THE PHONETICS–PHONOLOGY–SYNTAX INTERFACE: A COMPUTATIONAL IMPLEMENTATION

Tina Bögel, Romi Hill, Justin Hofenbitzer, Tianyi Zhao

University of Konstanz
firstname.lastname@uni-konstanz.de

ABSTRACT

This paper introduces a new, computationally implemented end-to-end system for German that takes a speech signal as input, interprets the phonetic data in phonological/prosodic terms, and makes the results available to a linguistically deep computational grammar. The grammar uses the provided information to disambiguate syntactically ambiguous structures, thus reducing overgeneration. A system evaluation showed promising results for this new combination of automatic speech signal analysis and computational grammars, which is a significant step towards a fine-grained linguistic analysis including all grammar modules and hence towards real automatic speech understanding.

Keywords: German, syntactic ambiguities, end-to-end system, automatic speech understanding

1. INTRODUCTION

Systems that allow for automatic speech recognition (ASR) and the identification of prosodic events are often used in phonetic and prosodic research (see, e.g., [1]). For research on German, MAUS [2, 3] is frequently utilized to automatically annotate segments and words. For the identification of prosodic events (e.g., accents or boundaries), available systems include the *Prosodizer*, which assigns pitch accents and boundary tones during speech recognition and synthesis [4, 5], and the prosody module of the Verbmobil system, which integrates word-based annotation and classification of boundaries and accents for German dialogues [6]. [7] trained a number of classifiers on acoustic, phonological, and basic morphosyntactic attributes of German reaching recognition accuracy rates of up to 86% for the occurrence of accents, and 93% for the occurrence of larger boundaries.

Concerning speech synthesis, these approaches went beyond the sole interpretation of acoustic cues and additionally included basic morphosyntactic information (e.g., part-of-speech tags). As this was found to considerably improve accuracy rates, it can be deduced that the algorithms would benefit even more if a fine-grained linguistic analysis of the underlying string were available in the form of syntactic or semantic representations. This information could in turn be associated with specific prosodic events (e.g., grouping or focus). The same is true for ASR which would benefit greatly from the inclusion of deep linguistic information, to allow for real automatic speech understanding (ASU).

While linguistically deep computational grammars (CGs) are available for text-based input and for a number of frameworks (a.o., LFG [8], HPSG [9]), these grammars are unable to process spoken language. As such, these CGs would profit considerably if prosodic information was available for linguistic interpretation.

This paper introduces a new system that bridges the gap between the automatic recognition of prosodic events on the one hand, and CGs on the other. The implementation includes a representation of the speech signal in phonetic and phonological/prosodic terms, where the latter categorical representation enables the CGs to prosodically disambiguate syntactically ambiguous structures. As a consequence, the CG is able to return the correct and linguistically fine-grained representation and thus takes a huge step towards real automatic speech understanding.

2. THE DATA

In the following syntactically ambiguous German sentence, the NP_2 (*der Freundin* 'the friend') can either be part of a genitive structure (meaning a) or constitute a separate dative argument (meaning b).

(1) Sie sahen, dass [der Partner]$_{NP1}$
They saw that the.MASC.NOM partner

[der Freundin]$_{NP2}$ fehlte
the.FEM.GEN/DAT friend was.missing

a) "They saw that the friend's partner was missing."
b) "They saw that the friend missed the partner."

Such syntactically ambiguous structures result in overgeneration, i.e., the CG returns several possible solutions (Figure 1).

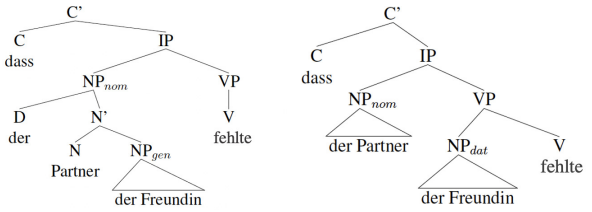

Figure 1: Two syntactic trees resulting from example (1): Genitive structure on the left, dative structure on the right.

These structures can be disambiguated by means of prosody [10] and several studies have shown this to be true for German as well [11, 12]. In a production experiment, [13] showed that speakers place a prosodic phrase boundary between the two NPs in the dative structure in (1), but not in the genitive. Acoustic indicators for the boundary included the lengthening of the last syllable of the first NP ($p < 0.01$), a rise with a following reset in F_0 ($p < 0.01$), and occasionally a pause ($p < 0.05$) between the two NPs in the dative structure.

3. THEORETICAL ANALYSIS

The implementation presented here is couched within the (generative, nontransformational) framework of Lexical-Functional Grammar (LFG) [14, 15]. The modular architecture of LFG proposes different representative structures for separate linguistic aspects (e.g., trees for syntax, see Figure 1), which constrain each other through mathematically well-defined functions. LFG comes with XLE, a state-of-the-art grammar development platform [16, 17], which allows researchers to build industrial strength CGs covering a wide range of languages.[1] However, while these grammars are well-established for syntactic and semantic analyses of texts, they have so far been unable to process spoken language.

This paper follows the theoretical approach to the interface proposed in [18, 13] which discusses the integration of speech signal information into the overall grammar, and develops a representation of the phonetic and prosodic information: the *p-diagram*, a syllable-based representation of the speech signal over time. The prosodic information provided in the p-diagram can subsequently be processed by other modules (e.g., syntax or semantics), thus allowing for a deep linguistic analysis. Under this approach, the input to the grammar consists of a speech signal annotated with syllables. In a first step, each of these syllables receives a vector which, in addition to the related segments, encodes acoustic information relevant to this syllable, e.g., duration or mean F_0. This information is stored at the *signal level* in the p-diagram. Figure 2 shows the p-diagram fragment for the six syllables related to the string *der Partner der Freundin* (ex. (1)), where, e.g., the vector for the first syllable of *partner* has the index S_2, the segments [pa6t] (in SAMPA [19]), the syllable length (0.25s), and a mean F_0 value (181Hz) at the signal level (lower part of Figure 2).

...	...	...	...	...	...		interpretation
PHRASING	-	-	$)_\varphi$	$(_\varphi$	-	-	↓
SEMIT_DIFF	...	-1	6.8	-4.3	-1.9	2.6	
GToBI	-	L*	+H H-	-	L*	+H	
DURATION	0.15	0.25	0.25	0.13	0.31	0.19	signal
FUND. FREQ.	192	181	269	209	188	218	↓
SEGMENTS	[de:6]	[pa6t]	[n6]	[de:6]	[fROYn]	[dIn]	
VECTORINDEX	S_1	S_2	S_3	S_4	S_5	S_6	

Figure 2: The p-diagram's signal level and interpretation level for *der Partner der Freundin* (example (1))

Based on this initial (phonetic) information from the speech signal, a categorical (phonological) *interpretation level* can be derived and added to the p-diagram (upper part of Figure 2), which can include, e.g., differences between adjacent semitones, prosodic phrase boundaries, or a (G)ToBI annotation. For example, a strong rise in F_0 and a following drop (S_2–S_4) and a comparatively long duration on the last (unstressed) syllable of *Partner* (as seen at S_3: [n6]) are strong indicators for a phonological phrase boundary and also justify an L*+H accent. As a result, PHRASING = $)_\varphi$ is added to the syllable's vector at the interpretation level and the L*+H accent is distributed over the associated syllables for the GToBI attribute.

While the p-diagram representation was developed with regard to LFG, it is an encapsulated representation that can be plugged into any modular framework. It is furthermore very adaptable: It could be based on segments or could include any other attribute of interest (e.g., intensity or F_{1-3}). The advantage of the p-diagram is that it allows for a compact and formal representation in terms of ordered vectors with attributes and associated values which makes it accessible to other modules of grammar (see [18] for formal details).

4. COMPUTATIONAL IMPLEMENTATION

The computational implementation follows the theoretical approach to the interface between signal and grammar, as presented in the previous section. Its accuracy is demonstrated by prosodically disambiguating syntactic structures as given in (1).

4.1. Information extraction and preparation

Input to the implementation is a speech signal annotated with (SAMPA) syllables in Praat [20].[2] In a first step, a Praat script extracts information that is relevant for the p-diagram's signal level (Fig. 2): segments, duration, and the mean F_0-value for each syllable vector.

For a more fine-grained analysis of the pitch, the script divides each syllable into five even-spaced subintervals, takes the mean F_0-values of each subinterval and turns the values into semitones, thus effectively normalizing duration and pitch. Each subinterval is also tagged for position within the syllable, either as central or as preceding or following a syllable boundary.

4.2. Interpreting the pitch

In a second step, the raw values from the speech signal are interpreted in terms of categories in order for them to become 'meaningful' for other modules of grammar. In addition to the semitones and the differences between these semitones indicating falls and rises, the implementation also uses residuals of a linear regression calculated based on the pitch values of a given speech signal. The residuals return the distance each value has from this line and are thus a good measure to describe deviations from the average while at the same time including the tendency of the signal's general pitch contour.

Taken together, semitones and residuals allow for the detection of deviations from the norm in the signal, i.e., maximums (H) and minimums (L). To avoid microprosodic effects, the distance between any Hs and Ls has to consist of at least one syllable. Slopes to and from a H/L tone are calculated based on a ratio between the semitones of adjacent Ls and Hs and the distance (the number of subintervals) that lies between them. The resulting values indicate whether the associated slopes are steep or flat.

In order to mark both categories, accent and slope, in one representation, the following system was devised, where each level of L or H is characterised by a particular height and shape of the slopes leading to it (lead) and following it (tail).

Cat.	Max/Min	lead	tail
H4/L4	Max/Min	steep	steep
H3/L3	Max/Min	steep	flat
H2/L2	Max/Min	flat	steep
H1/L1	Max/Min	normal	flat

Table 1: (Part of the) system of pitch accents and slopes in the computational implementation

H4 and L4 thus represent accents where the lead and the tail show a strong rise/fall respectively, while H1 and L1 have a relatively flat lead and tail. L2/L3 and H2/H3 are positioned between these two extremes, with each having a slightly different shape depending on the slopes. These tone values are stored in the interpretation level of the p-diagram (Fig. 2), where they replace the traditional GToBI values in order to facilitate (and simplify) the automatic interpretation by other modules of the grammar.

4.3. Matching against the lexicon

In order to acquire the correct syntactic string, the syllable-based segmental string is matched exhaustively against a lexicon which includes phonological and morphosyntactic material in form of a finite-state transducer [21]. Following [22], the lexicon stores information on the individual segments and the metrical frame, i.e., (in the case of German) the number of syllables, lexical stress and prosodic word status for each word in the p(honological)-form. This p-form is associated with a specific s(yntactic)-form which is then used for the syntactic parse. Figure 3 shows the p-form and s-form for the noun *Freundin*, a trochaic prosodic word, and the determiner *der*, a single, prosodically underspecified syllable.

s-form		**p-form**	
(↑ PRED)	= 'Freundin'	SEGMENTS	/f R OY n d I n/
↑ NUM)	= sg	METR. FRM	$(' \sigma \sigma)_\omega$
(↑ GEND)	= fem		
(↑ PRED)	= 'der'	SEGMENTS	/d e 6/
(↑ CASE)	= {gen | dat}	METR. FRM	σ

Figure 3: (Simplified) lexical entries for *der* and *Freundin*

Once the segmental string (*de6.pa6t.n6.de6. fROYn.dIn*) is exhaustively matched against the lexicon, the syntactic string (*der Partner der Freundin*) becomes available for syntactic parsing. In addition, the lexical p-form information can be included in the p-diagram (e.g., information on lexical stress or prosodic word/clitic status).

4.4. The p-diagram

Figure 4 shows an automatically created p-diagram for the string *der Partner der Freundin* based on a speech signal with a dative construction. As discussed in Section 3, each vector includes the segments, the duration, and the mean F_0-value for the associated syllable. The p-diagram also includes the lexical p-form information by

pros_phrase	pp(σ	σ	(σ	σ))pp	pp(σσ	(σ	σ)	(σ	σ))pp
pitch_tones		L2		H4		L2	H1		
lex_stress	–	–	x	–	–	x	–	x	–
F0_mean	225.62	193.49	198.90	267.53	219.35	194.02	213.77	176.27	85.71
duration	0.17	0.16	0.33	0.18	0.14	0.30	0.20	0.28	0.22
segments	das	de:6	pa6t	n6	de:6	fROYn	dIn	fe:l	t@
Vector_index	1	2	3	4	5	6	7	8	9

Figure 4: P-diagram for a dative interpretation of the string *der Partner der Freundin* ('the partner of the friend')

indicating lexically stressed syllables with x and by adding the prosodic unit information to the attribute PROS_PHRASE (prosodic phrasing). While each function word (*dass, der*) is indicated by an underspecified syllable σ, the nouns' prosodic word status is indicated by the syllables within a set of unmarked brackets: (σ σ).

In a next step, the system then calculates different high and low tones (see Section 4.2) and prosodic phrase boundaries $_{pp}(\ )_{pp}$ which can be determined based on the acoustic cues reported in Section 2: F_0 movement, duration, and pauses. At this stage, the placement of prosodic phrase boundaries was only determined by F_0 movement; the estimation whether a particular syllable is longer or shorter than expected is work in progress and will be added shortly to the system.

Figure 4 shows that the system gives a fairly accurate categorical representation of the speech signal. Interesting points of debate are, e.g., the question whether the low tone L2 associated with vector 2 (GToBI: L*), which occurs just before the syllable boundary, should be 'moved' to vector 3 where the syllable carries lexical stress, or whether an additional attribute for 'early' or 'late' L/H tones would be more useful.

4.5. Disambiguation

The syntactic string determined in Section 4.3. is parsed with an LFG-CG for German and returns the expected overgeneration as seen in Figure 1. However, now that the information on prosodic phrase boundaries is available via the p-diagram the grammar is able to prosodically disambiguate the syntactically ambiguous structure. For space reasons, the interested reader is referred to [23] for details of the syntactic implementation.

5. EVALUATION

In order to evaluate the implementation, the recordings from [13] were used to create a 'gold standard'. In an online perception experiment, 32 native German speakers were asked to rate the recordings from the production study on a scale from one to five, where each endpoint corresponded to a particular meaning (1 represented dative and 5 genitive case). All recordings were randomized and assigned to different experimental lists. Participants were asked to listen to each assigned recording and to indicate which meaning they thought was associated with the signal. Each sentence was rated by at least two listeners. Only the sentences that were correctly rated twice (i.e., where the case of the produced sentence matched the case perceived by the listeners) were included in the gold standard and used for the evaluation.

The resulting 72 (59 dative, 13 genitive) recordings were annotated with SAMPA syllables and used as input for the complete implementation. The system was able to determine the correct syntactic structure (dative or genitive) in 55 cases (76,4%) based on the occurrence or absence of a prosodic phrase boundary after the first NP.

6. DISCUSSION AND CONCLUSION

This paper introduced a new end-to-end system, which takes a speech signal annotated with syllables as input, extracts the different acoustic cues, and calculates pitch accents and prosodic phrase boundaries based on this information. The prosodic structure in the resulting representation is used by a computational LFG-Grammar to disambiguate syntactically ambiguous structures. The implementation thus enables these formerly text-only processing grammars to process spoken language as well, and closes the gap between automatic speech recognition and linguistically deep computational grammars. As such, it takes a major step towards real automatic speech understanding.

An initial evaluation of the German system showed promising results which are expected to improve even further once additional factors (e.g., duration) are added to the interpretation. Implementations in two further languages (English and Urdu) with other crucial linguistic phenomena at the interfaces are underway and are expected to challenge and improve other aspects of the system.

7. ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with projects BO 3388/3-1 and BO 3388/4-1.

8. REFERENCES

[1] Y. Xu, "ProsodyPro - A tool for large-scale systematic prosody analysis." Laboratoire Parole et Langage, France, 2013.

[2] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech and Language*, vol. 45, pp. 326–347, 2017.

[3] F. Schiel, "Automatic phonetic transcription of non-prompted speech," in *Proceedings of ICPhS*, San Francisco, 1999, pp. 607–610.

[4] N. Braunschweiler, "Automatic detection of prosodic cues," Ph.D. dissertation, University of Konstanz, 2003.

[5] ——, "The prosodizer – automatic prosodic annotations of speech synthesis databases," in *Proceedings of Speech Prosody*, Dresden, Germany, 2006.

[6] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The prosody module," in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed. Berlin [a.o.]: Springer, 2000, pp. 106–121.

[7] A. Schweitzer and B. Möbius, "Experiments on automatic prosodic labeling," in *Proceedings of INTERSPEECH*, Brighton, UK, 2009.

[8] M. Butt, H. Dyvik, T. H. King, H. Masuichi, and C. Rohrer, "The parallel grammar project," in *Proceedings of the COLING-2002 Workshop on Grammar Engineering and Evaluation*, 2002.

[9] A. Copestake, *Implementing Typed Feature Structure Grammars*. Stanford: CSLI Publications, 2002.

[10] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 2956–2970, 1991.

[11] M. Żygis, J. M. T. Jr., C. Petrone, and D. Pfütze, "Acoustic cues of prosodic boundaries in German at different speech rate," in *Proceedings of the International Congress of Phonetic Sciences (ICPhS), Melbourne, Australia*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds., 2019, pp. 999–1003.

[12] A. Gollrad, E. Sommerfeld, and F. Kügler, "Prosodic cue weighting in disambiguation: case ambiguity in German," in *Proceedings of Speech Prosody*, Chicago, 2010.

[13] T. Bögel, "German case ambiguities at the interface: production and comprehension," in *Prosody in Syntactic Encoding*, ser. Linguistische Arbeiten, G. Kentner and J. Kremers, Eds. Berlin: De Gruyter, 2020, no. 573, pp. 51–84.

[14] M. Dalrymple, *Lexical Functional Grammar*. San Diego [a.o.]: Academic Press, 2001.

[15] J. Bresnan and R. M. Kaplan, "Lexical-Functional Grammar: a formal system for grammatical representation," in *The Mental Representation of Grammatical Relations*, J. Bresnan, Ed. Cambridge, MA: MIT Press, 1982, pp. 173–281.

[16] M. Butt, T. H. King, M.-E. Niño, and F. Segond, *A Grammar Writer's Cookbook*. Stanford, CA: CSLI, 1999.

[17] R. Crouch, M. Dalrymple, R. M. Kaplan, T. H. King, J. T. Maxwell III, and P. Newman, *XLE documentation*. Palo Alto, CA: Palo Alto Research Center, 2022, online documentation.

[18] T. Bögel, "The syntax–prosody interface in Lexical Functional Grammar," Ph.D. dissertation, University of Konstanz, Konstanz, 2015.

[19] J. C. Wells, "SAMPA computer readable phonetic alphabet," in *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, R. Moore, and R. Winski, Eds. Berlin, New York: Mounton de Gruyter, 1997, pp. 684–732.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program, Version 5.3.56]," 2013, available at http://www.praat.org/ [retrieved 15.09.2013].

[21] K. R. Beesley and L. Karttunen, *Finite State Morphology*. Stanford, CA: CSLI Publications, 2003.

[22] W. J. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production," *Behavioral and Brain Sciences*, vol. 22, pp. 1–75, 1999.

[23] T. Bögel, "The prosody-syntax interface: a computational implementation," in *Proceedings of the LFG22 conference*. CSLI Publications/University of Konstanz, 2022.

[24] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "The CELEX lexical database (CD-ROM)," in *Linguistic Data Consortium*. Philadelphia: University of Pennsylvania, 1995.

[1] See the XLE-Web Interface which features a number of different computational LFG grammars that can be used interactively: https://clarino.uib.no/iness/xle-web.

[2] In principle, this step can be automatized as well with the help of MAUS and CELEX [24], but as this paper is concerned with automatic speech understanding, the focus lies on the intersection between signal and grammar.