Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

Consistency in Transcription and Labelling of German Intonation with GToBI

Martine Grice
Matthias Reyelt
Ralf Benzmüller
Jörg Mayer
Anton Batliner

Universität Saarbrücken
Universität Braunschweig
Universität Stuttgart
Universität München

Vm

Report 153
October 1, 1996

October 1, 1996

Martine Grice
Matthias Reyelt
Ralf Benzmüller
Jörg Mayer
Anton Batliner

Institut für Phonetik
Universität des Saarlandes
66041 Saarbrücken

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstr. 22
38092 Braunschweig

Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Azenbergstr. 12
70174 Stuttgart

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

Tel.: (0531) 391 - 2479
Fax: (0531) 391 - 8218
e-mail: reyelt@ifn.ing.tu-bs.de

Gehört zum Antragsabschnitt: 14.6 Prosodische Etikettierung

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 N0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Contents

1 Introduction

The proliferation of German language databases over the last few years has led to the development of a number of machine-readable signal-aligned systems for the transcription and labelling of German intonation and prosody. These are summarised by their developers in a survey [1] carried out for the One Day Workshop on Prosodic Labelling in Stockholm, August 1995. Amongst them are the pitch contour-based system of Kohler [2], the multi-level parameter approach of Heuft and Portele [3], and the pitch level-based approaches of Reyelt and Batliner [4], Grice and Benzmueller [5], and Mayer [6]. The latter three systems are related to the ToBI (Tones and Break Indices) system developed within the English speaking community [7, 8].

The pitch level-based approaches mentioned above are currently being used to annotate a wide range of databases in the German language. Since they are largely compatible, it was decided to create a consensus core set of symbols which could be used in order to facilitate sharing of transcribed corpora. We refer to this system as German ToBI, or GToBI. Partial mappings between GToBI and the contour-based approach have been considered in [9] and [10].

2 GToBI

2.1 Preliminaries

As in the English ToBI (henceforth EToBI), the GToBI system makes use of two tones, H and L. These may have a prominence-lending function, being grouped together into pitch accents. Alternatively, they may have a delimitative function, acting as final edge tones of intermediate phrases and intonation phrases.

All three German systems upon which this consensus is based have these two levels of phrasing, although one system ([6], based on [11]) does not have a tone directly attached to the intermediate phrase edge. Their inventories of pitch accents are, however, more diverse. In selecting a consensus set, we retained distinctions even if they were not common to all systems, favouring overspecification; it is simpler at a later stage to automatically collapse two categories into one, than it is to introduce a distinction, which would require later relabelling. The consensus set is outlined below.

2.2 Tones and their phonetic realisation

In the model on which ToBI is based [12], all tones are phonetically manifested as points in frequency and time which are interpolated between.

The scaling of these tones when they are combined into accent or edge tone clusters is not always transparent. It is affected by the two operations, upstep, which, after a H- intermediate phrase edge tone, automatically raises the pitch of intonation phrase edge tones; and downstep, which lowers the pitch of accent H or intermediate phrase H tones. In ToBI, unlike in [12], downstep does not apply automatically; it requires a special diacritic '!' before the H tone concerned.

GToBI Pitch Accents There are six basic pitch accents:

- H* 'peak accent'
- L* 'low accent'
- L*+H 'valley accent plus rise'
- L+H* 'rise from low up to peak accent' (peak on or just after the accented syllable)
- H+L* 'step-down from high to low accent' (valley clearly at or near bottom of speaker's range).
- H+!H* 'step-down from high to mid accent' (scaling of !H* same as other !H tones)

Five of the basic accents contain H tones which can be downstepped. This increases the inventory from 6 to 11 accents.

GToBI Edge Tones There are two intermediate phrase (ip) edge tones: L-, an F0 minimum low in the range, and H-, which has roughly the same F0 value as the peak corresponding to the most recent H tone in the phrase. This is true in combination with intonation phrase (IP) edge tones too. Since an IP edge tone never occurs without a preceding ip edge tone, they are given in combination. The first two combinations are affected by upstep.

- H-L% plateau
- H-H% plateau followed by sharp rise at end of phrase
- L-H% low followed by rise to mid at end of phrase
- L-L% extra low at end of phrase

Comparing H- with H-L%, the main difference is not tonal, but rather relates to perceived boundary strength, as encoded in EToBI by labels in the parallel Break Index tier. The only appeal to Break Indices made in the consensus system is in the introduction of a label to mark discrepancies between perceived boundary strength and tonal cues (roughly equivalent to the ToBI Break Index 2 or '-'). In addition to this discrepancy label, transcribers were also allowed to signal uncertainty, inserting '?' after the label concerned.

It will be evident from the inventories above that the GToBI tonal categories are similar to those in EToBI. This is not because all languages can be described with such similar inventories of pitch accents and edge tones, but because English and German are closely related languages which share a similar rhythm and intonation structure. There are differences in their inventories of pitch accents and in the phonetic realisation of the pitch accent categories they share, especially in relation to the timing of F0 events. This has a bearing on the weighting of criteria used for the selection of individual tonal elements (see [13] for more details). A difference in the definition of intermediate phrase is that in GToBI ip's do not have to contain an accent. When accentless, they are subordinate to a preceding or following accented intermediate phrase within the same intonation phrase.

3 The Experiment

An experiment was carried out in which 13 transcribers across three sites independently labelled a common corpus of speech data.

3.1 Speech Data

The corpus used for labelling comprised 304 seconds of speech and 733 orthographic words. It contains representative samples of speech data from databases already being labelled for other purposes at the sites participating in the experiment. Samples were either paragraph-length stretches of read speech or sequences of task-oriented dialogue turns, where a task or subtask is introduced and brought to completion. No isolated utterances were selected. The experimental conditions replicate conditions for routine labelling where labellers have access to enough context for "tuning in" to each speaker's range and vocal characteristics.

In the task-oriented dialogues, participants were able to speak freely with no intervention by third parties. They were of two types: (1) dialogues involving free exchange of information via the auditory channel as to a route on a map, (2) dialogues involving the scheduling of meetings via audio and visual channels in a set-up where participants pressed a button whilst speaking. There were three

Dialogue	words	Read speech	words
Scheduling	318	News	116
Map task	101	Story	82
		Guide	116

Table 1: Speech data used in labelling experiment

Site	GToBI developer	some experience with related systems	no prior experience
SB	1	0	4
BS	1	2	1
ST	1	2	1

Table 2: Labellers taking part in experiment

types of read speech: (1) A paragraph from a German classic read by a trained actor, taken from a published CDRom (2) Two news items from a German national radio station, and (3) Two paragraphs from a tourist guide, read by an untrained speaker.

3.2 Labellers

Labellers were all native speakers of German, studying or working at the universities of Saarbruecken (SB), Braunschweig (BS) and Stuttgart (ST). Some were experienced in labelling with related systems [4, 5, 6]. The labellers' profiles are given in table 2.

3.3 Experimental procedure

Training All labellers were required to work through a training manual [13], which was compiled for the consensus set of pitch accents and edge tones. These materials describe the different categories of pitch accent and edge tones, giving for each item separately (a) a schematic representation, (b) a set of criteria for its selection, and (c) pointers to a number of files containing prototypical examples. Prototypical examples were chosen from either read or spontaneous speech, rather than specially produced stimuli, and consisted of those examples where developers were in agreement not only as to the label used, but also that the instance of the category being exemplified was representative.

Labelling Labelling was carried out using either *ESPS xwaves* speech analysis software or *fish*, a free package for the display and annotation of speech [14]. Labellers worked independently and were not allowed to discuss utterances in the experimental data-set.

4 Results

4.1 Inter-Transcriber Consistency

Following the procedure used for EToBI [15], inter-transcriber consistency was measured by comparing the labels placed by transcribers on each potential site for a tonal element (on and after each word for pitch accents and edge tones respectively). Transcribers' labels were compared in pairs; the comparison of a pair of labels on or after each word counts as a transcriber-pair-word. The measure of inter-transcriber consistency is "the percentage of transcriber-pair-words exhibiting agreement on a particular element potential site) in the transcription" [15 :125]. It is shown in [15] that this is a stringent metric: when three out of four transcribers agree on a label, only three of the six transcriber-pair-words generated match, thus producing an agreement of 50%.

In this experiment 733 words were transcribed by 13 labellers, totalling 9499 transcribed words. The number of transcriber-pair-words (excluding cases where a transcriber was compared against self) was 57174 for pitch accents and 57174 for edge tones.

Pitch Accent Labelling Agreement. The overall inter-transcriber consistency for pitch accents was 71%. Transcription involved the placing of one of the following ten pitch accent labels on each word: zero accent, H*, !H*, L*, L*+H, L*+!H, L+H*, L+!H*, H+L*, H+!H* (the theoretically possible !H+!H* and !H+L* not having been selected). Part of this agreement is on whether or not an accent was present on a word (i.e., whether zero or one of the other labels was transcribed). Looking at this separately, agreement as to the presence of an accent was 87%, and where two transcribers agreed that a word was accented, the agreement as to which accent was present was 51% (33% of the disagreement involving confusion between L+H* and H*, see section 4.4). Since basic accents and their downstepped counterparts are closely related, we also computed the inter-transcriber consistency across transcriber-pair-words where downstep was not taken into account. In this case agreement was 74%. Agreement as to whether tones were downstepped or not was 82%.

1	85.6%	6	85.9%*	11	84.4%
2	86.6%*	7	85.2%	12	82.2%
3	81.6%	8	87%	13	83%
4	86.8%	9	85.1%		
5	83.5%	10	85.7%*		

Table 3: Consistency of individual transcribers - all tonal elements

Edge tone labelling agreement. The overall inter-transcriber consistency for edge tones was 86%. In this type of labelling transcribers placed after each word either no label, one of two ip edge tones in isolation (L-, H-), or one of five combinations of ip and IP edge tones (L-L%, L-H%, H-H%, H-L%, !H-L%). The theoretically possible !H- in isolation and !H-H% were not used. The above score was obtained by taking as one category each edge tone label or label combination used by the transcribers. Calculated in this way, agreement at intonation phrase boundaries involves not just the agreement of the intonation phrase edge tone itself, but also that of the preceding intermediate phrase edge tone. If both the ip and IP tones in a transcriber-pair-word do not match, disagreement is registered. Transcribers agreed as to the strength of the boundary 86% of the time.

4.2 Differences across labellers

For the purposes of comparison, we replicated as far as possible the method for pooling across all tonal categories used in [15], calculating the consistency between each transcriber and the remaining 12. This meant assembling a agreement matrices for all accents and all edge tones (this time treating intonation phrase and intermediate phrase edge tones separately). Pooling across all tonal categories, pitch accents and edge tones, and treating each category as distinct (i.e. not merging any categories, such as downstepped accents with their non-dowstepped counterparts) the consistency scores for individual transcribers are in table 3. The mean score is 84.8%.

The three asterisked values are those of the three developers taking part in the experiment. Since the less experienced transcribers outnumber the developers, it could be expected that the developers would not have higher consistency scores than the other transcribers, but it might be expected that there would be greater variability among the less experienced set. If we isolate the three developers and compare each transcriber to the other two in turn, the mean score is 88.9%. Taking the group of less experienced transcribers as a separate group, their mean score

Distinction	GToBI	EToBI
accented/unaccented	87%	80.6%
pitch accent labelling	71.2%	68%
pooled tonal labelling	84.8%	81.4%
pooled tonal labelling - no downstep	85.7%	82.9%

Table 4: GToBI and EToBI agreement compared

is somewhat lower, at 84.0%.

4.3 Results compared with EToBI

The agreement across transcribers in GToBI was calculated in the same way as in [15] . Results are given alongside EToBI scores in table 4.

Although the sources of variability are not identical, our corpus containing 733 words (compared with 489), being labelled by 13 (compared with 26) transcribers, these percentages can give a general indication as to comparability between GToBI and the English ToBI system.

4.4 Confusions between pairs of accents

Since the lowest agreement was obtained on pitch accent labelling, we investigated which pairs of basic pitch accents were most often disagreed upon, or confused. For this analysis we collapsed basic pitch accents with their downstepped counterparts and took into account only those transcriber-pair-words where transcribers agreed that an accent was present. We calculated the percentage of the total disagreement which each pair accounted for. The highest percentage, 33% (N=2641) was accounted for by the H*/L+H* pair, followed by L+H*/L*+H (14%, N=1077) and H*/H+!H* (13%, N=1030). However, these scores are not weighted according to how often individual accents are transcribed, or indeed how often they are confused in general. To do this, we expressed the number of confusions between the two accents in the pair as a percentage, not of the total number of confusions, but of those involving either or both of the accents concerned. In this calculation, L+H*/H* accounted for 28%, L*+H/L* 17% (N=771), L+H*/L*+H 16% and H*/H+!H* 15%. All other pairs had confusions which were equal to or less than chance.

It is clear that L+H*/H* is the pair most often confused, however the numbers are relativised. The main difference between these two accents is that there is a sharp rise, often coupled with an expanded pitch range, on the former, and the option of a gradual rise on the latter. The timing of the peak within the accented syllable is not distinctive; it is usually late in L+H* but may also be late in H*. It appears to be the extent of the rise which labellers have difficulty categorising. L+H* is also frequently confused with L*+H. Here there is a sharp rise in both accents. The difference is one of timing, in that the L and H tones are realised later in L*+H, and there is a sharp rise in both cases . A similar timing distinction has been investigated by Kohler [16], and found not to be categorically perceived. Although it is not difficult to find prototypical cases of any of the three accents, they clearly overlap considerably.

Two other slightly less frequent confusions are L*/L*+H and H*/H+!H*. Both of these confusions could involve neutralisation. Before a H- edge tone, it is only possible to distinguish between L*+H and L* when the stretch between the accent and the phrase edge is sufficiently long. H+!H* is often neutralised with !H* (merged with H* in this calcuation) when the high pitch on the preaccentual syllable is closely following high pitch attributable to another tone.

5 Discussion and conclusion

These results indicate that GToBI is already adequate for the transcription of databases in German. Inter-transcriber consistency is comparable to that obtained in a similar study using the English ToBI system. In compiling the inventory of accents, we included a number of labels with a view to later merging. However, we have seen that there are no clear candidates for merging. The most obvious one, given its confusability, L+H*, was frequently confused with two different accents, which meant that it could not be simply treated as a subcategory of either of those two pitch accents.

There is an indication that improved training might reduce the number of disagreements, since the developers were more consistent among themselves than the other labellers. However, this difference was slight, indicating that it is possible for non-experts to gain operational skill with GToBI. This is a necessary prerequisite for a system which is to be used for multi-site large scale database annotation[1].

6 REFERENCES

1. Mayer J, Questionnaire for prosodic labelling systems/ approaches in Ger-

[1]The speech data used for this experiment and GToBI training materials can be obtained by ftp, for details contact `benzm@coli.uni-sb.de` or `mgrice@coli.uni-sb.de`

man, (http://www.ims.uni-stuttgart.de/phonetik.joerg/stockholm/questionnaire.html).

2. Kohler K, PROLAB - The Kiel System of Prosodic Labelling, Proc. ICPhS, Vol. 3:162-165, 1995.

3. Heuft B et al, Parametric Descrption of F0 Contours in a Prosodic Database, Proc. ICPhS, Vol 2:378-381, 1995.

4. Batliner A and Reyelt M, Ein Inventar prosodischer Etiketten für VERBMOBIL, Verbmobil Memo 33-94, 1994.

5. Grice M and Benzmüller R, Transcription of German using ToBI Tones - The Saarbrücken System, Phonus 1, 33-51, 1995.

6. Mayer J, Transcribing German Intonation - The Stuttgart System, ms, University of Stuttgart.

7. Beckman M and Ayers G, Guidelines for ToBI transcription, version 2.0, Ohio State University, 1994.

8. Beckman M and Hirschberg J, The ToBI Annotation conventions, Ohio State University, 1994.

9. Kohler K, TOBIG and PROLAB - Two prosodic transcription systems for German compared, Workshop on Prosodic Labelling, ICPhS Stockholm, 1995.

10. Grice M and Benzmüller R, First attempt at a comparison between PROLAB and the German ToBI systems, Workshop on Prosodic Labelling, ICPhS Stockholm,1995.

11. Fery C, German Intonational Patterns, Niemeyer, 1993.

12. Pierrehumbert J, The Phonology and Phonetics of English Intonation, MIT PhD Dissertation, 1980.

13. Benzmueller R and Grice M, Trainingsmaterialien für GToBI, Saarbrücken, February 1996.

14. Reyelt M, Ein flexibles Programmpaket zur Visualisierung von Sprachdaten, In Fellbaum, K (ed) Proc. Elektronische Sprachsignalverarbeitung, 358-365, Berlin, 1994.

15. Pitrelli J, Beckman M and Hirschberg J, Evaluation of prosodic transcription labeling reliability in the ToBI framework, Proc. ICSLP, Yokohama, 1994.

16. Kohler K, Terminal Intonation Patterns in Single Accent Utterances of German: Phonetics, Phonology and Semantics, in Kohler K (ed), Studies in German Intonation, AIPUK 25, 1991.