Speech Prosody 2016
31 May - 3 June 2016, Boston, USA

Prosodic and syntactic structures in spontaneous English speech

Anna Dannenberg[1], *Stefan Werner*[2], *Martti Vainio*[3]

[1]Department of Modern Languages, University of Helsinki, Finland
[2]School of Humanities, University of Eastern Finland, Finland
[3]Institute of Behavioural Sciences, University of Helsinki, Finland

anna.dannenberg@helsinki.fi, stefan.werner@uef.fi, martti.vainio@helsinki.fi

Abstract

In this paper we examine prosodic and syntactic structures of spontaneous English speech. By wavelet-based analysis, the prosodic structure of speech can be visually represented as a tree diagram. Combined with automatic syntactic parsing, this enables a novel method to compare prosodic and syntactic hierarchical structures in spoken language.

In our research we segmented a sample of spontaneous American English speech prosodically and syntactically and produced tree diagrams of both prosodic and syntactic structures, automatizing the process as completely as possible. The demarcation and internal structure of both kinds of segments were then analyzed in various respects.

The results indicate significant differences in prosodic and syntactic structures of spontaneous speech. The most notable divergence is in the predominant direction of branching: syntactic trees of spoken English language tend to be mainly right-branching, whereas in prosodic trees left- and right-branching structures alternate. The typical positions of prosodic and syntactic boundaries in spontaneous English speech also differ considerably from each other. It thus seems that prosodic and syntactic structures of spontaneous speech mostly follow different patterns in their appearance and probably also in their formation.

Index Terms: prosody, syntax, wavelet, phrasing, segmentation, structure, tree, hierarchy, spontaneous speech, English

1. Introduction

In linguistics, natural language is often called a unique product of human cognition. This is also the common starting point of linguistic research, concentrating on features such as memory constraints affecting syntactic structures, or the need for human interaction producing different pragmatic features of language. The result is usually some kind of a grammatical approach, be it a version of traditional phrase structure grammar or a more recent grammatical model such as cognitive or construction grammar.

On the other hand, especially spoken language is not only affected by human cognitive qualities but also by physiological ones. Therefore, to understand the structure of speech, one can not only rely on syntax; another important factor to be taken into account is prosody. This view has gained popularity in linguistics during the past decades, producing new syntactic models such as combinatory categorial grammar [1] that aims at unification of syntactic and intonational structures of language. Some researchers, like Croft [2], have even stated that spoken language should not be analyzed as clauses and sentences at all but rather as information units better reflecting the information processing of human mind.

In linguistic applications, however, prosodic approaches have been relatively rare. Most tools for automatic syntactic analysis, for instance, are built solely on the basis of traditional sentence-based syntactic theory. This hinders their use in analyzing spontaneous speech whose syntactic structure often deviates from the standard models of syntax. Naturally there have been attempts to utilize prosody in syntactic parsing, the pioneering work including e.g. the concept of implicit prosody by Fodor, where underlying prosodic patterns of language are used for syntactic disambiguation [3, 4]. The corresponding research about relations between prosody and syntax has vastly increased in popularity lately, definitely implying better linguistic applications in the future.

A typical problem in these kinds of approaches, though, is that prosodic patterns or features used in these studies have turned out to be hard to observe or measure directly, especially in unplanned spontaneous speech. Prosody is generally assumed to have some kind of a hierarchical structure, but there have been no means to directly visualize and estimate this hierarchy. Therefore these models have not yet been automatized effectively enough to produce consistent and reliable results for spontaneous speech.

Our solution to this shortcoming is Continuous Wavelet Transform (CWT) [5]. A CWT based tool can be used to visualize prosodic hierarchies of spoken language. By an appropriate choice of a scale space, CWT can be used to reveal the suprasegmental structures of arbitrarily long time frames, which makes it a viable tool for analyzing prosodic hierarchies of units even longer than an utterance.

Our CWT based method for analyzing prosodic structure of speech has been developed by Vainio et al. and described in detail in [6, 7]. It applies the weighted sum of f0, energy and segmental durations to represent prosodic signals of speech in a multidimensional time-frequency scale-space akin to spectrograms, so that the subsurface structures of the prosodic variables (time waveform, f0 contour, energy envelope) are rendered visible. The method can be further enhanced with lines of maximum amplitude, LoMA [8, 9], resulting in a visual representation of the prosodic hierarchies of speech. An example is shown in Fig. 1. We have previously applied the method for spontaneous Finnish speech in [10].

The prosodic structures that we are outlining here are primarily related to the stress structure and chunking of speech, not tune or melodic contour. The resulting hierarchical structure is thus prominence-based, prosodic boundaries being represented as negative prominence in the same scale-space.

In this research, we have used a CWT based tool to analyze the prosodic structure of spontaneous speech. The results have

10.21437/SpeechProsody.2016-13

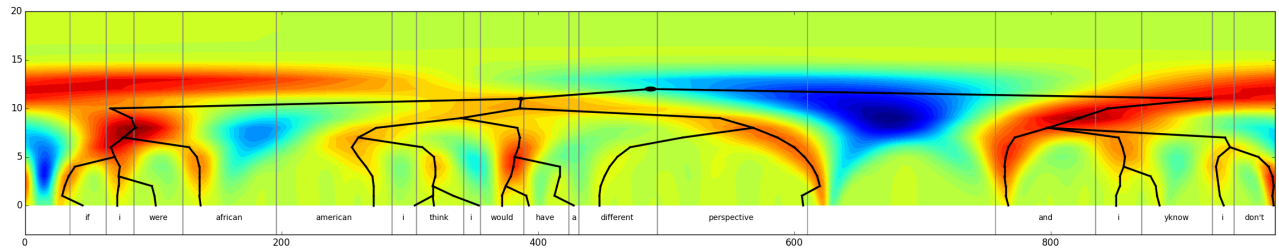

Figure 1: *Prosodic structure of a spontaneous speech utterance visualized by CWT and LoMA.*

then been compared with grammatical analysis of the same data to examine the relation of prosody and syntax in spoken language.

2. Data and methods

For this research, we applied data from the Buckeye corpus (Ohio State University) [11]. Our corpus was comprised of a total of 10076 words of spontaneous speech, and it included samples taken from informal interviews of five female speakers of American English, each sample consisting of ca. 2000 words.

In preprocessing, comments of the interviewer were deleted, and the speech was divided into turns based on speaker changes and obvious pauses. Each sample was divided into 30–50 speaker turns, their length varying from 2 to 350 words with the mean length of 51 words.

In our research, we hope some day to be able to utilize fully automatic methods for both prosodic and syntactic segmentation of the data. Thus far, however, it is not possible for the syntactic segmentation for two main reasons. First, since automatic parsers are primarily developed for standard written language, they are not well suited for handling spontaneous speech. Any discontinuities, repetitions or other nonstandard occurrences in speech thus tend to cause problems in automatic syntactic analysis. Second, there is a circular problem: identifying syntactic sentences or clauses typically presupposes some information of the relations between words, while identifying these relations requires presence of the syntactic boundaries. So far we have not found a way to perform both these tasks simultaneously, without previous knowledge of either of them. Therefore we had to apply manual and statistical methods for the syntactic segmentation.

2.1. Syntactic segmentation and parsing

The syntactic segmentation of the data was performed by 20 informants, all of whom were either students of or recently graduated from English language at the University of Eastern Finland. All the informants were native Finnish speakers, but were considered to have good to excellent skills in English.

In the assignment, the informants were asked, demonstrated by some examples, to tag every syntactic sentence and clause boundary in the text. The instructions were based on common grammatical parsing criteria, such as a clause typically consisting of a finite verb and other constituents closely attached to it. The instructions were not very detailed but rather guidelines, because in spontaneous speech there would be many exceptions to deal with in any case. In ambiguous cases, the informants were encouraged to make their own interpretations of the structure according to the context. The informants had no access to the spoken data but only the transcription, so they had to perform the segmentation task without help of any acoustic cues.

Each of our five samples was segmented by four informants independently of each other. Of these four segmentations, the one that differed most from the other three (i.e., had most boundaries that the other informants had not marked, or vice versa) was omitted. For the final segmentation, each boundary tagged by at least two of the three remaining informants was then accepted. This assured that each segment separated by syntactic boundaries in both ends was interpreted as a full independent syntactic segment by at least one informant, thus preventing inconsistent tagging in cases where a certain structure could have several completely different syntactic interpretations. The final number of phase boundaries was 1965, 70% of which were marked by all three informants, suggesting that the segmentations were relatively consistent with each other.

In the precursory analysis, it unfortunately appeared that "syntactic segmentation" was not perfectly understood or performed by all informants. Of the boundaries tagged by the informants, 60 (3,1%) were not syntactic sentence or clause boundaries at all but rather separating e.g. non-clausal phrases, and the number of dubious cases related to filler particles, hesitations etc. was even higher. Perhaps even worse, a lot of indisputable clause boundaries had been left unmarked (or had been tagged by only one informant and therefore left out of the final segmentation). We decided to leave the excessive boundaries intact, but 182 missing indisputable syntactic boundaries were added to the segmentation afterwards, increasing the amount of boundaries by ca. 9%. For the uncertainty reasons, we also decided to combine sentence and clause boundaries, thus only leaving one level of syntactic units to examine.

The grammatical analysis was performed using the Stanford statistical parser for English language. Syntactic segments were analyzed by the parser, producing their grammatical structure, and then transformed into tree form to facilitate comparison with prosodic structures. The Stanford parser is considered to perform reasonably well with natural language, but since it is created and mostly trained using written language, it turned out to have notable difficulties in handling spontaneous speech data, producing parses that clearly deviated from the actual grammatical structure of the segments. The data also had to be manually re-preprocessed before parsing (e.g. add punctuation and remove repeated words to avoid false interpretations). Therefore only a small sample of the data could be parsed, despite the method itself being automatic.

2.2. Prosodic segmentation and structural analysis

For our research, the data was prosodically segmented using a wavelet based tool. The whole process was fully automatic and unsupervised.

The weighted sum of normalized f0, energy and segmental

durations was used as an input signal for the CWT. Prosodic boundaries were determined by tracking minima across scales in the resulting scalograms, lines of minimum amplitude [8]. In the analysis, four levels of prosodic boundaries were specified, but since we decided to use only one level of syntactic boundaries in this research, all the prosodic boundaries were also combined into one level by assigning them an equal weight.

The prosodic structural analysis was performed by CWT enhanced with unpruned lines of maximum amplitude, producing visual tree images of the prosodic hierarchies.

3. Results

The result of the syntactic segmentation was a total of 2147 syntactic boundaries, 1965 tagged by the informants and 182 added afterwards. The number of boundaries per sample ranged from 368 to 500 boundaries. The mean length of a syntactic unit for the whole data was thus 4.7 words, the longest one consisting of 19 words.

The CWT analysis resulted in 1700 prosodic boundaries for the whole data, the mean length of a prosodic unit being 5.9 words. Most of the prosodic units consisted of less than 10 words, but there were a couple of units exceeding 30 words, the longest one having the total of 44 words.

3.1. Prosodic and syntactic boundaries

Of all the boundaries marked in the data, 906 were co-occurrencies of prosodic and syntactic ones. This is 53.3% of all the prosodic boundaries and 42.2% of the syntactic boundaries. The result was remarkably less than expected: in our earlier Finnish data [10], the percentage of shared boundaries of the syntactic ones was as high as 70%. A slightly different method had been used for the prosodic segmentation of the Finnish data, though, so there were now much fewer prosodic segments in the data with a notably longer mean length. In addition, the Finnish data had been divided into noticeably shorter speaker turns, which probably too strongly affected the syntactic segmentation. Despite this, the English data had significantly less shared boundaries than we had anticipated. The distribution of shared boundaries in our data was, nevertheless, clearly not random (in χ^2 test, $p < 0.001$).

Of all the separate segments, only a very small minority was limited by a combination of prosodic and syntactic boundary in both ends. These segments included a lot of exclamations, fillers and formulaic expressions (*yes, no, y(ou) know*), whereas information-carrying longer clauses and sentences tended to have internal prosodic and syntactic breaks in different positions.

The most common locations for combined prosodic and syntactic boundaries in our data included changes of topic or other noticeable hesitations in the speech. They were typically represented by some kind of syntactic discontinuity accompanied by a prosodic break. For example, most of the speaker turn changes included both a prosodic and a syntactic boundary; not all of them, though, since in some occasions the syntactic structure could be interpreted to continue uninterrupted in the following turn. However, the speaker turns did not have a major effect on the overall results: if the syntactic boundaries located at a speaker turn change were completely omitted, the percentage of shared boundaries would still be 38.9% (as compared with the total of 42.2%).

For the prosodic boundaries without a syntactic one, a couple of typical contexts were found. One of them was after a conjunction, whereas a syntactic boundary was situated before the conjunction. The same result had been perceived in our earlier Finnish data, supporting the assumption that in spoken language, the role of conjunctions somehow differs from that of standard written language. In spontaneous speech, a conjunction at the end of an utterance or speaker turn often shows an intention to continue despite the prosodic break. The choice of a turn-final conjunction can also tell about the speaker's attitude towards the topic, even if the structure is left open both grammatically and prosodically. Another common context for a solitary prosodic boundary was occasions where the same word or a group of words was repeated two or more times. In these cases, hesitating often causes a prosodic break in the flow of speech either in the middle of the repetition or directly after it, but grammatically these repeated words are interpreted as belonging to the same syntactic unit.

Concerning syntactic boundaries and their co-occurrence with the prosodic ones, one surprising detail showed up. Of the syntactic boundaries tagged by the student informants, almost 45% were accompanied by prosodic ones, but the same only pertained to 13% of the syntactic boundaries added to the data afterwards. In other words, almost nine out of ten clause or sentence boundaries that the informants had failed to identify were also prosodically unemphasized by the speaker. There would be nothing unusual in this observation, was it not for the fact that the informants had no access to the spoken data but merely had to rely on textual cues in their segmentation task. It seems that the informants, while reading the transcription to find syntactic boundaries, somehow more or less unconsciously were able to interpret it as speech, thus paying more attention to the sections where they could "hear" a prosodic break. This observation reminds of the concept of implicit prosody [3, 4], where some kind of prosodic structures are noticed to regulate even the interpretation of written language. We cannot examine this observation here in detail, but it is definitely worth more elaborate research in future, opening an interesting view on the human perception of transcribed spontaneous speech.

3.2. Internal structure of prosodic and syntactic segments

After tagging and analyzing the prosodic and syntactic boundaries in the spoken data, the internal structures of prosodic and syntactic segments in speech were further examined and compared by automatic grammatical and prosodic methods.

Of the resulting prosodic and syntactic tree structures (Figures 2 and 3), even at first glance it could be seen that they did not have much in common. Practically for any segment of three words or more, the syntactic and prosodic trees differed remarkably from each other. The most important difference was in the very basic structure of the trees: syntactic clause structures of spoken English tend to be almost exclusively right-branching, whereas in prosodic trees right- and left-branching structures alternate, the latter ones perhaps being even slightly more common in our data.

In syntax of spoken language, the predominance of right-branching structures mainly results from the limitations of human working memory. In left-branching syntactic structures, the initial items have to be stored in working memory of both speaker and hearer until the end of the structure. This makes them much harder to process, causing the speakers to favor right-branching structures instead (e.g. [12]). As for prosodic structures in our data, on the contrary, left-branching prosodic structures tended to dominate especially in the beginning of utterances. This may be related to the speech rate: in the begin-

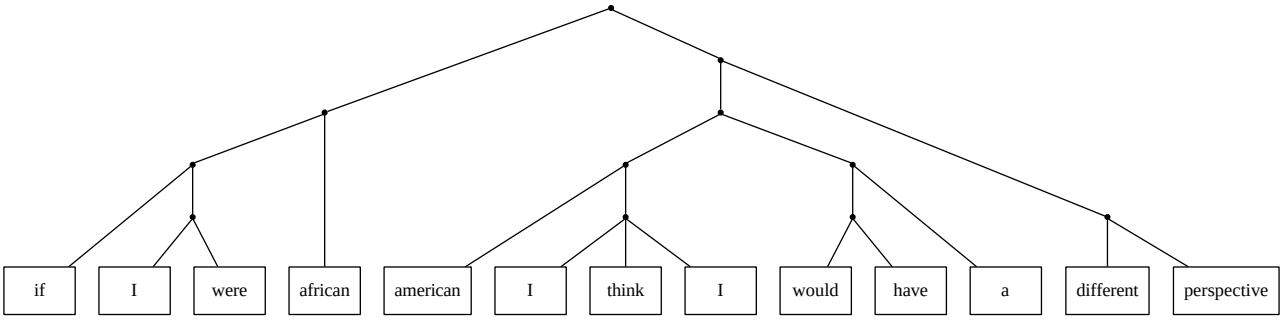

Figure 2: *Simplified prosodic tree structure.*

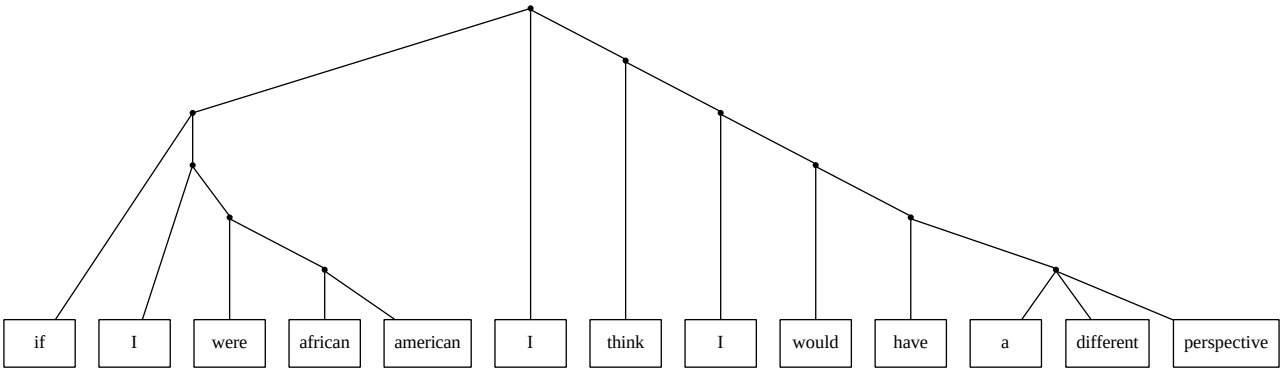

Figure 3: *Simplified syntactic tree structure.*

ning of a speech sequence, the words tend to combine prosodically, producing left-branching tree structures where the first words may appear almost amalgamated. Towards the end of the sequence, the pace typically slows down due to the final lengthening (e.g. [13]), resulting in prosodically separate words that in the tree structure appear as a right-branching composition.

The CWT+LoMA method, however, while providing good results on read English speech [8], has not been formally evaluated on spontaneous speech. Therefore it is so far probably wise to refrain from any far-reaching conclusions about the prosodic structures based solely on CWT.

4. Discussion

In this research, we used a wavelet-based method to reveal the prosodic structures of spontaneous speech. The resulting prosodic boundaries and trees were then compared with corresponding syntactic structures.

As a conclusion, our spontaneous English speech data turned out to have less shared prosodic and syntactic boundaries than expected, but still remarkably more than would be achieved randomly. This result proposes some kind of a connection between syntactic and prosodic structure of spoken language, especially considering the fact that our informants were more apt to identify those syntactic boundaries accompanied by prosodic ones. The internal hierarchical structures of prosodic and syntactic units, nevertheless, had practically no resemblance to each other.

The dissimilarity between prosodic and syntactic structures can be interpreted to suggest that prosodic and syntactic structures of spoken language are affected by different aspects of speech production. In spoken language syntax, cognitive constraints such as working memory produce certain kinds of hierarchies, while many phenomena behind prosodic structures, such as speech rate, are also strongly affected by physiological constraints.

On the other hand, the most severe critique towards our methods and results can be conducted towards our choice of syntactic model. The syntactic analysis in our research mainly follows the traditional phrase structure grammar. It was initially chosen purely to strive for automatic syntactic analysis, for which no usable tools are available based on different syntactic models. Since our results show such remarkable discrepancies between syntactic and prosodic structures, though, it can be taken as a sign that this kind of a syntactic approach might not truly reflect the structure of spoken language. Therefore traditional sentence-based syntax may not be a good method for analyzing spontaneous speech at all.

In our future work we thus have to pay more attention to the choice of syntactic methods, probably heading for more usage-based models such as dependency or construction grammars. The wavelet-based prosodic analysis method must also be formally evaluated on spontaneous speech. So far, however, it seems that syntactic analysis of spontaneous speech can truly benefit from prosodic methods.

5. Acknowledgements

The authors would like to thank Antti Suni for his valuable contribution in development and use of the CWT based analysis method.

6. References

[1] M. Steedman, *The Syntactic Process*, Cambridge (Mass.): MIT Press, 2000.

[2] W. Croft, "Intonation units and grammatical structure," *Linguistics*, vol. 33, no. 5, pp. 839–882, 1995.

[3] J. D. Fodor, "Learning to parse?," *Journal of Psycholinguistic Research*, vol. 27, no. 2, pp. 285–319, 1998.

[4] J. D. Fodor, "Prosodic disambiguation in silent reading", *Proceedings of the North East Linguistic Society 32*, M. Hirotani (ed.), Amherst: GSLA, University of Massachusetts, pp. 112–132, 2002.

[5] S. Mallat, *A wavelet tour of signal processing*, Academic Press, 1999.

[6] A. Suni, D. Aalto, T. Raitio, P. Alku and M. Vainio, "Wavelets for intonation modeling in HMM speech synthesis," *Proceedings of the 8th ISCA Speech Synthesis Workshop (SSW8)*, Barcelona, Spain, pp. 285–290, 2013.

[7] M. Vainio, A. Suni and D. Aalto, "Continuous wavelet transform for analysis of speech prosody," *Proceedings of the TRASP 2013 (Tools and Resources for the Analysis of Speech Prosody)*, Aix-en-Provence, France, pp. 78–81, 2013.

[8] A. Suni, D. Aalto and M. Vainio, "Hierarchical representation of prosody for statistical speech synthesis," *Computer Speech and Language* (submitted), arXiv preprint arXiv:1510.01949, 2015.

[9] M. Vainio, A. Suni and D. Aalto, "Emphasis, word prominence, and continuous wavelet transform in the control of HMM-based synthesis," *Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, Berlin Heidelberg: Springer, pp. 173–188, 2015.

[10] A. Dannenberg, M. Vainio, A. Suni and S. Werner, "Prosodic and syntactic segmentation of spontaneous speech: a preliminary study," *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK, paper number 978, 2015.

[11] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume and E. Fosler-Lussier, *Buckeye Corpus of Conversational Speech* (2nd release), www.buckeyecorpus.osu.edu, Columbus, OH: Department of Psychology, Ohio State University (Distributor), 2007.

[12] M. L. Hummert, "Maintaining competence in the face of resource limitations: the role of schema complexity in aging and communication," *Generative Mental Processes and Cognitive Resources*, U. von Hecker, S. Dutke and G. Sedek (eds.), Dordrecht: Kluwer Academic Publishers, pp. 157–174, 2000.

[13] P. Wagner, *The rhythm of language and speech: constraints, models, metrics and applications*, online publication, http://pub.uni-bielefeld.de/publication/1916845, 2008.