

DIMA – ANNOTATION GUIDELINES FOR GERMAN INTONATION

Frank Kügler¹, Bernadett Smoliboeki¹, Denis Arnold², Stefan Baumann³, Bettina Braun⁴, Martine Grice³, Stefanie Jannedy⁵, Jan Michalsky⁶, Oliver Niebuhr⁷, Jörg Peters⁶, Simon Ritter³, Christine T. Röhr³, Antje Schweitzer⁸, Katrin Schweitzer⁸, Petra Wagner⁹

¹University of Potsdam; ²University of Tübingen; ³University of Cologne; ⁴University of Konstanz; ⁵ZAS Berlin; ⁶Oldenburg University; ⁷University of Southern Denmark; ⁸Stuttgart University; ⁹Bielefeld University
kuegler@uni-potsdam.de

ABSTRACT

This paper presents newly developed guidelines for prosodic annotation of German as a consensus system agreed upon by German intonologists. The DIMA system is rooted in the framework of autosegmental-metrical phonology. One important goal of the consensus is to make exchanging data between groups easier since German intonation is currently annotated according to different models. To this end, we aim to provide guidelines that are easy to learn. The guidelines were evaluated running an inter-annotator reliability study on three different speech styles (read speech, monologue and dialogue). The overall high κ between 0.76 and 0.89 (depending on the speech style) shows that the DIMA conventions can be applied successfully.

Keywords: German, intonation, annotation, guidelines, inter-annotator reliability.

1. INTRODUCTION

We present a consensus system for a prosodic annotation of German, developed by intonologists of German over the past four years. *DIMA* stands for *Deutsche Intonation, Modellierung und Annotation* and is rooted in the framework of autosegmental-metrical (AM) phonology [2, 13, 19, 26]. Our goal is to gain a phonetically informed phonological annotation in a way that spans different variants of the AM framework. The general aim is to achieve compatible annotations of (corpus) data, thus facilitating the exchange of data. In order to increase exchangeability and compatibility in particular with existing data we envision automatic mappings from DIMA to the phonological systems used by different working groups such as GToBI [10], GToBI(S) [20], KIM [14] and off-ramp analyses like [8, 24].

The motivation for a consensus system for the annotation of German intonation lies in a diverse usage of these different phonological models, as illustrated in (1). The interpretation of the low pitch before the accentual H* tone is either attributed to a low leading tone [10] or to a rightward-spreading low initial boundary tone [24]; the falling pitch after

the accentual H* is either interpreted as a low boundary tone [10] or as a low trailing tone [20, 24]. The DIMA system is confined to the representation of those aspects of tonal structure which are accounted for in all the phonological models mentioned. In (1), for example, the DIMA annotation will represent the high tonal target as an accentual tone and the initial and final targets as boundary tones, whereas the low targets before and after the accentual peak will not be assigned to a specific tone class, such as a leading tone, a trailing tone, or a phrase accent. We hope that this under-specification, as first suggested in [11], will make it easier to exchange annotated data and corpora. In addition, the DIMA system should be easy to learn.

- (1)
- a. L+H* L-% [10]
Mein ZAHN tut weh. 'My tooth is hurting.'
- b. %L H*L L% [24]

2. PRELIMINARIES

The symbols used for annotation were borrowed from the classical ToBI system [1]. We propose three distinct layers of intonational events as well as one layer for comments as illustrated in Figure 1 using Praat [3]. The distinct layers indicate phrase boundaries, tones and corresponding diacritics, and prominences. As a crucial departure from other systems, these layers are annotated independently of each other. A prerequisite is labelled text at the levels of words and (stressed) syllables. Table 1 lists the inventory of symbols used for DIMA annotation.

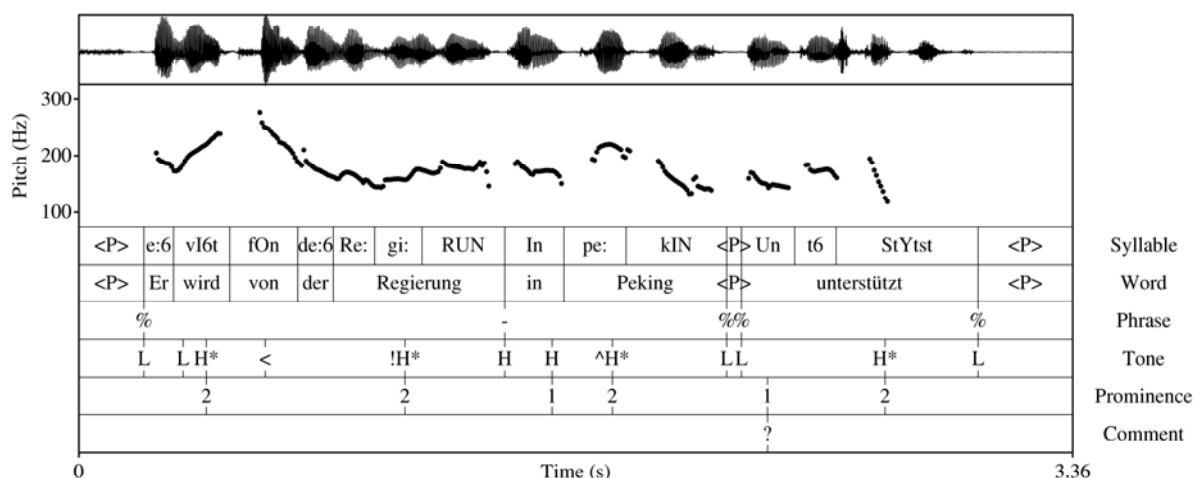
Table 1: Symbols for prosodic DIMA annotation.

Layer	Symbols
Phrase	% -
Tone	H* L* L H ! ^ < >
Prominence	1 2 3
Comments	e.g. ?

2.1 Phrase boundaries

Two types of phrase boundary are distinguished: A prosodic phrase with a strong boundary (%) and one

Figure 1: Illustration of DIMA-annotation layers and annotated intonational events for the utterance *Er wird von der Regierung in Peking unterstützt* ‘He is supported by the government in Beijing’. Segmental annotations in SAMPA.



with a weak boundary (-). Based on the prosodic hierarchy [21] we assume that a prosodic phrase with a weak boundary is dominated by a phrase with a strong boundary, hence two levels of phrasing. Auditory-phonetic criteria for the presence of a boundary are a pause, phrase-final lengthening and tonal movement, pitch reset, and other prosodic phenomena such as laryngealisation. The decision on the type of boundary depends on the number of co-occurring criteria and thus their perceptual strength. Figure 1 shows two prosodic phrases with a strong boundary in one utterance since perceptual impression suggests a bundle of the mentioned boundary criteria. The first phrase also contains a weak phrase break.

2.2 Tones

The tonal layer distinguishes between accentual and non-accentual tones. Two types of tone, H and L, are interpreted relative to each other (cf. Fig. 1).

- An asterisk marks accentual tones (H* / L*), non-accentual tones do not carry an asterisk (H / L).
- Downstep (!) or upstep (^) indicate the height of accentual / non-accentual tones relative to a preceding H tone (!H*, !H, ^H*, or ^H).
- The occurrence of a tonal target outside of the tone bearing syllable is indicated by the displacement label < (actual target pointing to the associated syllable to the left; Fig.1) or > (actual target pointing to the right).

2.3 Prominences

The prominence layer distinguishes three levels of perceived prominence (cf. [14]). Non-prominent syllables are not annotated.

- 1 = Weak prominence: Typically caused by metrical strength or tonal events. Examples for level 1 prominence are post-focal prominences in a reduced pitch register [17], partial deaccentuation [14], rhythmically determined accents [4], phrase accents [12][11], or post-lexical stress (*‘Druckakzent’*) [9].
- 2 = Strong prominence: Typically caused by syllables that are associated with a pitch accent, irrespective of the position of the accent in the phrase (cf. accents in first phrase of Fig. 1).
- 3 = Emphasis, extra strong prominence: Assigned for a clear and distinct marking of prominence beyond the strong prominence of a pitch accent. This level of prominence does not refer simply to a prosodically marked focus or the nuclear accent of the phrase, but often to an attitudinal, emphatic production [16], [22].

2.4. Comments

Like in [1] a layer for comments allows to indicate uncertainties about prosodic labels by means of a question mark (cf. Fig. 1), or to indicate phenomena that cannot be captured otherwise.

3. THE ANNOTATION PROCESS

The prosodic annotation is carried out in a number of steps, from left to right, in three distinct layers that need to be annotated independently. For instance, a prominence label does not necessarily entail a co-occurring tonal label. The annotation process is as follows:

1. “Phrase” layer – identify phrase boundaries: Identify and label the start and end of a strong

boundary (% ... %). If any, identify and label a weak boundary within that phrase (% ... - ... %). The hierarchical representation of phrases implies that a phrase with a weak boundary may never occur outside of a phrase with a strong boundary.

2. “Prominence” layer:
Add a prominence label within the respective syllable [25].
3. “Tone” layer – label the tones from left to right:
 - (a) Assign a left boundary tone below the phrase label. The default left-edge boundary tone is L. If the phrase starts with a distinctly high pitch, assign a high boundary tone (unless the high contour can be explained by an H tone in the first syllable). The phrase ends with a tone on the right boundary below the phrase label (H or L). If the end of a prosodic phrase coincides with the beginning of the next phrase, two tone labels need to be provided – but only if the tonal values differ. For example, a phrase may end with high pitch, and the following phrase starts with low pitch (HL). Otherwise, one tonal label is sufficient (see the weak boundary in Fig. 1).
 - (b) Accentual H* or L* tones are labelled at the F0 peak or valley of the accented syllable. If this target occurs outside of the syllable, label the accentual tone in the middle of the accented syllable and use the appropriate displacement label “<” or “>” (cf. first accent in Fig. 1). Note that accentual tones must co-occur with a prominence label; move the prominence label accordingly if necessary.
 - (c) Relevant F0 turning points that are perceived before and/or after the accentual tone indicate the presence of a tone; these non-accentual tones are either L or H.

Note some implications and further rules:

H tone labels can be modified with the diacritics for downstep “!” and upstep “^”, which are interpreted locally in relation to a preceding H tone in the same phrase.

Prominence labels can occur with and without a tonal label.

Prototypically, a prosodic phrase with a strong boundary contains at least one prominence of level 2 and one accentual tone H* or L*. DIMA allows for exceptions (e.g. prominences without tones or phrases without prominences), which are likely to occur in spontaneous speech data.

4. INTER-ANNOTATOR AGREEMENT

To evaluate the quality of the proposed consensus system we ran an inter-annotator agreement study on three different speech styles with two annotators. We thus tested our claim that the annotation

guidelines are transparent and easy to apply, such that we reach a high inter-annotator agreement.

4.1 Speech data

The data for the inter-annotator agreement study was composed of three different speech styles, i.e. read speech, and spontaneous monologue and dialogue. Read speech examples were taken from a news broadcast [7] and the dialogues were part of the Kiel corpus [15], [23]. The monologues were taken from a corpus of advisory speech in the context of mobile phones, which in total consists of 13 monologues on different topics, e.g. multimedia or business applications of mobile phones [18]. The monologues were non-scripted speech produced by two professional salesmen.

4.2 Procedure

Two graduate students who are familiar with the acoustic analysis of speech, intonation analysis and GToBI were trained with DIMA in two separate sessions of about one and a half hours each. The first session involved a thorough explanation of the distinct annotation layers and conventions. About 15 phrases from the monologues served as training materials. Note that in-depth training materials still have to be developed. The second training session was a discussion of the training materials and problems that arose by annotating the training speech samples. Both annotators annotated approximately one minute of speech in each data set.

4.3 Reliability measurement

Inter-annotator agreement refers to Cohen’s Kappa (κ) [6], which calculates the agreement between two annotators considering the agreement that would be predicted by chance. Although the interpretation of κ is under discussion, we consider a $\kappa > 0.8$ as high quality of annotation agreement, and a $0.67 < \kappa < 0.8$ ‘allowing for tentative conclusions’ [5].

4.4 Results and discussion

Table 2 shows an overview of total word counts and prosodically annotated words across the three speech styles. The total number of words that received a prosodic label (annotated words) differs from the number of words that received a prosodic label by both annotators (agreed words) showing some degree of disagreement. Although read speech and monologue data seem to allow about 10% higher agreement on average than dialogue, this may be explained by the fact that the dialogue speech contained a number of phrases where it was hard to

decide whether prominence and/or tone was present at all. These phrases contained whispered speech or repetitions of words as individual prosodic phrases with a strong boundary.

For the comparison of reliability measures across the three speech styles, all labels of the three distinct annotation layers entered the analysis. Results revealed an overall reliable inter-annotator agreement (Table 3). Read speech seems to pose more difficulties to reach inter-annotator agreement than spontaneous speech, which yields higher coefficients for annotation agreement.

Table 2: Number of words per speech style split by total word count, annotated words receiving a prosodic label, and agreed words labelled by both annotators (total number and percentage).

Speech style	Words	Annotated words	Agreed words
read news	124	55	40 (72%)
dialogue	289	98	64 (65%)
monologue	171	62	45 (73%)

Table 3: Reliability measures (κ) per speech style.

Speech style	Kappa
read news	0.76
dialogue	0.89
monologue	0.83

Table 4: Reliability measures for boundary and corresponding tones, and prominence and corresponding tones, according to speech style.

Speech style	Boundary & Tones κ	Prominence & Tones κ
read news	0.94	0.65
dialogue	0.93	0.81
monologue	0.92	0.74

Table 5: Reliability measures (κ) and ratio of actual observed agreement (p_0) for boundary, tone, and prominence layer according to speech style.

Speech style	Boundary κ	Tone κ (p_0)	Prominence κ (p_0)
read news	0.72	0.38 (63%)	0.36 (78%)
dialogue	0.90	0.68 (83%)	0.41 (80%)
monologue	0.77	0.27 (60%)	0.46 (91%)

Comparing the individual prosodic events across speech styles we calculated reliability measures a) for prosodic boundaries and their tonal labels, and b) for prominence ratings and the corresponding tones separately (Table 4). The agreement for boundaries and corresponding tones was very high. This shows

that boundaries were detected reliably, both in general and across different speech styles. The agreement for prominence and corresponding tones was lower, yet reliable, for spontaneous speech, as the $\kappa > 0.67$ shows. The reduction of complexity in annotation as proposed in DIMA thus leads to a high inter-annotator agreement, as was also shown for ToBI, where a relatively high agreement was achieved for accentual tones only [27].

Analysing each layer of annotation separately, we observe a dramatic reduction of reliability for the layers of tone and prominence (Table 5). However, the ratio of actually observed agreement (p_0) is high, which shows a weakness of the Kappa statistics when analysing data categories with large differences in their distribution. For instance, level 2 prominence occurs most frequently in annotated data since, prototypically, each proper prosodic phrase with a strong boundary contains at least one level 2 prominence. Hence, prominences at levels 1 and 3 are much less frequent. This kind of skewed distribution leads to a low κ despite the observed high agreement (p_0) of 80 to 90%. A similarly skewed distribution of tonal categories arises because accentual tones occur much more frequently than non-accentual tones, the latter depending on the presence of an accentual tone.

5. CONCLUSION

This paper reported on a consensus system for the prosodic annotation of German, set-up in order to achieve compatible data annotations from different research groups working in the field. The consensus system represents those aspects of tonal structure which are accounted for in the different phonological models used. We obtained high coefficients for annotation agreement, which are as good or even better than for similar annotation systems like [27]. We conclude that the proposed consensus system can be applied successfully. A website of the DIMA project presents detailed guidelines for transcription and will be updated with further developments of the system and training materials: <http://dima.uni-koeln.de/>.

Acknowledgements

This research was supported by German Research Association (DFG) grants to some of the authors: SFB 632, projects D5 and T2, SFB 732, projects A4 and INF, SPP 1234, as well as DFG projects GR 1610/5 and BA 4734/1, and a fund from the German Ministry for Education and Research – BMBF Grant Nr. 01UG0711.

6. REFERENCES

- [1] Beckman, M. E., Ayers-Elam, G. 1997. Guidelines for ToBI Labelling, Version 3. Ohio State University. http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf.
- [2] Beckman, M. E., Pierrehumbert, J. 1986. Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255–309.
- [3] Boersma, P., Weenink, D. 2013. Praat: doing phonetics by computer [Computer program].
- [4] Calhoun, S. 2010. How does informativeness affect prosodic prominence? *Language and Cognitive Processes*, 25, 1099–1140.
- [5] Carletta, J. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.
- [6] Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46
- [7] Deutschlandradio 2014. Nachrichtensendung, 15.00, 14.06.2014. Deutschlandfunk, Köln. (<http://www.deutschlandfunk.de/nachrichten.353.de.html>)
- [8] Féry, C. 1993. *German Intonational Patterns*. Tübingen: Niemeyer.
- [9] Grice, M., Baumann, S. to appear. Intonation in der Lautsprache: Tonale Analyse. In Primus, B. & Domahs, U. (eds.), *Handbuch Laut, Gebärde, Buchstabe*. De Gruyter, Reihe Sprachwissen.
- [10] Grice, M., Baumann, S., Benzmüller, R. 2005. German Intonation in Autosegmental-Metrical Phonology. In Jun, S.-A. (ed.), *Prosodic Typology*, 55–83. Oxford: OUP.
- [11] Grice, M., Baumann, S., Jagdfeld, N. 2009. Tonal association and derived nuclear accents: The case of downstepping contours in German, *Lingua*, 119: 881–905.
- [12] Grice, M., Ladd, D.R., Arvaniti, A. 2000. On the place of phrase accents in intonational phonology. *Phonology*, 17:2, 143–185.
- [13] Gussenhoven, C. 2004. *The Phonology of Tone and Intonation*: Cambridge: CUP.
- [14] Kohler, K. J. 1991. A Model of German Intonation. In *AIPUK 25. Studies in German Intonation*, 295–360. Kiel: IPdS.
- [15] Kohler, K. J. 1996. Labelled data bank of spoken Standard German: The Kiel Corpus of Read/Spontaneous Speech. *Proc. ICSLP*, 73–77.
- [16] Kohler, K. J. 2004. Prosody Revisited: FUNCTION, TIME, and the LISTENER in Intonational Phonology. *Proc. Speech Prosody 2004*, Nara, Japan, 171–174.
- [17] Kügler, F., Féry, C. submitted. Postfocal downstep in German. Submitted to *Language and Speech*.
- [18] Kügler, F., Smolibocki, B., Stede, M. 2014. Information status and prosody in a corpus of non-scripted spoken German. Poster at *Linguistic Evidence 2014*, Tübingen.
- [19] Ladd, D. R. 1996/2008. *Intonational Phonology*. Cambridge: CUP.
- [20] Mayer, J. 1995. Transcription of German intonation: the Stuttgart System. University of Stuttgart: <http://www.ims.uni-stuttgart.de/institut/arbeitsgruppen/phonetik/papers/STGTsystem.pdf>
- [21] Nespor, M., Vogel, I. 2007. *Prosodic phonology*. Berlin: Mouton De Gruyter.
- [22] Niebuhr, O. 2010. On the phonetics of intensifying emphasis in German. *Phonetica* 67, 170–198.
- [23] Niebuhr, O., Kaernbach, C., Pfitzinger, H., Schmidt, G. 2015. The Kiel Corpora of "Speech & Emotion" - A Summary. *Proc. 41st DAGA conference*, Nuremberg, Germany.
- [24] Peters, J. 2014. *Intonation*. Heidelberg: Winter.
- [25] Peters, B., Kohler, K. J. 2004. Trainingsmaterialien zur prosodischen Etikettierung mit dem Kieler Intonationsmodell KIM. MS, http://www.ipds.uni-kiel.de/kjk/pub_exx/bpkk2004_1/TrainerA4.pdf
- [26] Pierrehumbert, J. B. 1980. *The phonology and phonetics of English intonation*, PhD Thesis, MIT.
- [27] Yoon, T., Chavarría, R., Cole, J., Hasegawa-Johnson, M. 2004. Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. *Proc. ICSLP 2004*, 2729–2732.