gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil
Verbundvorhaben

Why Sentence Modality in Spontaneous Speech is More Difficult to Classify and why this Fact is not too bad for Prosody

A. Batliner, C. Weiand,
A. Kießling, E. Nöth

L.M.-Universität München

F.–A.–Universität Erlangen–Nürnberg

Vm

Report 46
Dezember 1994

Dezember 1994

A. Batliner, C. Weiand,
A. Kießling, 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: `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 H/0 und 01 IV 102 C 6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

From Proc. ESCA Workshop on prosody, Lund, 1993, pp. 112–115.

Why Sentence Modality in Spontaneous Speech is More Difficult to Classify and why this Fact is not too bad for Prosody

A. Batliner[1], C. Weiand[1], A. Kießling[2], E. Nöth[2]
[1] L.M.-Universität München, Institut für Deutsche Philologie, Schellingstr. 3, 80799 München, F.R. of Germany
[2] Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), Martensstr. 3, 91058 Erlangen, F.R. of Germany

"You crazy," said Max. It was either a statement or a question.
(John le Carré: Tinker Tailor Soldier Spy)

"So you're our man, then," he said. It was half statement, half question.
(Josef Skvorecky: The Engineer of Human Souls)

ABSTRACT
We show in this paper that the labeling of sentence modality in German, esp. of questions vs. non-questions, is more difficult for spontaneous than for read speech and easier for non-elliptic than for elliptic utterances. However, the prosodic marking *of sentence modality is more important in elliptic utterances that occur more often in spontaneous speech.*

INTRODUCTION
Until now, most research has been done on controlled, read speech (i.e., non-spontaneous speech, henceforth NSP), and so far, little work has been reported on spontaneous speech (SP) in German. In an experimental design for the recording of NSP, sentence modality, e.g. question/non-question (Q and NQ respectively), can be controlled beforehand via the careful construction of the linguistic context, explicit instructions or simply via punctuation marks. In SP, however, sentence modality has to be determined afterwards, using different criteria - syntactic, semantic, contextual, or prosodic; the corresponding cues are not always present, especially because SP often contains elliptic utterances. In this paper, we will concentrate on the marking of the Q/NQ dichotomy in SP and NSP as well as in elliptic and non-elliptic utterances (ELs and NELs respectively). Related work and comparable results for English are reported e.g. in [4].

MATERIAL AND EXPERIMENTAL DESIGN
Two pairs of speakers (3 female, 1 male) who didn't know that they were recorded for prosodic research had to solve different problems in a "blocks world". The experiment was designed in a way that resulted in absolutely SP (short clarification dialogs with many turn takings). The utterances were transliterated and classified along the lines of a formal syntactic model, cf. [1]. The four cross-classified main groups were Qs vs. NQs and ELs vs. NELs. From the whole material those utterances were chosen for further investigation that met the following criteria: a sufficient signal quality and no specific non-syntactic phenomena like hesitations which are normally only found in SP. We chose all Qs, all ELs, and out of the NQs all non-statements that met the criteria, and roughly the same number of NEL statements. After 9 months, the same 4 speakers read the chosen utterances - their own utterances and those of the partner, given in written form and embedded in a sufficiently large context. Recording conditions were comparable to a quiet office environment. The 1329 utterances (approx. 30 minutes of speech, 1/3 SP, 2/3 NSP) were digitized with 12 Bit and 10 kHz. The number of the four main sentence types is the following (in parenthesis, NELs/ELs): Qs: 566 (332/234), **statements**: 623 (266/357), **commands**: 128 (108/20), **exclamations**: 12 (9/3); i.e. NQs in total: 763 (383/380). Using three different F0 algorithms, a F0 contour was computed and corrected manually to obtain a reference contour. From the corrected F0 contour the following features were

extracted: Onset, offset, maximum, minimum, range, mean, standard deviation, and regression coefficient. These features were normalized with respect to the average F0 value of the utterance. A perception experiment was performed where 10 naive listeners had to classify each utterance as Q or NQ. For more details, cf. [2] and [3].

CLASSIFICATION OF NQs VS. Qs

The classification problem was already mentioned in the introduction. We assume that for ELs, the prosodic marking is more important than for NELs, because other features such as e.g. word order are missing. This assumption is reasonable but as far as we can see it has up to now not been verified for German. It would, however, almost be a sort of "self-fulfilling prophecy" if the object of investigation (prosodic marking) is used as crucial criterion for the classification. There is no simple way out of this "classification paradox". We decided therefore to use three different classification procedures:

1. **Linguistic classification**, where the sentences were classified according to a formal syntactic model by an expert who listened to the utterances as well (formal classification without contextual knowledge).
2. **Perceptual classification**, where a group of naive listeners had to determine the sentence modality of the utterances presented in isolation ("out of the blue"-sentences).
3. **Context classification**, where the sentences were classified by another expert with the help of contextual features (content criteria and dialog structure, e.g., what does the speaker know, what is the reaction of the listener, etc.) and with the help of syntactic features, but without listening to the utterances, i.e. without prosodic knowledge (functional classification).

The context classification was conducted for the SP part of the material; their NSP counterparts could be grouped automatically into the same class because they were embedded into the same context. We established four classes, NQs and three Q classes:

1. **NQs:** All utterances that are not followed by an answer, a confirmation, etc.; it is obvious that the speaker is in possession of the information at stake but not the partner.
2. **possible Qs (Q*poss*):** Utterances followed by an answer; the context shows that both speaker and partner are in possession of the information at stake. The context and/or lexical information (e.g. modal particles) give no clues whether the speaker is confident about that what he/she says or not. Quite often the speaker is simply paraphrasing something the partner has said just shortly before.
3. **probable Qs (Q*prob*):** Utterances followed by an answer, but not clear-cut Qs; the context shows that, in contrast to Q*poss*, the speaker obviously does not know whether he is right or wrong, but the partner does. Often, the speaker uses a modifying particle, e.g. *vielleicht* (*perhaps*).
4. **Qs:** clear-cut questions, i.e. utterances followed by an answer, etc., mostly with an agreement of contextual and grammatical criteria (e.g. WH-questions). It is obvious from the context that the information needed by the speaker is in possession of the partner but not of the speaker.

The following example can illustrate both Q*poss* and Q*prob*: speaker: *"The green block is on the red one."* – partner: *"Yes, that is right."*. Depending on the different contextual information, cf. above, the first sentence is assigned either to Q*poss* or to Q*prob*. With only syntactic information, the first sentence had to be classified as a clear-cut statement. The reaction of the partner makes it possible that the first utterance could be a declarative Q. Without prosodic and/or contextual information, the conflict cannot be solved, because almost any statement can be followed by a confirmation or by a negation.

RESULTS AND DISCUSSION

As for the context classification, a systematic difference between ELs and NELs can be seen in figure 1 for SP. Note that the classification for the NSP counterparts is identical, cf. above: in the clear-cut categories NQs and esp. in Qs, there are more NELs than ELs. It is the other way round in the two other categories (approx. 25% of the cases); i.e. ELs are really less clear-cut than NELs.

In figure 2 and 3, the height of the F0 offset in semitones (st) subtracted by the F0 mean of the utterance as the most stable prosodic feature indicating the Q/NQ-dichotomy, is plotted for the four context categories. For NSP (figure 3), there is almost a linear relationship between offset and Q-proneness: the more Q-prone, the higher the offset. There is, however, no difference in SP between ELs and NELs for NQs; for Qs in SP (figure 2), the offset is markedly higher in ELs than in NELs.

In figure 4, the perception results are compared with the context classification; as almost no difference could be noticed between SP and NSP, they are plotted together. The ordinate shows the frequency of the cases, the abscissa perceived NQs and Qs for the four context classes. A perceived NQ is defined if less than five out of the ten listeners classified an utterance as Q; the other cases are classified as Q. In approx. 5% of the cases, cf. the small bars for NQ and Q, there is disagreement between context and perceptual classification due to an inherent difficulty in the context classification and/or an equivocal prosodic marking of the utterances; for details, cf. [3].

Figure 5 and 6 put the F0 offset in relation to the perception experiment. The abscissa shows the number of listeners that categorized an utterance as Q, the ordinate shows – analogously to figure 2 and 3 – the average of the height of the F0 offset in semitones (st) in relation to the F0 mean of the utterance. There were not many scores in the region between 2 and 8 and extreme values would have a distorting influence on the mean of the offset. This region is therefore combined and projected onto the value 5. For ELs, there is a linear relationship between F0 offset and Q-score: the higher the offset, the more listeners classified the utterances as Qs. The linearity is more pronounced for NSPs (figure 6) than for SPs (figure 5), and for SPs, the offset is markedly higher in the rightmost region, i.e. for Qs. For NELs, this relationship is much less clear. Obviously, Q-proneness is marked

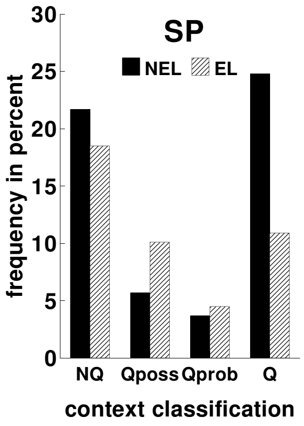

Figure 1

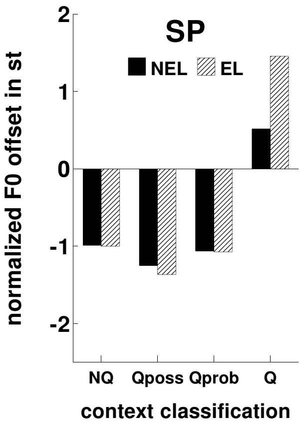

Figure 2

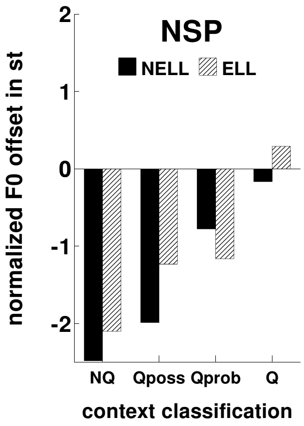

Figure 3

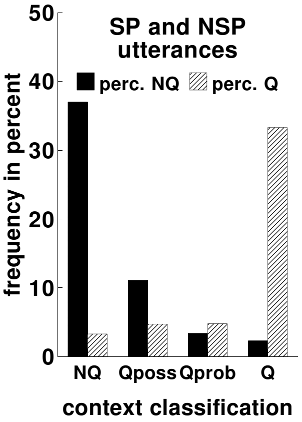

Figure 4

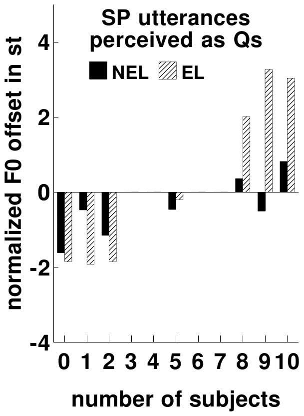

Figure 5

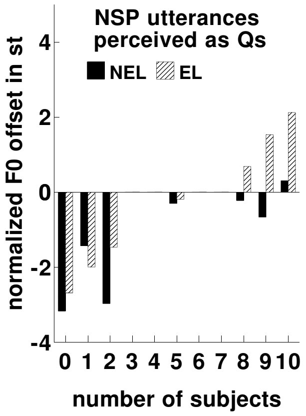

Figure 6

much more with prosodic means in ELs than in NELs.

FINAL REMARKS

Coming back to the first part of the title of this paper, it is now clear why sentence modality in SP is more difficult to classify than in NSP: even if the Q/NQ-dichotomy holds for most of the utterances, one should say goodbye to a straightforward and clearcut dichotomy. In quite a number of cases (approx. 20%, cf. Q*poss* and Q*prob* in figure 1 and figure 4), contextual and prosodic features point towards a category in between Qs and NQs that is illustrated in the two quotations above: sometimes, the category can not be decided upon (le Carré, Q*poss*), sometimes, it is really just something in between (Skvorecky, Q*prob*). That holds especially for ELs. Note that ELs do occur much more often in SP than in NSP; in our material, however, both are strictly parallelized. In real life, this difference will thus show up even more clearly. There was no pronounced difference between NSP and SP, although NSP behaved more regularly. There is, however, throughout a difference between ELs and NELs: sentence modality in ELs is more often marked by prosodic means. This fact corroborates the second part of our title: as ELs do occur quite often in SP, prosody will be needed much more in automatic speech recognition – if one really wants to deal with SP.

Acknowledgements

This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research project ASL/VERBMOBIL and by the *Deutsche Forschungsgemeinschaft* (*DFG*). Only the authors are responsible for the contents of this paper.

References

[1] H. Altmann, A. Batliner, and W. Oppenrieder, editors. *Zur Intonation von Modus und Fokus im Deutschen.* Max Niemeyer Verlag, Tübingen, 1989.

[2] A. Batliner, B. Johne, A. Kießling, and E. Nöth. *Zur prosodischen Kennzeichnung von spontaner und gelesener Sprache.* In G. Görz, editor, *KONVENS 92*, Informatik aktuell, pages 29–38. Springer-Verlag, Berlin, 1992.

[3] A. Batliner, A. Kießling, and E. Nöth. *Die prosodische Markierung des Satzmodus in der Spontansprache – Methodologie und erste Ergebnisse.* Technical Report: ASL-Süd–TR–14–93/LMU, Februar 1993.

[4] N. Daly and V. Zue. *Statistical and Linguistic Analyses of F0 in Read and Spontaneous Speech.* In *Int. Conf. on Spoken Language Processing*, volume 1, pages 763–766, Banff, Canada, 1992.