PROSODIC AND SYNTACTIC SEGMENTATION OF SPONTANEOUS SPEECH: A PRELIMINARY STUDY

Anna Dannenberg[a], Antti Suni[b,c], Martti Vainio[b], Stefan Werner[d]

[a] Department of Modern Languages, University of Helsinki
[b] Institute of Behavioural Sciences, University of Helsinki
[c] Department of Signal Processing and Acoustics, Aalto University
[d] School of Humanities, University of Eastern Finland
anna.dannenberg@helsinki.fi

ABSTRACT

In this paper we examine prosodic and syntactic segmentation of spoken Finnish language. Syntactic sentence or clause is generally mentioned as one of the basic units of language, but it can be questioned whether it is a good unit for analysing the structure of spontaneous speech.

By wavelet-based analysis, the prosodic structure of speech can be represented as a tree diagram, making it possible to compare prosodic and syntactic hierarchical structures of spoken language. As a first step, we compare here syntactically and prosodically defined speech segments and their boundaries. Our preliminary results show many similarities but also discrepancies between the prosodic and syntactic segments of spontaneous Finnish speech. These results serve as a good starting point for further comparison of syntactic and prosodic structures of spoken language.

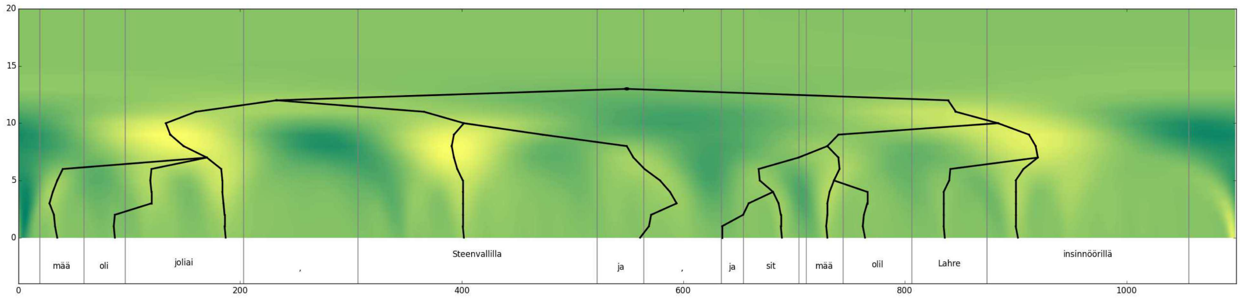
Keywords: prosody, syntax, wavelet, phrasing, Finnish

1. CONCEPT OF ”SENTENCE” IN LINGUISTIC THEORIES

Sentence is generally recognized as one of the most basic units of language. As e.g. Halliday [2] puts it: ”Sentence and word are the two grammatical units that are recognized in our folk linguistics; and this incorporates a piece of good common sense.” Sentence is widely used as the basic unit of language analysis, and much of both ancient and contemporary grammar is based on it.

There is one severe problem in the concept of ”sentence”, though: in most grammatical theories it is defined so that it is not well suited for analysing spoken language, especially spontaneous speech. In common use, ”sentence” is often understood as an orthographic unit, a sequence of written text extending from a capital letter to a full stop. Naturally, this definition does not fit to spoken language at all, since in speech there are neither capital letters nor full stops. Another common way of defining sentence in linguistics is through grammar, as in e.g. [1], [3], [5]. The typical minimum requirement for a grammatical sentence is subject and predicate together with their arguments, which leaves out a great deal of spoken language since every independent utterance of speech does necessarily not have these parts of construction. A third – and amazingly common – alternative is that ”sentence” is not defined in the theory at all but is nevertheless used as a central concept of it, as if being based on some kind of general knowledge about the structure of language. Perhaps a little surprisingly, this is the case with e.g. Halliday [2]. Usually this means a traditional grammatical approach, which may work well with most of written language but leads to obvious problems with spontaneous speech.

Figure 1: An example of a prosodic tree diagram derived from CWT analysis, and the corresponding prosodic and syntactic segmentation.



```
[S] [P] mää oli         joliai , [P] Steenvallilla [S] ja , [P] ja  sit  mää
        I   be+IMPF+1SG some+ADESS   Steenvall+ADESS   and      and then I

olil [P]    Lahre     insinnöörillä .
be+IMPF+1SG Lahti+GEN engineer+ADESS
```

”I was [working as a servant] at a person called Steenvall, and then I was at an engineer named Lahti.”

[S] = syntactic clause boundary, [P] = CWT-indicated prosodic boundary

Problems thus arise when one tries to analyse the structure of spontaneous speech with methods developed for standard (written) language. If the method is based on a basic unit called ”sentence”, what can one do if there are no such units in the material? One of the main challenges is automatical parsing of spoken language, since most of the parsing tools are developed with clearly identifiable syntactic sentences as a starting point. Therefore, alternatives for the concept of sentence are needed for successful analysis of structures of spoken language.

2. CONTINUOUS WAVELET TRANSFORM

There is very little doubt that prosodic structure is hierarchical, but there have been relatively few attempts to directly visualize this hierarchy. The continuous wavelet transform (CWT) [4] offers a way to represent prosodic signals (f0, energy, etc.) in a multidimensional time-frequency scale-space akin to spectrograms. The advantages in transforming prosodic signals with wavelets are similar to viewing the ordinary oscillogram as a spectrogram: the structures that are not visible in the surface (time waveform, f0 contour, energy envelope) are rendered visible. In the case of spectrogram, phoneticians are typically interested in the formant structure and other relatively short term – segmental – features of speech. In prosody, the interest is typically directed towards longer time scales that vary from segments to whole utterances and speaking turns (or paragraphs if we are dealing with read speech). Due to constant time frequency resolution, a spectrogram is a relatively poor instrument for studying the suprasegmental structure of signals. The scale space of a CWT analysis, on the other hand, can be chosen so that it can reveal the suprasegmental structures to arbitrarily long time frames that can cover units even longer than an utterance.

Suni et al. have recently developed a CWT based method that they have used to estimate word prominences and to control prosody in text-to-speech synthesis [7, 9]. In order to interpret the inherent hierarchy in speech they have further applied a simple algorithm that produces discrete tree structures from the prosodic signals [8, 10]. Here we use those trees, based on so called lines of maximum amplitude (LoMA) to study prosodic phrasing in Finnish. Figure 1 shows an utterance analysed with CWT and LoMA. The analysis is based on a compound signal that has been produced from a continuous (interpolated) fundamental frequency, energy envelope, as well as durations signal that has been calculated using differences from mean durations on the level of syllables.

3. CORPUS AND PROCESSING

Our small experimental corpus consists of spontaneous speech from dialect interviews conducted by Institute for the Languages of Finland [6]. It contains 395 utterances from two speakers, both native speakers of different dialects of Finnish.

The utterances have been analysed with CWT and LoMA, producing a tree diagram for every utterance. The tree diagrams have then been used for dividing the utterances into prosodic segments. The segmentation has been conducted manually, based on how the words have been grouped into branches in the tree structure.

For comparison, the same utterances have been segmented according to their syntactic properties. The segmentation follows the traditional grammatical parsing of Finnish language. The segments include full independent and coordinate main clauses, full subordinate clauses and corresponding clause fragments. The syntactic segmentation and classification has been conducted manually by a skilled native speaker of Finnish.

Because one of our aims is to examine the suitability of CWT analysis for this kind of purposes, the prosodic segmentation has been executed purely along the visual tree diagrams, without any reference to the auditive data after the CWT analysis has been performed. In the syntactic segmentation, auditive data has been used for disambiguation of some grammatically ambiguous cases.

4. RESULTS OF PROSODIC AND SYNTACTIC SEGMENTATION

Our experimental corpus of 395 utterances has a total of ca. 4950 words (depending on what is counted as a ”word” in e.g. hesitations or self-corrections). It has 933 syntactic boundaries (denoted by [S] in our data) and 1506 prosodic boundaries (denoted by [P]), resulting in the mean length of ca. 3.3 words for a prosodic segment and 5.3 words for a syntactic segment. The mean length of prosodic segments is somehow arbitrary, since it depends of the specificity in which small branches of a tree diagram have been interpreted as belonging to same or different segments, but it can still be stated that prosodic segments are shorter in average and also show less variation in length than syntactic clauses.

In our corpus, there are 653 co-occurrences of both prosodic and syntactic segment boundary, which is 70 % of all the syntactic boundaries and 43 % of all the prosodic boundaries. The deviation of this observation from a random sample is statistically highly significant ($p < 0.001$). It is thus evident that especially for syntactic boundaries, the most typical environment is together with a prosodic one. Therefore it is reasonable to concentrate on the occasions where the segment boundaries do not co-occur.

4.1. Solitary syntactic boundaries

Of the 280 syntactic boundaries in the corpus that occur without a prosodic boundary, the most typical instance is that after the syntactic boundary there is only a single conjunction before the next prosodic boundary (Example 1, Figure 1). These occasions form 40 % of all the solitary syntactic boundaries in the corpus, coordinate conjunctions being a little more common in this context (58 instances) than subordinate ones (55).

The result is expectable, as it illustrates the common feature of the spoken language that a conjunction is prosodically grouped to the previous prosodic entity, although grammatically it is more closely connected with the following clause. These lone conjunctions may also occur sentence-finally, thus telling about the speaker's purpose to continue even though it would not happen immediately.

(1) **[P]** **[S]** minä aatteli **[S]** että **[P]** se ajaa jänestä **[S]** mutta **[P]** se ajoki , **[P]** mettäsikaa **[S]** ja .
[P] **[S]** I thought **[S]** that **[P]** he was chasing a hare **[S]** but **[P]** he was chasing , **[P]** a wild boar **[S]** and .

Rest of the syntactic boundaries occurring without a wavelet boundary can not be grouped as clearly as the solitary conjunction cases. The only repeated pattern is a combination of a syntactic boundary and a lone particle *ni* (”so”, ”well”), a colloquial particle often used to begin a main clause after a subordinate one (Examples 2 and 3). This is a reverse phenomenon to the single conjunctions, a sign that the turn will be continued with a main clause (as opposed to a subordinate or coordinate clause).

(2) [P] [S] kum minä tulin siältä [S] ni , [P] siin oli , [P] kettu , kualluk keskel [P] tiätä .
[P] [S] when I came from there [S] so , [P] there was , [P] a fox , dead in the middle [P] of the road .

(3) - - [S] enkä [P] o enää kulkennu [S] ni , [P] em minä tiärä .
- - [S] and I haven't [P] been wandering anymore [S] so , [P] I don't know .

4.2. Solitary prosodic boundaries

Since CWT-based prosodic segments in our corpus are shorter in average than syntactic ones, grammatical clauses are more often broken by prosodic boundaries than vice versa. Usually prosodic boundaries do not break smaller syntactic phrases, though, but rather are situated between them. For example, prosodic boundaries seem to occur in the middle of a noun phrase or a preposition/postposition phrase significantly less often (p = 0,025) than in a random sample.

Some typical places for solitary prosodic boundaries include between a verb and its arguments (Examples 1, 2 and 4, Figure 1), or between an auxiliary and a main verb (Example 3). Some relatively independent phrases, such as adverbial phrases, frequently form separate prosodic entities inside a syntactic clause (Example 5). Self-corrections or hesitations also often result in a prosodic break (Figure 1), as well as discourse particles (Example 4).

(4) [P] [S] ja tua noi , [P] mää palveli [P] sil sill insinööril [P] sit vähänn aikaa [S] ja .
[P] [S] and well , [P] I worked as a servant [P] for that that engineer [P] then for a while [S] and .

(5) [P] [S] mu niit oli loukkui paljo mettäs [P] semmottis karjapolvuil .
[P] [S] but there were many traps in the forest [P] on those cattle paths .

The above mentioned cases of a conjunction in the end of a prosodic segment naturally lead to solitary prosodic boundaries as well; see Example 1 and Figure 1.

5. DISCUSSION

In this preliminary research, we have not yet gained deeper insight into the internal structure of prosodic segments acquired by CWT. This will be the next step of our study, once the method has been fully assessed and evaluated.

Thus far it seems nonetheless that prosodic and syntactic segments of spoken language resemble each other in some respect. Syntactic sentence boundaries typically co-occur with prosodic ones, and though prosodic boundaries also often break syntactic clauses, they are not situated arbitrarily but usually between smaller syntactic phrases. Similarly, if there is a syntactic boundary in the middle of a prosodic segment, it often tells about some kind of discrepancy between syntactic and discourse structure.

Syntax and prosody thus complement each other in the structure of speech, which reinforces our assumption that prosodic methods are needed for improving the results of spoken language parsing. The wavelet-based method seems a good candidate, since it offers unforeseen possibilities to examine in detail the prosodic structure of spoken language. Visualising speech prosody with a scale-space presentation is a viable tool for a linguistic research on prosody as it produces an objective rendition of the given signal. This helps to understand interaction between physiology and cognition in the development and use of human language.

6. REFERENCES

[1] Chomsky, N. 1969 [1957]. *Syntactic structures.* 8th printing. The Hague/Paris: Mouton.

[2] Halliday, M. A. K. 1985. *An introduction to functional grammar.* London: Edward Arnold.

[3] Langacker, R. W. 1991: *Concept, image and symbol: the cognitive basis of grammar.* Berlin: Mouton de Gruyter.

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

[5] Sag, I. A., Wasow, T. 1999: *Syntactic theory: a formal introduction.* Stanford: CSLI Publications.

[6] Samples of spoken Finnish: an online speech corpus (2014). Version 1.0. Helsinki: Institute for the Languages of Finland. Retrieved in December 2014, http://urn.fi/urn:nbn:fi:lb-1001100134.

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

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

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

[10] Vainio, M., Suni, A., Aalto, D. (in press). Emphasis, word prominence, and continuous wavelet transform in the control of HMM-based synthesis. In: *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis.* Springer.