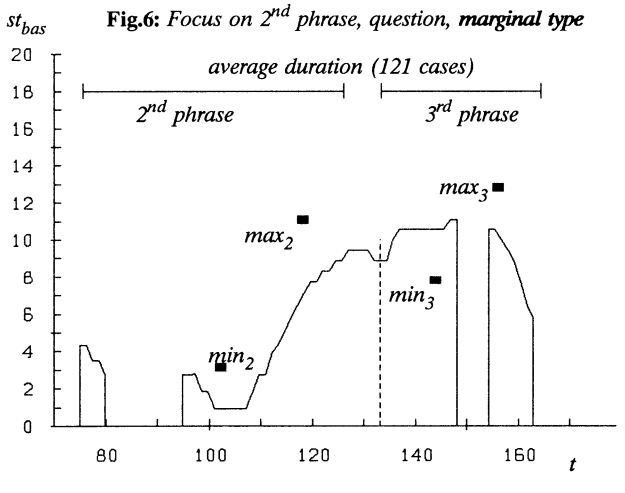

Fig.6: *Focus on 2nd phrase, question,* ***marginal type***

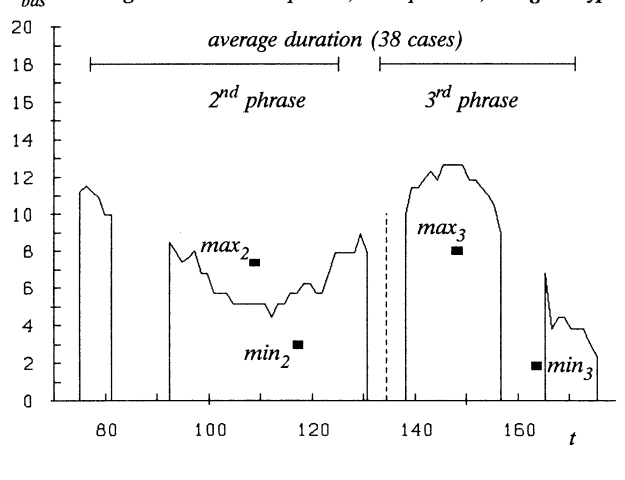

Fig.7: *Focus on 3rd phrase, non-question,* ***marginal type***

control module. Consequently the control module may request a verification of the hypotheses with respect to the prosodic parameters. In this case the phrase boundaries for each hypothesis are known.

For example, imagine an automatic system to handle dialogs about train schedule information. During a dialog the system has to decide between two competing hypotheses:

1) *Da* ***fährt*** *noch einer?* vs. 2) *Der fährt um* ***ein*** *Uhr?*
There ***leaves*** *one more?* *It leaves at* ***one*** *o'clock?*
Is there another one ***leaving?*** *Does it leave at* ***one*** *o'clock?*

Finding the right position of the FA can be crucial for the selection of the right hypothesis, as the phonetic structure of the two utterances is very similar.

Furthermore it should be noted that we concentrated on *how exactly* the prosodic parameters are used to mark the FA. We did not intend to present the most robust and easily extractable features that *represent* these parameters. In other words, we consider it a first step to show that for example the speaking rate has a significant influence on the actual value of the prosodic parameters *duration* and should consequently be taken into account by the prosodic module of an ASU system. It is a second step to decide which is the most appropriate *and* automatically extractable transformation for the normalization of the speaking rate.

Apart from the practical problem of transferring basic research into "real life" applications, we regard the results of the two different but converging approaches (strategy 1 & 2 above) to be relevant for ASU. Whereas in the case of speech synthesis only one typical realization (central type) for each different category is necessary, in ASU each acceptable type of realization (central **and** marginal types) must be accounted for, especially in a speaker independent application. Knowledge about the frequency of usage will inevitably lead to better performance in an ASU system.

CONCLUSION

Purpose of this study was to find out how focus is marked intonationally in German. We have shown that all 3 intonational parameters are used for this task (in order of importance: F_0, duration, and intensity). Speaker- or utterance-specific transformations of the features improved their relevance. Using two different approaches, a statistical and a "psychological" one (average values and perception experiments), we arrived at central (≙ mostly used) and marginal (≙ rare but acceptable) types. The results indicate that FA is marked differently in questions and non-questions. Speaker-specific ways to mark the FA were observed. Generally, the focus could be predicted with a high probability (up to 96 %), depending on the chosen constellation and/or transformation.

REFERENCES

/1/ H. Altmann, "Zur Problematik der Konstitution von Satzmodi als Formtypen", in J. Meibauer (ed.), "Satzmodus zwischen Grammatik und Pragmatik", Niemeyer, Tübingen, pp. 22-56, 1987

/2/ H. Altmann (ed.), "Intonationsforschungen", Niemeyer, Tübingen, 1988

/3/ H. Altmann / A. Batliner / W. Oppenrieder (eds.), "Zur Intonation von Modus und Fokus im Deutschen", Niemeyer, Tübingen, 1989 (to appear)

/4/ A. Batliner, "Fokus, Modus und die große Zahl. Zur intonatorischen Indizierung des Fokus im Deutschen", in /3/, 1989

/5/ A. Batliner / E. Nöth / R. Lang / G. Stallwitz, "Zur Klassifikation von Fragen und Nicht-Fragen anhand intonatorischer Merkmale", Proc. DAGA 1989, (to appear)

/6/ A. Cohen / J. 't Hart, "On the Anatomy of Intonation", Lingua 19, pp. 177-192, 1967

/7/ E. Nöth, "Prosodische Information in der automatischen Spracherkennung - Berechnung und Anwendung", Dissertation, Universität Erlangen, 1989 (to appear)

/8/ J. Pierrehumbert, "The Phonology and Phonetics of English Intonation", dissertation, M.I.T., 1980

The research in München was financed by the *Deutsche Forschungs-Gemeinschaft (DFG)*, the research in Erlangen by the *Bundesministerium für Forschung und Technologie (BMFT)*.