ANNOTATION OF GERMAN INTONATION: DIMA COMPARED WITH OTHER ANNOTATION SYSTEMS

Frank Kügler[1], Stefan Baumann[2], Bistra Andreeva[3], Bettina Braun[4], Martine Grice[2], Jana Neitsch[4],
Oliver Niebuhr[5], Jörg Peters[6], Christine T. Röhr[2], Antje Schweitzer[7], Petra Wagner[8]

[1]Goethe-University Frankfurt, [2]University of Cologne, [3]Saarland University, [4]University of Konstanz, [5]University of Southern Denmark, [6]The University of Oldenburg, [7]University of Stuttgart, [8]Bielefeld University
kuegler@em.uni-frankfurt.de

ABSTRACT

Annotating intonation is a considerable challenge, since not only intonational form but also its meaning are complex in terms of their internal make-up and contextual variation. Since the advent of the autosegmental-metrical approach to intonation in the 1980s, the annotation of intonation has continued to be a matter of debate, witnessed by the current discussion around the proposed International Prosodic Alphabet (IPrA), with a reported need for a more surface-related annotation that serves as a basis for phonological categorisation. The DIMA system accounts for such a level by providing a phonetically informed annotation of an intonation contour that nevertheless reflects its phonological core. DIMA is a consensus system for the annotation of German intonation that analyses intonation at three distinct levels: phrasing, tones and prominences. The present paper compares DIMA with other annotation systems such as GToBI, ToGI, IViE, KIM, RaP, and IPrA.

Keywords: Annotation, intonation, AM phonology

1. INTRODUCTION

This paper compares DIMA (*Deutsche Intonation: Modellierung und Annotation*), a consensus system for annotating German intonation, with other annotation systems. DIMA integrates phonetic and phonological criteria in the process of mapping the continuous speech signal onto discrete labels. It can thus be considered a 'phonetically informed phonological annotation system' and aims to apply cross-linguistically. A core property is that a proper phonological analysis of the data in terms of *on-ramp* [9] or *off-ramp* models [8,12,23] of intonation can be postponed until a later stage [17,19]. The idea of a surface-related tier is found in a number of systems for intonation annotation [3,6,8,14,15], but unlike those systems, DIMA decomposes the complex signal on three independent layers: phrase boundaries, tones and prominences.

The systems compared here (except for KIM) have their roots in AM Phonology [10,20] in which modulations of speech melody are treated as a sequence of interpolated tonal targets (H(igh) and L(ow) tones) that may be grouped into categorically distinct, abstract composite tones. Functionally, tones are organized into pitch accents (highlighting) and boundary tones (delimiting) that are associated with a particular meaning. These models provide a *phonological* representation of intonation that is separate from the details of *phonetic* implementation.

DIMA is special in providing a phonetically-oriented, perceptually grounded, 'pre-phonological' annotation for German. It is theory-neutral, insofar as in later phonological analysis it allows for a translation into model-specific types of tonal label sequence. Table 1 shows the inventory of annotation symbols [17].

Two phrase levels are distinguished: phrases with a strong ('%') or a weak boundary ('-'). Diacritics indicate changes in pitch register of whole phrases, either lowered ('!') or raised ('^'). Disfluencies causing a perceptual phrase break are annotated with '&'. The question mark '?' indicates uncertainties at all layers.

Accentual tones labelled as either H* or L* within the accented syllable are distinguished from non-accentual tones labelled at perceptually salient F0 peaks and valleys in the vicinity of accentual tones. Hence, DIMA neither uses any leading nor trailing tones; the interpretation of complex pitch accents is a matter of (later) phonological analysis. Lowered ('!') or raised ('^') tones relative to preceding tones are indicated by diacritics to the left of the tone label. If the tonal target pertaining to an accentual tone is realized outside the prominent syllable, the diacritics '<' or '>' are used to indicate the location of the actual pitch target.

Finally, DIMA distinguishes between three prominence levels: '1' weak prominence (cued tonally or by other cues), '2' normal prominence (usually correlated with tonal events) and '3' extra-strong prominence [2]. The fourth level, i.e. no prominence, is not explicitly marked in DIMA.

Table 1: Symbols for a DIMA annotation [17].

Layer	Phrase	Tone	Prominence
Label	% -	H* L* H L	1 2 3
Diacritics	! ^ & ?	! ^ < > ?	?

2. DIMA AND OTHER SYSTEMS

We will compare DIMA with different intonation systems for German and other languages, like German ToBI (GToBI) [9], German ToDI (ToGI) [23], IViE [7,8], the Kiel Intonation Model (KIM) [16,22], RaP [5] and the proposed International Prosodic Alphabet (IPrA) [14]. The comparison considers the criteria ‘phrase boundaries’, ‘prominences’, ‘tonal structure’, and ‘position on the phonetics-phonology scale’. Annotated examples comparing the German annotation systems mentioned above can be found in [31].

GToBI has become the standard annotation system for German intonation. It is a phonological model that is nevertheless more surface-oriented than the American English original ToBI (cf. [3]). ToGI (‘Transcription of German Intonation’) is an adaptation of ToDI [11] to German. Some of the differences between ToDI- and ToBI-style annotations are rooted in different conceptions of intonational phonology. For example, ToDI tone classes are defined in purely structural terms, whereas classical ToBI uses both structural and phonetic criteria. IViE (‘Intonational Variation in English’) was developed for comparative linguistic research in British English varieties. It departs from classical ToBI in providing a more clearly constrained accent inventory due to the need for a more transparent comparative transcription system for non-standard varieties. Both ToGI and IViE annotations account (mostly) for the pitch movement leading *off* the accented syllable (*off-ramp* analysis), which is more in line with the British School characterizations of falls and rises. In contrast, (G)ToBI and IPrA (see below) annotations use bitonal accents primarily to account for the pitch movement leading *towards* an accented syllable (*on-ramp* analysis).

KIM is one of the earliest phonological intonation models of German, and is often classified as contour-based as it regards limited sets of rising-falling peaks and (falling-)rising valleys as the basic building blocks. However, it is actually the local F0 minima and maxima inside the peaks and valleys whose alignment is considered to be directly phonological. The RaP (‘Rhythm and Pitch’) system is an alternative to ToBI-based prosody annotations. It aims at improving on certain issues with ToBI such as listeners’ difficulty to categorically differentiate similar pitch shapes or differences between break indices without any perceptual disjuncture, or the potentially categorical function of pitch span that cannot be adequately captured in ToBI [4]. Finally, IPrA [14] attempts to provide a set of pre-phonological, yet ‘phonetically categorical’ tone labels, aiming to label surface tonal variation of underlying phonological contours. DIMA’s perceptually-based manual annotation will finally be compared to automatic annotation methods.

2.1. Phrase boundaries

All systems combine a phrase boundary with either tonal (DIMA, GToBI, ToGI, KIM, IPrA) and/or rhythmical labels (RaP, IViE). However, only DIMA marks phrase boundaries on a separate tier (see Table 2). In GToBI, ToGI, KIM and IPrA the label for the level of phrasing is combined with tonal diacritics: e.g., a low intonation phrase boundary is marked as ‘L-%’ in GToBI, while the same contour is represented by ‘&2.’ in KIM, indicating a terminal fall to a very low value of the speaker’s pitch range.

Like DIMA, GToBI, ToGI and RaP explicitly differentiate between two levels of phrasing (corresponding to strength levels of prosodic boundaries), while IViE only has one. IPrA has at least three levels of phrasing, the intonation phrase (IP), the intermediate phrase (ip), and the accentual phrase (AP).

Both DIMA and RaP rely on perceptual disjuncture when annotating phrase boundaries, but DIMA allows for additional cues that annotators may take into account when making their judgments, such as pauses, pitch resets or voice quality changes. Moreover, in RaP boundary tones are only specified if they are markedly low or high. In contrast, DIMA requires a tonal labelling of each boundary.

Like DIMA, KIM and IPrA explicitly mark pitch register differences of whole phrases. Neither GToBI, ToGI, RaP nor IViE account for such differences.

Table 2: Comparing phrasing annotation.

Criteria:	DIMA	GToBI	ToGI	IViE	KIM	RAP	IPrA
Separ. phrase tier	✓	-	-	-	-	-	-
Phrase levels	2	2	2	1	-	2	3
Initial Bound.To	✓	(✓)	✓	✓	✓	-	(✓)
Final Bound.To	✓	✓	✓	✓	✓	✓	✓
Register changes	✓	-	-	-	✓	-	✓

2.2. Prominences

Four systems, namely DIMA, IViE, KIM and RaP, annotate tonal movements and rhythmic prominences on separate tiers. Consequently, prominent syllables do not need to co-occur with a corresponding tone or accent. In fact, prominent syllables in DIMA and RaP may either carry a pitch accent (marked by an accentual tone) or not (marked by a non-accentual tone or no tone at all). IViE additionally distinguishes prominent (stressed/accented) syllables (transcribed with a ‘P’) from non-prominent ones, while DIMA, KIM and RaP further account for differences between weak and strong prominences. RaP does not allow for an extra-strong 'emphatic prominence' which is part of DIMA and KIM [2,17,22]) (see Table 3).

A major difference between the prominence annotations in DIMA, and IViE and RaP is the interpretation of the association between prominences and tones. In DIMA, the prominence levels (none, 1, 2, 3) are entirely perceptual and may be aligned with a corresponding (non-)accentual tone or not. However, the DIMA guidelines state that a 'typical' pitch-accented syllable corresponds to prominence level 2 but they also allow for the annotation of non-tonal strong prominences, see [22] for examples. In contrast, IViE and RaP expect each starred tone to correspond to a metrical beat (i.e. a prominence) on the prominence tier (IViE) or on the rhythm tier (RaP), which in RaP expresses an additional perceptual strengthening of the corresponding metrical beat. That is, the RaP system differentiates four levels of prominence, namely *none, weak non-tonal prominence, strong non-tonal prominence* and *strong tonal prominence*, indicating that the decoupling of prominence and tones is not systematic here.

The other systems, GToBI, ToGI, and IPrA do not explicitly differentiate between prominence levels. They do use the star notation, however, which implies the presence of a prominent (stressed or accented) syllable. We do not regard the star as a specific expression of prominence marking here.

Table 3: Comparing prominence annotations.

Criteria:	DIMA	GToBI	ToGI	IViE	KIM	RAP	IPrA
Separ. prom. tier	✓	-	-	✓	✓	✓	-
Prom. marking	✓	-	-	✓	✓	✓	-
Levels of prom.	4	-	-	2	4	(4)	-
Prominence independent of tones	✓	-	-	-	✓	(✓)	-

2.3. Tonal structure

All systems use at least one tier for the annotation of the tonal structure (see Table 4). While IViE and IPrA distinguish between a phonetic and a phonological level of annotation, the majority of systems assume one tonal tier. In general, the extent to which an annotation can be called 'phonological' differs (cf. section 2.4). IViE transcribes the pitch movements surrounding prominent syllables on the phonetic (or target) tier (lower case letters for non-accentual tones and upper case letters for accentual tones) and the phonological categorisation of these pitch movements on the phonological tier. IPrA uses a broad phonetic annotation with 'phonetically categorical' [14] tone labels to label surface tonal variation of underlying phonological contours, and a corresponding phonological tier.

The central difference between DIMA and all other systems concerns the phonological level of annotation. This level is based on (mostly) right-headed (*on-ramp*) accent types in GToBI, IPrA and RaP, left-headed (*off-ramp*) pitch accents in ToGI and IViE, and timing-dependent contours in KIM. IPrA views the set of ToBI labels as phonetic categories comparable to the IPA symbols for segmental transcription, which should be capable of capturing the contrastive pitch events of any language. GToBI is less surface-oriented, its tonal inventory is at the same time underspecified (initial boundary tones, e.g., are only annotated in the case of an exceptionally high beginning of an intonation phrase) and rather restrictive, e.g. in that the choice of IP boundary tones is fairly limited. In line with the stronger phonological perspective mentioned above, the reasons for these limitations lie in the assumption that only those categories should be transcribed that are attested to indicate a meaning difference in a given variety. Hence, the two basic functions of intonation, highlighting and delimiting, are reflected at a combined level of annotation for pitch accents and boundary tones in GToBI. DIMA departs from this view in disentangling the complexity of the intonation signal (i) by breaking up (potential) accent types into accentual and non-accentual tones and (ii) by keeping boundary tones separate and linking them to the phrase tier.

In order to allow for a less restricted interpretation and representation of the intonation contour, DIMA reduces the phonological analysis of intonation to the determination of the tonal 'core' of an accent (the 'starred tone') and focusses more on phonetic aspects of the contour (e.g. the exact position of F0 minima and maxima). The latter aspect is a central property of KIM as well (indicating early, medial and late peaks); KIM defines the local minima and maxima with regard to actual physical values in the (macro-prosodic) F0 contour instead of conceptualizing these tonal targets as abstract high or low entities at the level of perception. In line with such a more fine-grained level of analysis, which aims at providing the prerequisite for a closer investigation of, for instance, differences in peak alignment, the DIMA annotation is based on the syllable. In contrast, a GToBI annotation takes the word as the basic textual domain of annotation – like most AM models.

IPrA symbols are pre-defined in acoustic shape such that a label decision can be made without a phonological analysis, and on the basis of the speech signal. Thus, labelling is perceptually less informed, and it remains to be shown to which extent phonetic timing distinctions between language varieties would provide any benefit for comparative research purposes (e.g. rising accents in Northern and Southern German [1] that would result in phonetically different accent categories, i.e. L*+H in Southern and L+H* in Northern German).

Table 4: Comparing tonal structure and annotation.

Criteria:	DIMA	GToBI	ToGI	IViE	KIM	RAP	IPrA
Tone tiers	1	1	1	2	1	1	2
Types of tones/accents	a	b	c	a/c	d	b	b

a) accentual and non-accentual tones; b) preference for right-headed pitch accents; c) preference for left-headed pitch accents; d) local minima and maxima and their timing information.

2.4. Position on the phonetics-phonology scale

The systems compared here ascribe different degrees of importance to a phonetic or phonological analysis (1). The explicitly surface-related tiers of DIMA, IViE and IPrA can be allocated at the phonetic side of the scale. DIMA is yet more phonetically-oriented than IViE / IPrA as it decomposes not only the complex signal but even tonal events. On the other side of the scale, GToBI, the phonological tier of IViE, and ToGI represent phonological analyses of intonation, with ToGI being the clearest case of a phonological model. KIM is placed in the middle of the scale in (1) because it takes phonetically specific tonal targets and shapes as the pivots for phonological distinctions of paralinguistic meanings within the two contour classes of peaks and valleys.

In sum, DIMA, IViE, and IPrA annotate the intonational events on separate, surface-oriented tiers. This enables a differentiation between phonological and phonetic aspects of intonation and provides information about the mapping between phonological categories and their phonetic implementation.

DIMA, IViE, IPrA KIM RaP GToBI, IViE, ToGI
(1) phonetic ↔ phonological

2.5. Advantages of manual annotation

The most recent systems, i.e. IPrA and DIMA, advocate a more phonetically-oriented annotation, potentially as a basis for further phonological analysis. This raises the question whether automatic methods of annotation should be preferred. After all, if annotation not only relies on listening but also on scrutinizing the F0 contour, automatic methods might be more successful and consistent than human annotators. There is indeed a growing body of work on automatic detection or recognition of pitch accents and phrase boundaries [13,24–30]. The state of the art for American English for instance is around 65-70% accuracy for pitch accent recognition and 73-87% for phrase boundary recognition. However, these rates are only obtained because the most frequent categories (including ‘no accent’ or ‘no boundary’ cases) are recognized very well, while moderately frequent and infrequent accent or boundary categories are often not detected. Furthermore, it should be noted that automatic methods have mostly focused on ToBI-style phonological categories, while DIMA involves a more phonetically-oriented annotation. It remains to be seen how automatic methods perform for such labels.

Hence, we argue that automatic methods are useful in speech technology tasks, but that they are not (yet) good enough to replace a manual annotation of the tones and phrase tiers, which implies both an auditory and a signal-based analysis. Certainly, manual (i.e. perceptual) annotation requires the training of annotators and the assembly of valid training materials. These are currently being developed. The phonetic nature of DIMA will probably ease its learning, as first inter-rater reliability studies showed substantial agreement between annotators [18].

3. CONCLUSION AND OUTLOOK

With DIMA, we believe, it is not only possible to make annotations using different models of German intonation comparable but also to initiate work on under-described languages and language varieties, as well as on second language and child speech. The annotation process of DIMA does not require a complete phonological analysis of a language and its intonational grammar. The idea is to separate the annotation from the phonological interpretation and thus to provide a basis for a translation of DIMA labels into different intonation models [17]. As a possible application, in L2 speech, for instance, a complete phonological grammar cannot be obtained in the first place since residuals of the L1 and non-matches of L2 categories are characterizing properties of L2 speech [21]. With this in mind, DIMA allows for a phonetically-oriented annotation, which can be interpreted phonologically in that DIMA events can be traced back to L1, or forth to L2 categories. Future research will reveal whether DIMA is transferable to other languages and/or language varieties.

4. ACKNOWLEDGEMENTS

We thank Stephanie Berger for her support. This research was funded by grants from the DFG to some of the authors: KU 2323/4-1, BA 4734/1-2, SFB 1252 “Prominence in Language”.

5. REFERENCES

[1] Atterer, M., Ladd, D.R. 2004. On the phonetics and phonology of "segmental anchoring” of F0. Evidence from German. *Journal of Phonetics* 32, 177–197.

[2] Baumann, S., Niebuhr, O., Schroeter, B. 2016. Acoustic Cues to Perceived Prominence Levels. Evidence from German Spontaneous Speech. In: *Proc. Speech Prosody 2016*, 711–715.

[3] Beckman, M. E., Hirschberg, J., Shattuck-Hufnagel, S. 2005. The Original ToBI System and the Evolution of the ToBI Framework. In: Jun, S.-A. (ed.), *Prosodic Typology*. Oxford: OOUP, 9–54.

[4] Breen, M., Dilley, L.C., Kraemer, J., Gibson, E. 2012. Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory* 8, 277–312.

[5] Dilley, L.C., Brown, M. 2005. The RaP (rhythm and pitch) labeling system. v. 1.0, MIT.

[6] Gilles, P. 2005. *Regionale Prosodie im Deutschen*. Berlin: De Gruyter.

[7] Grabe, E. 2004. Intonational variation in urban dialects of English spoken in the British Isles. In: Gilles, P., Peters, J. (eds.), *Regional Variation in Intonation*. Tübingen: Niemeyer, 9–31.

[8] Grabe, E., Nolan, F. 2001. Modelling intonational Variation in English. The IViE system. In: *Prosody 2000. Speech Recognition and Synthesis*, Krakow, 51–57.

[9] Grice, M., Baumann, S., Benzmüller, R. 2005. German Intonation in Autosegmental-Metrical Phonology. In: Jun, S.-A. (ed.), *Prosodic Typology*. Oxford: OUP, 55–83.

[10] Gussenhoven, C. 2004. *The Phonology of Tone and Intonation*. Cambridge: CUP.

[11] Gussenhoven, C. 2005. Transcription of Dutch intonation. In: Jun, S.-A. (ed.), *Prosodic Typology*. Oxford: OUP, 118–145.

[12] Gussenhoven, C. 2016. Analysis of Intonation: the Case of MAE_ToBI. *Laboratory phonology: Journal of the Association for Laboratory Phonology* 7, 1–35.

[13] Hasegawa-Johnson, M., Chen, K., Cole, J., Borys, S., Kim, S.-S., Cohen, A., Zhang, T., Choi, J.-Y., Kim, H., Yoon, T. 2005. Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus. *Speech Communication* 46, 418–439.

[14] Hualde, J.I., Prieto, P. 2016. Towards an International Prosodic Alphabet (IPrA). *Laboratory phonology* 7, 1–25.

[15] Jun, S.-A., Fletcher, J. 2014. Methodology of studying intonation. In: Jun, S.-A. (ed.), *Prosodic typology II*. Oxford: OUP, 493–519.

[16] Kohler, K.J. 1991. A Model of German Intonation. In: Kohler, K. J. (ed.), *Studies in German Intonation*, Kiel, 295–360.

[17] Kügler, F., Baumann, S. 2019. Annotationsrichtlinien DIMA. V4.0, Köln [http://dima.uni-koeln.de].

[18] Kügler, F., Baumann, S., Schweitzer, A., Wagner, P. 2017. Reliabilität zwischen Annotatoren in DIMA – Ein Vergleich der Annotation von Experten und einer Trainingsgruppe. Poster at P&P-13, Berlin.

[19] Kügler, F., Smolibocki, B., Arnold, D., Baumann, S., Braun, B., Grice, M., Jannedy, S., Michalsky, J., Niebuhr, O., Peters, J., Ritter, S., Röhr, C.T., Schweitzer, A., Schweitzer, K., Wagner, P. 2015. DIMA - Annotation guidelines for German intonation. In: *Proc 18th ICPhS, Glasgow*, paper 317.

[20] Ladd, D.R. 2008. *Intonational Phonology*, 2nd ed., Cambridge: CUP.

[21] Mennen, I. 2015. Beyond Segments: Towards a L2 Intonation Learning Theory. In: Delais-Roussarie, E., Avanzi, M., Herment, S. (eds.), Prosody and Language in Contact: L2 Acquisition, Attrition and Languages in Multilingual Situations. Berlin: Springer, 171–188.

[22] Niebuhr, O. 2019. The Kiel Intonation Model - KIM. In: Barnes, J., Shattuck-Hufnagel, S. (eds.), *Prosodic Theory and Practice*. Cambridge: MIT Press.

[23] Peters, J. 2018. Phonological and semantic aspects of German intonation. *Linguistik online* 88, 88–107.

[24] Ren, Y., Kim, S.-S., Hasegawa-Johnson, M., Cole, J. 2004. Speaker-independent automatic detection of pitch accent. In: *Proc. Speech Prosody*, Nara, 521–524.

[25] Rosenberg, A. 2010. Classification of prosodic events using quantized contour modeling. In: *Proc. HLT-NAACL*, 721–724.

[26] Schweitzer, A. 2010. *Production and Perception of Prosodic Events-Evidence from Corpus-based Experiments*. PhD Thesis, Stuttgart.

[27] Sridhar, V.K.R., Bangalore, S., Narayanan, S. 2008. Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework. *IEEE Trans. Audio Speech Lang. Process*. 16.

[28] Stehwien, S., Vu, N.T. 2017. Prosodic Event Recognition Using Convolutional Neural Networks with Context Information. In: *Interspeech 2007*, 2326–2330.

[29] Sun, X. 2002. Pitch accent prediction using ensemble machine learning. In: *Proc. ICSLP-2002*, 953–956.

[30] Wang, X., Takaki, S., Yamagishi, J. 2016. Enhance the Word Vector with Prosodic Information for the Recurrent Neural Network Based TTS System. In: *Interspeech 2016*, 2856–2860.

[31] http://dima.uni-koeln.de/?page_id=353 (annotated examples comparing German annotation systems)