Interspeech 2018
2-6 September 2018, Hyderabad

An Empirical Analysis of the Correlation of Syntax and Prosody

Arne Köhn[1], *Timo Baumann*[2], *Oskar Dörfler*[1]

[1]Natural Language Systems Group, Department of Informatics, Universität Hamburg, Germany
[2]Language Technology Institute, Carnegie Mellon University, Pittsburgh, USA

{koehn,3doerfle}@informatik.uni-hamburg.de, tbaumann@cs.cmu.edu

Abstract

The relation of syntax and prosody (the syntax–prosody interface) has been an active area of research, mostly in linguistics and typically studied under controlled conditions. More recently, prosody has also been successfully used in the data-based training of syntax parsers. However, there is a gap between the controlled and detailed study of the individual effects between syntax and prosody and the large-scale application of prosody in syntactic parsing with only a shallow analysis of the respective influences. In this paper, we close the gap by investigating the significance of correlations of prosodic realization with specific syntactic functions using linear mixed effects models in a very large corpus of read-out German encyclopedic texts. Using this corpus, we are able to analyze prosodic structuring performed by a diverse set of speakers while they try to optimize factual content delivery. After normalization by speaker, we obtain significant effects, e. g. confirming that the subject function, as compared to the object function, has a positive effect on pitch and duration of a word, but a negative effect on loudness.

Index Terms: syntax, prosody, corpus-based analysis, linear mixed effects models

1. Introduction

Spoken language features two structural systems – syntax on the language side and prosody on the speech side – with a strong interrelation that has been a long-running topic of scientific research. Our contribution is to bring together large-scale corpus-based analysis using automatic tools with the breakdown of individual effects of the interplay to underpin linguistic insight.

Looking at pausing and constituency, it was first observed that most pauses appear at major syntactic constituent breaks [1], and that pause duration relates to the strength of breaks between constituents of a read sentence [2]. Using controlled experiments rather than observations, Price et al. [3] found that syntactic ambiguities can be recovered by listeners from the prosodic realization, and remark that this happens "primarily based on boundary phenomena, although prominences sometimes play a role" [3]. In a similar way, Beach [4] shows that prosody in a partial sentence can be used to (partially) predict sentence structure, in particular subject/object opposition, and using synthesized speech (thus limiting other influences). Weber et al. "conclude that in addition to manipulating attachment ambiguities, prosody can influence the interpretation of constituent order ambiguities" [5], hence pointing towards much broader influences, using eye-tracking in laboratory phonetics settings. Neurolinguists, likewise, have found evidence for the integrated processing of prosody and syntax in the human brain [6], in particular of intonation and sentence finality. Thus, we conclude that there is ample evidence from detailed analysis and controlled studies on the influences between prosody and syntax and the merit of prosody to differentiate syntactic ambiguities. In particular, prosody helps to differentiate constituent boundaries (in particular through pausing), attachment ambiguities, and syntactic function (potentially through features more closely related to prominence like duration, loudness and pitch).

The merit of prosody has indeed been used to improve syntax parsing, by including cues into PCFG parsers [7, 8] and very recently in neural networks-based dependency parsing [9]. While the addition of discrete prosodic symbols has been helpful in tasks like speech recognition and understanding [10], this was not helpful for parsing [7]. In contrast, a neural networks-based syntax parser can use a continuous representation of prosody and shows a merit on overall parsing accuracy for pausing, word duration, loudness and pitch [9].

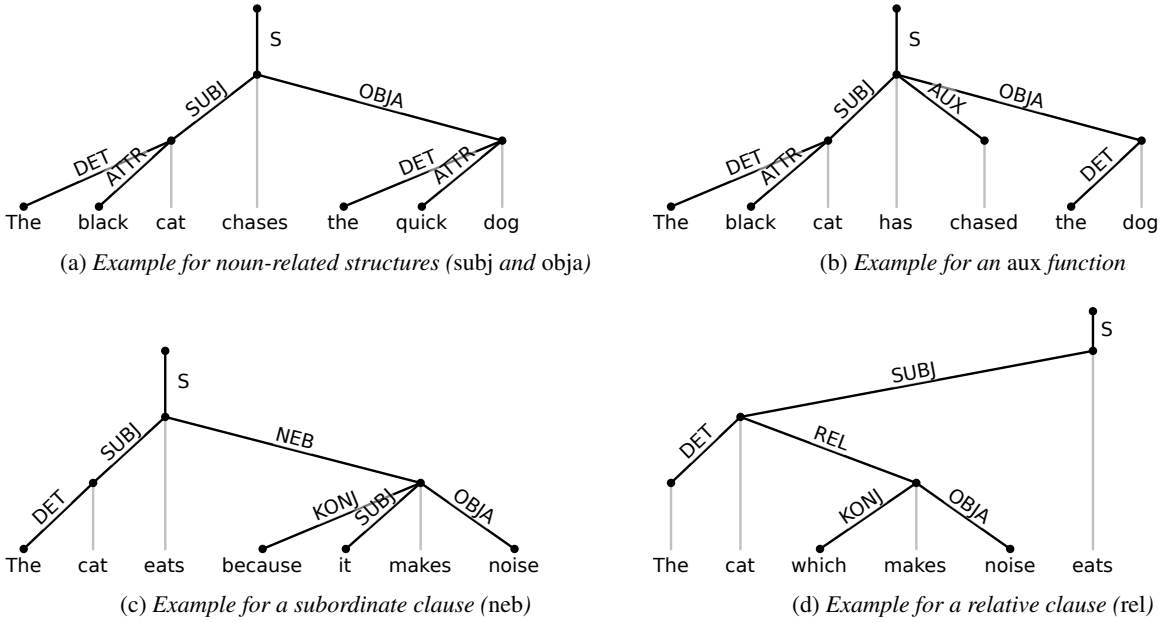
The work on acoustic parsing, while based on large corpora of syntactically annotated language has two limitations: (a) The resulting parser models cannot easily be analyzed in terms of which prosodic aspects help towards resolving what syntactic ambiguity. The models merely show that overall, prosody helps to improve parsing performance but do not explain concretely which syntactic structures correlate with which prosodic properties. (b) The work has only been performed on conversational language [11] and prosody was shown to be particularly helpful for parsing sentences with disfluencies. This leaves open the question of whether prosody would also help for parsing more canonical, read language which does not contain such disfluencies, or whether the positive influence of prosody is largely valuable to recover from disfluencies.

We analyze, in a large and diverse corpus of read German encyclopaedic material as collected from the Spoken Wikipedia[1] [12], the relation of syntactic dependency structure and prosodic realization using linear mixed effects models [13]. We analyze the correlation of prosodic features measured on a word and the syntactic function assigned to that word (or the word's head). We select the conditions to test based on linguistic intuition and find significant results for a number of oppositions (such as whether a noun is used as a subject or object). Similar in spirit to our work, prosodic features and their relation to broad syntactic features such as content vs. function words are analyzed in [14]. Although that work uses a small corpus of conversations, they similarly use automatic alignments and automatic annotations (in that case parts of speech).

The remainder of this paper is structured as follows: in Section 2, we explain in detail the method of using a linear mixed effects model for finding significant differences in prosodic characteristics of words with different syntactic functions. In Section 3, we describe the data that we use in the experiment which we describe and evaluate in Section 4. We discuss our results and conclude in Section 5.

[1] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Spoken_Wikipedia; also contains links to other languages.

10.21437/Interspeech.2018-2530



(a) *Example for noun-related structures (*subj *and* obja*)*

(b) *Example for an* aux *function*

(c) *Example for a subordinate clause (*neb*)*

(d) *Example for a relative clause (*rel*)*

Figure 1: *Dependency trees exemplifying the syntactic structures we use to predict prosodic features.*

2. Method and Statistical Model

We focus our analysis of the syntax–prosody interface on finding (significant) correlations between syntactic functions of words and prosodic features of those words. We ignore the question of causality in this paper (and given the neurolinguistic research cited above there probably is no clear causality but a coordinated interplay). Also, we focus on very simple features that can be automatically extracted: tempo, avg. pitch, avg. loudness and the duration of a following pause. These features can be measured easily and objectively and do not take into account any linguistic structure (unlike, e. g. [15]).

We find correlations by performing a likelihood ratio test between two linear models: (a) the basic model is fitted to predict the outcome, e. g. the mean pitch of the word using non-lexicalized textual features such as length of the word and position in the sentence. (b) The extended model performs the same task but uses the syntactic function as an additional feature. For each model (a) and (b) we compute how well it fits the data (i. e. compute the likelihood of the model). If the extended model fits the data better than the basic model, the syntactic function yields a benefit in predicting prosody even after accounting for the features contained in the basic model. This benefit can be expressed by dividing the likelihood of the basic model by the likelihood of the extended model and this likelihood ratio is used to compute *p*-values using the likelihood ratio test.

Different speakers perform differently, and recordings of the same speaker might differ.[2] To account for this, we take into account speaker and recording ID and each linear model is built using fixed effects (coefficients are shared between articles) and random effects (coefficients are optimized for each article to capture speaker characteristics), which yields a linear mixed effects model. We use the features themselves as predictors, as well as the interactions between features (i. e. feature combinations). The extended models only add one binary categorial feature each, namely to which of two syntactic groups an example belongs. To obtain a magnitude of change between the two groups, the coefficients from the basic model are used in the extended model and only the coefficient for the syntax information is fitted. This coefficient is the average difference between the two groups after controlling for the features of the basic model and is reported as effect in Cent/dB/ms in Table 1. Hypothesis testing is performed using the `lme4` R package [16]; for an in-depth discussion of hypothesis testing with linear mixed effects models see e. g. [13].

We perform two types of tests: The first type of test compares words that could fill the same syntactic roles using their function in the sentence as a predictor for prosodic features. For example, nouns and pronouns can act both as object (*obja*) and as subject (*subj*) in a sentence (see "cat" and "dog" in the example in Figure 1a). To shorten the notation, we write A~B to denote a comparison of words in context A to words in context B (in this case, subj~obja). The second type of test compares words that fill the same syntactic function but are attached to words that fill different functions. As an example, the word "the" occurs twice in the example in Figure 1a, once as a determiner (*det*) for the subject and once as a determiner for the object. We write A↗(B~C) to denote a comparison of words having dependency label A attached to words filling role B with words having dependency label A attached to words filling role C (in this case, det↗(subj~obja)).

Our comparisons only entail pairs of syntactic roles where words that can fill one syntactic role can also fill the other, e. g. the subject and object roles can both be filled by nouns, proper nouns, names, or pronouns. We could e. g. test for subj~det, but due to the two groups being disjoint, there would be no meaningful interpretation of the results and the fitted models might just pick up systematic differences between nouns and determiners, via their dependency labels.

Our approach could be seen as needlessly complicated; a straightforward (but wrong) way of testing the effect of syntactic structure on prosodic features would be to extract pairs of syntax and prosody features and determine the correlation between both. However, this approach would yield spurious significance as

[2] Some speakers recorded over a span of ten years.

important mediating factors would not be modeled. For example, a difference in pitch between subjects and objects might simply stem from the fact that objects tend to occur later in a sentence than objects and pitch tends to decrease over the sentence as subglottal pressure decreases [17]. We would not really measure the influence of syntax but merely the influence of the word's position, expressed through the correlation with syntax.

3. Data and Setup

We use recordings from the Spoken Wikipedia as a sample of read *speech in the wild* rather than laboratory-collected speech samples. The Spoken Wikipedia project unites volunteer readers who devote significant amounts of time and effort into producing read versions of Wikipedia articles read by a broad speaker population. It can thus be considered a valid source of speech produced by ambitious but not always perfect readers. The large number of readers make the corpus a valid sample of speech (although gender is imbalanced with only ~10 % female readers).

The data has been prepared as a corpus [12].[3] Importantly, the annotation includes a linguistic sentence segmentation and tokenization and the relation of original and normalized text has been preserved, allowing the timings of the aligned normalized text to be mapped to each original text token, thus bridging the gap between speech and language processing.

We limit our analysis to the German sub-corpus of the Spoken Wikipedia which contains some 1000 articles totaling 386 h of audio (360 h after VAD) and 3 M word tokens read by 350 different speakers [12]. The alignment favors quality over coverage and hence only about 70 % of the word tokens have alignment information available. For simplicity, we limit our analysis to those sentences in which every word has alignment information which yields 31,803 fully aligned sentences[4] with a total of 348,062 word tokens in 57,265 word forms.

We enrich the annotation with a dependency tree as well as Part-of-Speech tags for each sentence in our dataset using TurboParser [18] trained for German on the Hamburg Dependency Treebank [19]. The reported performance of the parser is >93 % labeled accuracy [19]. We parsed the complete German Spoken Wikipedia and these parses as well as all other code and data needed to reproduce our findings are freely available[3].

Our 32k fully aligned sentences span 46 hours of speech. We extract two types of data: textual features to predict prosodic features, and the prosodic ground truths to be predicted. Regarding textual features, we compute, for every word in our dataset, the canonical word duration using the duration predictor by Hal Daumé III[5], which is trained on MaryTTS [20] timing output for the most frequent German words. In addition, we extract the position of the word in its sentence as well as the sentence length. For the prosodic features, we use the SNACK library to extract power (in dB) and pitch (in semitones normalized relative to the speaker's mean) and record the mean value of each for each word. The duration is extracted from the alignment. Finally, for every word, we record whether it is followed by a pause according to the alignment information, and if so, the length of the pause (or zero otherwise).

[3] Available at: `http://islrn.org/resources/684-927-624-257-3/` and `http://nats.gitlab.io/swc`.

[4] The published version of the corpus, unlike the pre-release that we use here, contains about twice as many fully aligned sentences.

[5] See `https://nlpers.blogspot.com/2015/09/how-longll-it-take-to-say-that.html`.

4. Experiment and Analysis

Our experiments can be categorized into two parts: experiments regarding nouns and experiments regarding verbs. We specify each experiment using the notation introduced in Section 2. Our detailed results for all comparisons are shown in Table 1.

We first look at nouns and at words modifying nouns. Nouns frequently occur in subject as well as in object position. We compare subjects to accusative objects as these – in contrast to other objects in German – do not have case markers, thus making sure that the same word forms can occur in both functions.

subj~obja In Figure 1a, this would be "cat"~"dog" (keep in mind that we mitigating effects of nouns being more/less frequent in one of the positions by fitting the data to predicted speaking durations). We find that subjects are spoken with significantly higher pitch (a fifth of a semitone higher) and we also find that subjects are spoken significantly longer (by about 6 % compared to avg. noun durations in the corpus). We also find that the average signal power is slightly lower for subjects. We do not think this (very small) effect is due to softer speech. Instead we speculate that the slower speech causes this.

det↗(subj~obja) Differences in pronunciation between subjects and objects could very well carry over to the determiners attached to subjects and objects, respectively. In Figure 1a, this would be the determiners attached to "cat" and "dog". Again, we limit the object position to accusative. We find a smaller effect on duration with determiners being spoken more quickly before subjects. We speculate that the extra time spent on speaking subjects is subtracted (to some extent) from preceding determiners. Also, there is significantly more pausing after the determiner and, presumably, most often before the following noun.

attr↗(subj~obja) Similarly to determiners, attributes (i. e. adjectives) could be pronounced differently depending on whether they modify a subject or an object. In Figure 1a, this would be "black" and "quick". Again, we find a significantly higher pitch (and likewise power) for attributes of subjects (but no significant effect on duration). Interestingly, the effect on pitch is higher for attributes than for the noun itself.

Regarding verbs, we test words filling the main verb function (*s*, for *sentence*[6]), verbs that form the head of a subordinate clause (*neb*, for *Nebensatz*, Figure 1c), verbs that form the head of a relative clause (*rel*, Figure 1d), and verbs attached to auxiliary verbs (Figure 1b).

s~neb Of the non-main verb types, *neb* exhibits the least divergence from main verbs. This fact is not surprising since subordinate clauses are structurally the most similar to main sentences. However, we find that all of *neb*, *rel*, and *aux* are spoken substantially longer than main verbs.

s~rel In contrast to subordinate clauses, relative clauses modify nouns, e. g. in Figure 1d, "makes" modifies "cat". As relative clauses tend to be interjected into sentences, they need to be distinguishable from the main sentence (and it may be efficient to mark this prosodically). Verbs heading a relative clause are, on average, spoken with lower pitch, less power, and lengthened, i. e. less pronounced. This could be a strategy to distinguish the additional information from the main content of the sentence. *rel* and *aux* have a significant tendency for pausing to follow.

s~aux Auxiliary verbs express tense or other grammatical aspects. If an auxiliary verb is used, the main verb has an infinite form and the auxiliary verb takes over the role of the finite verb. In the annotation scheme we use, auxiliary verbs form the head

[6] The attachment label of verbs in dependency grammars directly encodes the equivalent of constituency in phrase structure grammars.

Table 1: *Experimental settings, their significance level and the coefficients of the added syntactic predictors in the extended model*

			det↗	attr↗				aux↗...		
		subj~obja	...↗(subj~obja)		s~neb	s~rel	s~aux	(s~neb)	(s~rel)	(s~aux)
pitch	p-value	***	NS	**	NS	***	NS	NS	***	***
	effect in Cent	19.38	—	25.14	—	39.38	—	—	22.81	53.77
power	p-value	***	0.06	*	*	***	NS	NS	NS	NS
	effect in dB	-0.01	0.005	0.01	0.01	0.02	—	—	—	—
duration	p-value	***	***	NS	***	***	***	0.12	***	*
	effect in ms	35.42	-14.62	—	-17.8	-23.03	-29.32	15.88	19.03	30.04
pause	p-value	*	**	NS	NS	*	***	NS	***	NS
	effect in ms	3.81	6.14	—	—	-5.52	-10.17	—	10.73	—

Significance levels: ***: $< .001$, **: $< .01$, *: $< .05$; clearly non-significant results: 'NS'. Effect is positive iff first > last, e. g. pitch higher for *subj* than *obj*.

of clauses and sentences and the main verb is attached to it (see Figure 1b). Therefore, in this setting we compare finite verbs with infinite ones that are combined with a finite auxiliary verb. We speculate that pauses are less frequent after a main verb which may also have an influence on the duration.

aux↗(s~neb) shows no significant results: we cannot show that infinite verbs that co-occur with auxiliary verbs differ depending on whether they are embedded into a main sentence or subordinate positions. Given that subordinate clauses are structurally most similar to main sentences, this appears plausible.

aux↗(s~rel) Verbs in conjunction with an auxiliary verb behave differently in main clauses compared to relative clauses. Similar to verbs without an auxiliary, the pitch is lower in a relative clause. The duration of verbs followed by a finite verb is also lower in relative clauses than in main clauses.

aux↗(s~aux) Verbs can form a chain in German, such as "gegessen haben müssen", *have to have eaten,* lit. *eaten have must*. If a verb is deeply nested like this, i. e. at least two verbs in the verb chain will still follow, it is spoken with significantly lower pitch than a word not nested as deeply.

5. Discussion

We have performed, to the best of our knowledge, the largest scale analysis of the correlation of syntax and prosody in terms of the amount of data (31,803 sentences in 46 hours of speech read by 300 speakers).

We have limited our correlation analysis to a few linguistically selected comparisons and often find significant differences between the contrasted conditions. Given the small number of comparisons we have performed, we have not corrected for multiple comparisons. Most results would stand even after (very conservative) Bonferroni correction, given very strong significances obtained by the large number of data points in our study.

In particular, we find strong effects of subject/object function on prosodic features, most prominently pitch and duration. We also find highly significant differences between the functions of verbs in different types of clauses (similarly to [21]). While overall the effect sizes may seem small, most are well above the *just noticeable differences* for pitch [22] and tempo [23]. Some smaller effects might not be consciously 'noticeable', but could still subconsciously help disambiguation.

Using a corpus of read speech allowed us to infer syntactic structure using an automated parser trained for parsing factual texts. The occasional errors of the parser, and hence the syntactic annotations that we use are unlikely to have yielded false positive results. In contrast, given that errors tend to add random noise, they should make our estimates more conservative. However, parsers are known to be relatively weak in resolving attachment ambiguity. This is why we did not investigate these in our study. In the future, we would like to investigate whether the prosody that relates to such ambiguities (e. g. for PP-attachments) naturally falls into clusters and whether these clusters overlap with 'ground truth' for the attachment. This would allow us to investigate in detail the prosodic means by which speakers communicate attachment information (and would help to improve parsers accordingly). Spontaneous speech, with its richer interactional behaviours, might yield different results (e. g. marking syntax might be less relevant as the expectations for 'correct' syntax might be lower). In any case, automatic syntactic parsing for spontaneous spoken language is much harder to come by.

Our experiments were performed on a single language. Using Universal Dependencies [24] as a syntactic annotation schema would enable to research similarities and differences in the relation between prosody and syntax across languages, using the different languages contained in the Spoken Wikipedia.

We have looked at very basic acoustic/auditory features in isolation, rather than looking at more higher-level prosodic features such as 'prominence' or 'phrasing' which themselves are constituted by a complex interplay of the basic auditory features. Our statistical analysis is hence conservative and points out only the most direct and most relevant correlations. Feature combinations instead of simple flat representations have lead to a break-through in parsing [9]. Hence, we expect many more and more complex interplays in the syntax–prosody interface to be found in future work building on more complex notions of prosodic and syntactic features.

There are many factors that influence prosody and that we have not modeled in our study, putting our results even further on the conservative side. One such aspect could be the information structure and modeling that would potentially yield even stronger results. On the other hand, we could also simply apply the same methodology as outlined here to investigate the influence of information structure (e. g. givenness) on prosodic properties. Using the Spoken Wikipedia for this research comes with the additional benefit that tools for extracting e. g. coreference are readily available for Wikipedia-style texts.

Acknowledgements: The authors would like to thank countless members of the Wikipedia community for providing the written and spoken material used as the basis of this study. We thank the anonymous reviewers for their helpful comments.

6. References

[1] F. Grosjean and A. Deschamps, "Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation," *Phonetica*, vol. 31, no. 3-4, pp. 144–184, 1975. [Online]. Available: https://doi.org/10.1159/000259667

[2] F. Grosjean, L. Grosjean, and H. Lane, "The patterns of silence: Performance structures in sentence production," *Cognitive Psychology*, vol. 11, no. 1, pp. 58–81, 1979. [Online]. Available: https://doi.org/10.1016/0010-0285(79)90004-5

[3] P. J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *The Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 2956–2970, 1991. [Online]. Available: https://doi.org/10.1121/1.401770

[4] C. M. Beach, "The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations," *Journal of Memory and Language*, vol. 30, no. 6, pp. 644–663, 1991. [Online]. Available: https://doi.org/10.1016/0749-596X(91)90030-N

[5] A. Weber, M. Grice, and M. W. Crocker, "The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements," *Cognition*, vol. 99, no. 2, pp. B63–B72, 2006. [Online]. Available: https://doi.org/10.1016/j.cognition.2005.07.001

[6] K. Eckstein and A. D. Friederici, "It's early: Event-related potential evidence for initial interaction of syntax and prosody in speech comprehension," *Journal of Cognitive Neuroscience*, vol. 18, no. 10, pp. 1696–1711, 2006. [Online]. Available: https://doi.org/10.1162/jocn.2006.18.10.1696

[7] M. Gregory, M. Johnson, and E. Charniak, "Sentence-internal prosody does not help parsing the way punctuation does," in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Boston, USA: Association for Computational Linguistics, May 2004, pp. 81–88. [Online]. Available: http://www.aclweb.org/anthology/N04-1011

[8] J. G. Kahn, M. Lease, E. Charniak, M. Johnson, and M. Ostendorf, "Effective use of prosody in parsing conversational speech," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, Canada: Association for Computational Linguistics, Oct. 2005, pp. 233–240. [Online]. Available: http://www.aclweb.org/anthology/H/H05/H05-1030

[9] T. Tran, S. Toshniwal, M. Bansal, K. Gimpel, K. Livescu, and M. Ostendorf, "Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, Jun. 2018, pp. 69–81. [Online]. Available: http://aclweb.org/anthology/N18-1007

[10] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech recognition and understanding," in *Mathematical Foundations of Speech and Language Processing*, M. Johnson, S. P. Khudanpur, M. Ostendorf, and R. Rosenfeld, Eds. Springer, 2004, pp. 105–114. [Online]. Available: https://doi.org/10.1007/978-1-4419-9017-4_5

[11] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The NXT-format Switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language resources and evaluation*, vol. 44, no. 4, pp. 387–419, 2010. [Online]. Available: https://doi.org/10.1007/s10579-010-9120-1

[12] T. Baumann, A. Köhn, and F. Hennig, "The Spoken Wikipedia corpus collection: Harvesting, alignment and an application to hyperlistening," *Language Resources and Evaluation*, 2018, special issue representing significant contributions of LREC 2016. [Online]. Available: http://dx.doi.org/10.1007/s10579-017-9410-y

[13] M. Brauer and J. J. Curtin, "Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items." *Psychological Methods*, 2017. [Online]. Available: https://doi.org/10.1037/met0000159

[14] M. Heldner and B. Megyesi, "Exploring the prosody-syntax interface in conversations," in *Proceedings ICPhS 2003*, Barcelona, Spain, Aug. 2003, pp. 2501–2504.

[15] A. Rosenberg, "AutoBI - a tool for automatic ToBI annotation," in *Proceedings of Interspeech*, Makuhari, Japan, Sep. 2010, pp. 146–149.

[16] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015. [Online]. Available: https://doi.org/10.18637/jss.v067.i01

[17] C. E. Gelfer, "A simultaneous physiological and acoustic study of fundamental frequency declination," Ph.D. dissertation, City University of New York, New York, USA, 1987.

[18] A. Martins, M. Almeida, and N. A. Smith, "Turning on the turbo: Fast third-order non-projective turbo parsers," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 617–622. [Online]. Available: http://www.aclweb.org/anthology/P13-2109

[19] K. A. Foth, A. Köhn, N. Beuck, and W. Menzel, "Because size does matter: The Hamburg Dependency Treebank," in *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*. Reykjavik, Iceland: European Language Resources Association, May 2014, pp. 2326–2333.

[20] M. Schröder and J. Trouvain, "The German text-to-speech synthesis system MARY: A tool for research, development and teaching." *International Journal of Speech Technology*, vol. 6, pp. 365–377, 2003. [Online]. Available: http://doi.org/10.1023/A:1025708916924

[21] M. Lelandais and G. Ferré, "Prosodic boundaries in subordinate syntactic constructions," in *Proceedings of Speech Prosody 2016*, Boston, USA, May 2016.

[22] S. G. Nooteboom, "The prosody of speech: Melody and rhythm," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle and J. Laver, Eds. Oxford: Blackwell, 1997, pp. 640–673.

[23] H. Quené, "On the just noticeable difference for tempo in speech," *Journal of Phonetics*, vol. 35, no. 3, pp. 353–362, 2007. [Online]. Available: https://doi.org/10.1121/1.4784699

[24] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman, "Universal Dependencies v1: A multilingual treebank collection," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. Portorož, Slovenia: European Language Resources Association, May 2016, pp. 1659–1666.