Exploiting Prosody for Automatic Syntactic Phrase Boundary Detection in Speech

György Szaszák[1] *and András Beke*[2]
[1] Department of Telecommunication and Media Informatics,
Budapest University for Technology and Economics, Budapest, Hungary
[2] Research Institute for Linguistics,
Hungarian Academy of Sciences, Budapest, Hungary

ABSTRACT

The relation between syntax and prosody is evident, even if the prosodic structure cannot be directly mapped to the syntactic one and vice versa. Syntax-to-prosody mapping is widely used in text-to-speech applications, but prosody-to-syntax mapping is mostly missing from automatic speech recognition/understanding systems. This paper presents an experiment towards filling this gap and evaluating whether a HMM-based automatic prosodic segmentation tool can be used to support the reconstruction of the syntactic structure directly from speech. Results show that up to 85% of syntactic clause boundaries and up to about 70% of embedded syntactic phrase boundaries could be identified based on the detection of phonological phrases. Recall rates do not depend further on syntactic layering, in other words, whether the phrase is multiply embedded or not. Clause boundaries can be well assigned to intonational phrase level in read speech and can be well separated from lower level syntactic phrases based on the type of the aligned phonological phrase(s). These findings can be exploited in speech understanding systems, allowing for the recovery of the skeleton of the syntactic structure, based purely on the speech signal.

Keywords: *prosody, syntax, phonological phrase, boundary detection*

1 INTRODUCTION

A number of applications in automatic speech understanding require some analysis of the content prior or parallel to speech-to-text conversion referred to often as automatic speech recognition. In speech understanding, a pure transcription of the speech yielded by a speech-to-text converter (speech recognizer) would be insufficient, as the underlying meaning remains unextracted, uninterpreted. Of course, text-based analysers can be used for the speech-to-text output to assess meaning, however, this output can be unreliable depending on the difficulty of the speech recognition task, closely linked to several factors like environmental ones (noises in the speech signal, distortions), or speaking style (the "spontaneity" of speech) or in general the complexity of the recognition task (vocabulary, language model perplexity), etc. For some languages – like Hungarian, which is in the center of interest of the present study – both speech recognition (Szarvas et al., 2000) and text-based syntactical analysis (Babarczy et al., 2005) are difficult and work with significantly weaker performance compared to the English baselines due to the very rich morphology of the language. If recognition performance is poor, only highly unreliable data could be fed into a text-based syntactic or semantic analyser to assess the meaning. On the other side, if speech recognition works with good accuracy for a given task, the interpretation of the meaning can be still supported or constrained by other methods than text-based analysis, in order to add redundancy and create a more robust and more powerful system.

The speech signal itself carries information related to syntax, represented by speech prosody. This means that syntax and prosody interact, even if they cannot be mapped directly and unambiguously to each other (Selkirk, 2001). From a linguistic point of view, the majority of theories dealing with the syntax-prosody relationship conclude that syntax and prosody are closely related, but this relation cannot be expressed as a definite mapping between syntax and prosody. The *prosodic structure hypothesis* by Selkirk (2001) postulates that the prosodic structure of a sentence is related to (but not fully dependent on) the surface syntactic structure. In contrast, some theories argue that prosody is directly governed by the surface syntactic phrase structure (Kaisse, 1985), but evidence shows rather that the relationship syntax-

prosody is more difficult, especially as we are approaching the lower levels (or layers – the two terms are used as synonyms throughout this paper) of the prosodic hierarchy.

Approaching the same problem from point of view of human perception, several studies have proven that prosody is an important clue in syntactic parsing and that prosody constrains lexical access (Cristophe et al., 2004), which proves its essential role in human speech perception. Imaging techniques tracing human brain activity during speech perception by ERP (Event Related Potential) or PET (Positron Emission Tomography) measurements also support this hypothesis (Li-Yang-Lu, 2010), and it is suspected that prosody is a predictive clue for syntactic (and semantic) processing in human perception, justified by ERP tests allowing for the tracing of brain activity (Strelnikov et al., 2006). Indeed, brain areas situated in the dorsolateral prefrontal cortex, associated with the perception of prosody are very close to (or rather overlapping with) those responsible for syntactic analysis, forming a real prosody/syntax interface in the human brain (Strelnikov et al., 2006).

Evidence and practice in speech technology also shows the practical usefulness of the prosody/syntax interface. In text-to-speech conversion, syntactic analysis of a written sentence has become a common task prior to speech synthesis (Koutny-Olaszy-Olaszi, 2000), (Becker et al., 2006). The first initiatives date back even to the 1980s. The underlying assumption for this approach is that the required prosodic features of the sentence to be synthesized can be well predicted relying on syntactic analysis. In other words, one assumes that surface syntactic structure determines the prosodic structure. As this determination is only partial, applications were often restricted to well described domains, such as name and address synthesis (Silverman, 1993).

Despite the fact that the relation between syntax and prosody has widely been exploited in text-to-speech synthesis (Hirschberg, 1993), it is not explicitly included in speech recognition: although prosody is implicitly modelled in *segmental domain* (i.e a domain proportional to the length of a phoneme) through energy related features and implicit duration modelling, it remains the most often neglected in *suprasegmental domain* (i.e. the length of a word or group of words) as explained in Vicsi-Szaszák (2010). Some other studies have already also highlighted this technological gap (Batliner et al., 2006) long time

ago. Since then, several attempts were made to integrate prosody into speech recognition and understanding, focusing essentially on boundary detection tasks using an event detection like approach (Veilleux-Ostendorf , 1993), (Gallwitz al., 2002), (Shriberg et al., 2000). Iwano (1999) and Vicsi-Szaszák (2010) implemented alignment based segmentation and boundary detection upon this segmentation. Going one step further and mapping prosody to syntax and deducing some syntactical attributes based on prosody or perform disambiguation is even less frequently used, although is not completely unknown in speech understanding (Price et al., 1991), (Nöth et al., 2000). A fully statistical approach for information extraction based on prosody was presented by Shriberg-Stolcke (2004), without using labelled corpora to train machine learning based structural and pragmatic taggers, speaker and word recognizers.

These considerations lead and motivate us to experiment with – at least partial – recovery of the syntactic structure in speech based on prosody, an important clue when carrying out automatic interpretation of the content encoded in the speech signal, in order to assess its meaning. Applications so far in this domain are almost exclusively dedicated to disambiguation problems (Price et al., 1991), (Nöth et al., 2000), where prosody is used to select between ambiguous hypotheses (minimal pair sentences) given a speech sample and its different interpretations represented by different syntactic structures. The selection between minimal sentence pairs can be a realistic problem in automatized dialogues, however, it does not provide a globally useful framework to assess syntax based on prosody. Our goal is to fill this gap and implement and test a more globally applicable framework (in contrast to the work presented by Price et al. (1991)) for syntactic analysis based only on speech, capable of providing more detailed analysis based also on training with hand labelled data (in contrast to the work presented by Shriberg-Stolcke (2004)). The outcome of this activity can be useful in several technologies involving speech understanding, like dialogue-based automatized systems with speech interface, automatic interpretation (speech translation), and in general, in any application where analysis of meaning (focus detection, topic detection, keyword spotting, speech segmentation based on prosody, syntactic or semantic analysis, etc.) is crucial.

Instead of creating a more or less artificial corpus of minimal pair sentences, which usually provides only a moderate and often non-realistic corpus for analyzing purposes, a general, large speech corpus is used. The main interest is to analyse the nature of the relation between automatically performed prosodic analysis (phonological phrase boundary detection and classification) and automatically generated and previously disambiguated syntactic analysis (this latter represents the surface syntactic structure). The basic interest is to explore and evaluate to what extent different levels (called also layers (Selkirk, 2001)) in the prosodic and syntactic hierarchy can be mapped to each other, and to analyse further if any type of phonological or syntactic phrase exists which has a special impact on syntactic or phonological structure, respectively. An important side-outcome of the experiment is to evaluate to what extent automatic prosodic segmentation for phonological phrases can reflect the underlying syntactic structure.

Experiments to be presented in this paper were carried out for the Hungarian language, however, special emphasis is also put on the universality and possible extension of the approach for other languages.

This paper is organized as follows: First, syntactic analysis issues are revised, then the prosodic segmentation of speech is presented in details. Hereafter, experiments and results are presented for the reconstruction of the syntax based on speech prosody, followed by conclusions.

2 SYNTACTIC ANALYSIS

The syntactic analysis is performed on the transcripts of utterances which will be used for the evaluation of prosody-to-syntax mapping experiments. The syntactic analysis – provided as a syntactic phrasing – serves as a reference when correspondence of the prosodic and syntactic structure is investigated. First, some basic Hungarian specificities are briefly presented, necessary for the comprehension of the syntactic analysis method used, which has to deal with rich morphology and relatively free word order. Syntactic analysis is presented next, with a short outlook explaining the necessary morphological considerations.

2.1 *Specificities of Hungarian syntax*

Hungarian is an agglutinating language, with a very rich morphology, and consequently, grammatical relations are expressed less by word order constraints and more by suffixes. This allows also for a relatively free word order, where word ordering is more submitted to the fine semantic tuning of the meaning, as case information is available via the suffixes. For example, in English and in many other languages a basic sentence would start with the subject (noun), followed by the verb, ended by the object (noun). In Hungarian, the object is differentiated by the objective case (usually suffix -t) and hence is identifiable as object even if it is moved within the sentence.

Hungarian sentences can be divided into a topic and a predicate (or comment) part (É. Kiss, 2002). The topic part either contains constituents whose denotation (normally, an individual) counts as given in the context, or those denoting entities, properties or eventualities constituting new information that are intended to be contrasted to their alternatives. In sentences with a narrow or contrastive focus, the focused unit is placed between the topic and the verb and must directly precede the latter – this is the focus position. In other words, constituents before the verb are associated with specific functions, whereas units following the verb do not normally express new information (comment part).

Given that Hungarian is an agglutinating language, i.e. grammatical information is expressed by suffixes rather than word order, the primary role of word order is to express information structure.

(a) Mária ismeri Józsefet.

(b) Józsefet Mária ismeri.

Thus, the utterance “Mary (Mária) knows Joseph (József)” can take the forms given in (a) and (b) without a semantic change (Mary is the subject and Joseph the direct object in both sentences), but the information structure is different: sentence (a) has either broad focus with an accent on all content words (also called either a neutral sentence or a sentence with verbal accent), or Mary is in focus and hence the accented unit in the sentence (verbs are deaccented if the focus position is filled). In the latter case, the sentence could be an answer to the question “Who knows Joseph?”. In sentence (b), Mary is in fo-

cus, but the word order indicates that this sentence is about Joseph (= the topic) and includes the option of contrastivity (i.e. "… but it is Rebecca who knows Isaac").

2.2 *Syntactic phrasing*

The syntactic analyser is a language-dependent tool (however, of course, its output is standardized as used in automatic machine translation tools for example). As experiments presented in this paper were done for the Hungarian language, the freely available Hunpars tool (Babarczy et al., 2005) was used as a syntactic analyser was used. This syntactic analyser uses a so-called *phrase-structure grammar*, completed by *lexical databases* and a *morphological analyser* to perform syntactic analysis of written sentences. The analyser outputs tagged and layered syntactic analysis hypotheses for each input sentence.

In a phrase structure grammar, words are grouped into syntactic phrases, which together form a hierarchic (or layered) structure (Gazdar et al., 1985). The identification of grammatical dependencies follows based on this hierarchical grouping. The syntactic phrasal structure is output by bracketing, preserving the hierarchy so that it can be easily converted into a tree-like representation (see an example in Fig. 1).

The phrase structure grammar used in Hunpars is *head-driven* (Pollard-Sag, 1994): each syntactic phrase has a head, corresponding to the word that determines the behavior of the phrase within the syntactic constituent (embedding syntactic phrase or the sentence) located one level up in the hierarchy. For example, the syntactic phrase 'a főkonzul lányát' (the consul's daughter + Acc) is a noun phrase (NP) headed by the noun 'lány' (daughter) – this means that this phrase is an embedded phrase which behaves as it were a single noun (in Acc).

The sentence shown in Fig. 1 could be bracketed as follows:
[[<<Gróf(NP)> Vásárhelyi(NP)> <Görögországban(NP)> <kötött ki(VV)> Clause)] és(Conj) [<titkárul(NP)> <szerződtette(VV)> <a(Art) <főkonzul(NP)> lányát(NP)> (Clause)] (Sentence)]

2.3 *Morphological analysis and disambiguation of syntactic analysis*

As explained so far, syntactic phrasing of a sentence needs morphological analysis, too, in order to identify the grammatical cases and

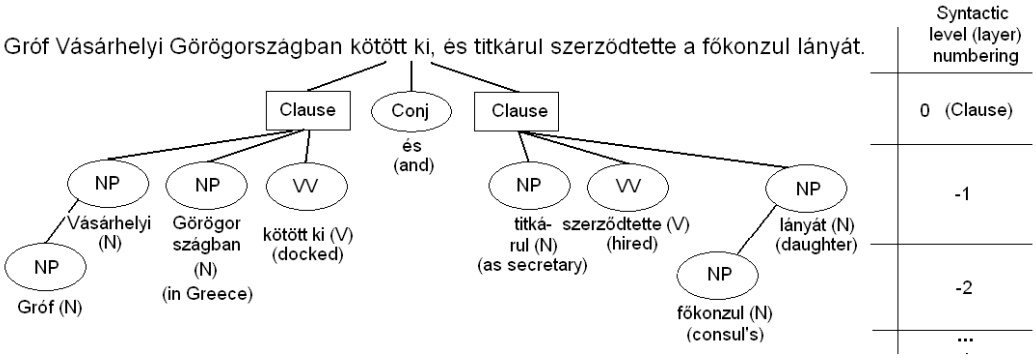

Figure 1: An example of syntactic phrase structure of the Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*" (*Count Vásárhelyi docked in Greece, and hired the daughter of the consul as his secretary*)

relations in which words are used actually. This is why a morphological analyser, called Hunmorph (Trón et al., 2005), is also used from within the tool Hunpars.

The rich morphology may also lead to homonymy: words with same spelling but with different meaning, eventually also two different stems with different suffixes may result in the same word having quite different meaning and being in different cases. This causes ambiguity during the automatic syntactic analysis. Some disambiguation is performed during syntactic analysis relying on the phrase structure grammar (Babarczy et al., 2005): based on a lexicon and some rules, a part of the concurring analysis hypotheses can be ruled out. The remaining ones, however, are all kept and output by the Hunpars tool. As further automatic disambiguation is not provided by the tool, in a case of multiple hypotheses the actually correct one was selected by an expert.

3 AUTOMATIC PROSODIC SEGMENTATION OF SPEECH

3.1 *Prosodic hierarchy model*

The model of the prosodic structure used in this work relies on the *prosodic structure hypothesis* (Selkirk, 2001). This model provides a hi-

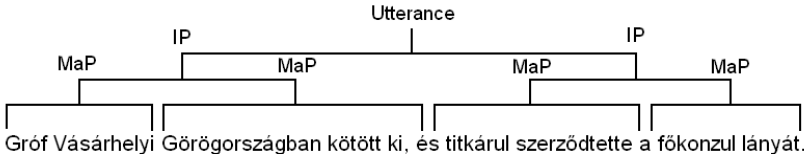

Figure 2: An example of canonical prosodic structure of a Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*"

erarchic view of the prosodic structure as follows top-down: utterances are composed of *intonational phrase*s (IP), which can be divided into *phonological phrase*s (PP). Selkirk's model differentiates between major (MaP) and minor (MiP) phonological phrases. Some studies argue that this distinction is not necessary (Ito-Mester, 2008). Indeed, the acoustic-phonetic realizations of major and minor phonological phrases seem to be very close to each other (at least for Japanese (Ito-Mester, 2008) and for Hungarian, as this issue can be language-dependent. This suggests creating a sort of recursion in the language, i.e. there is no significant difference between major and minor phonological phrases, but rather a general phonological phrase layer exists, which can embed further other phonological phrases and creates sublayers within the phonological phrase layer of the prosodic hierarchy model. However, in this work, phonological phrases are regarded as being identical with minor phonological phrases unless explicitly stated otherwise. This prosodic structure is often represented as a tree or bracketing of the utterance. An example is given in Fig. 2 for a Hungarian sentence (supposing the speaker uses the canonical prosodic patterns when uttering the sentence): "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*" (*Count Vásárhelyi docked in Greece, and hired the daughter of the consul as a secretary*). The canonical prosodic structure of this sentence could be written bracketed as: *[[<Gróf Vásárhelyi> <<Görögországban> <kötött ki és>>][<<titkárul> <szerződtette a>> <főkonzul lányát>]]*.

The prosodic hierarchy could be further refined, e.g. phonological phrases are composed of *phonological word*s, called sometimes prosodic words and so down to the syllable level, but units inferior to (minor) phonological phrases are beyond our interest in the current work, as our goal is to assess the syntax based on suprasegmental prosodic features. This means that units shorter than a phonological phrase (often

containing a single prosodic word) are not regarded as suprasegmental units in speech and fall out of interest here, as segmental domain processing of speech is a common task in automatic speech recognition (even if the used features are rather spectral or spectra-derivative and not (micro)prosodic ones), whilst suprasegmental domain processing is mostly missing in these applications (Vicsi-Szaszák, 2010) – as already mentioned in the Introduction. Another reason for focusing on the suprasegmental domain when assessing prosody is that segmental level exploitation of prosody seems to be highly language-dependent, for example in tonal languages, prosody has to be integrated into phoneme- or word-based speech recognizers in the segmental domain (Chang et al., 2000), or even in the non-tonal Japanese, prosody can be exploited in mora recognition as individual words are usually characterized by specific prosodic attributes allowing the identifications of word boundaries based on prosody (Hirose et al., 2001), (Iwano, 1999). However, whilst segmental domain use of prosody is language-dependent, the role of prosody in the suprasegmental domain is more universal and allows for the evaluation of a more general framework to assess it in this domain.

In the prosodic structure, upper-level prosodic units dominate lower-level ones, that is, for example, intonational phrase dominates the underlying phonological phrases. One of the phonological phrases belonging to the same intonational phrase usually constitutes a focal – stressed or somehow highlighted – part of the intonational phrase. In present study this means that the focus is realized with a higher F0 – local F0 peak. This phonological phrase can be called the head phrase. More generally, all phonological phrases are influenced by their location and role in the intonational phrase, which means that typical prosodic patterns can be associated with each phonological phrase. This allows us to cluster and classify individual phonological phrases and create a more or less disjunct set of phonological phrases in terms of their intonational contour, strength and location of the stress or prominence they carry, etc. In other words, clustering of phonological phrase types involves implicitly effects linked to upper level (Map or IP or utterance level) constraints and hence, phonological phrase models implicitly incorporate and reflect the upper prosodic structure of the utterance to some extent. For the Hungarian language, 6 different phonological phrase types were created in (Vicsi-Szaszák, 2005).

Prosodic label	Description
co	Clause onset PP
ss	Strongly stressed PP
ms	Medium stressed PP
ce	Low clause ending PP
cr	High ending (continuation rise) PP
ls	Low-stress PP

Table 1:
Phonological phrase types for Hungarian following Vicsi-Szaszák (2005)

They are listed in Table 1. It can be clearly seen that the distinction between them is based on the influence of higher level functions governed primarily by the intonational phrase they belong to.

The theoretic prototype of phonological phrases in Hungarian shows a smart rise of F0 at the stressed syllable, then a slightly descending contour follows. As Hungarian is a fixed-stress language (stress, if present, can almost always be found on the first syllable of the word stressed), location of the stress within the phonological phrase is not a distinctive feature.

As phonological phrases are constituents of intonational phrases, the higher level constituent influences their characteristics. Clause onset (*co*) and clause ending (*ce*) usually alter the standard phonological phrase intonational contour, so does the focus (strongly stressed phonological phrase, *ss*) and the continuation rise (*cr*). The continuation rise usually alters the subsequent phonological phrase, causing the stress to be often undetectable or turned into (low) stress (*ls*). Although Selkirk (2001) underlines that “prosody is strictly layered”, that is, higher level constituents immediately influence only the constituents located one level below, it is clear that even utterance-level constraints might have their effect on phonological phrases, as it is the case between low (ce) and high (cr) clause endings: alterations provoked by upper-level constraints propagate further down to the (minor) phonological phrase layer. These also mean that modelling done in the phonological phrase layer implicitly incorporates higher-layer information and may be used on these higher layers to perform some analysis. This hypothesis is also addressed in the paper.

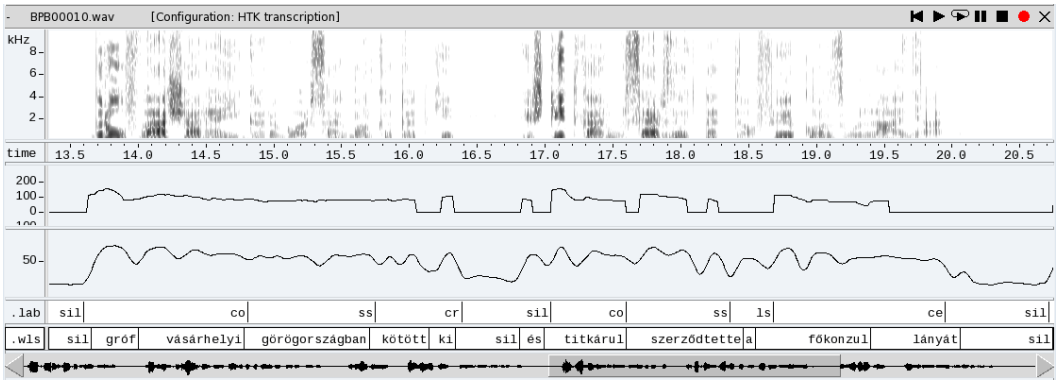

Figure 3: An example of the output of the prosodic segmenter for the Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*". Spectrogram, partly interpolated F0, long time energy, prosodic segmentation and word level segmentation is also given.

3.2 *Automatic alignment of phonological phrases*

For Hungarian, an automatic phonological phrase classifier and aligner software has been made available (Vicsi-Szaszák, 2010). The parallel classification and alignment operates theoretically like a Hidden Markov Model-based automatic speech recognizer used in word or phoneme segmentation mode, but the features used are prosodic ones (see subsection 3.3) and the models are those of the phonological phrases presented in Table 1. All these mean that this tool performs automatic segmentation for phonological phrases: detects hypothetised phonological phrase boundaries and classifies the phrases. An example output of this is shown in Fig. 3. The authors underline that in this alignment approach, continuous tracking of prosody over the whole speech signal is implemented instead of looking for discrete markers, indices of breaks or tones. This provides a soft and more flexible framework, which is believed to be also closer to human perception processes.

The alignment for phonological phrases operates only at one level (or layer) in the prosodic hierarchy, that of phonological phrases. However, as phonological phrases are influenced directly by intonational-phrase-level constraints and indirectly by utterance level constraints, the prosodic structure in terms of the layering of the pro-

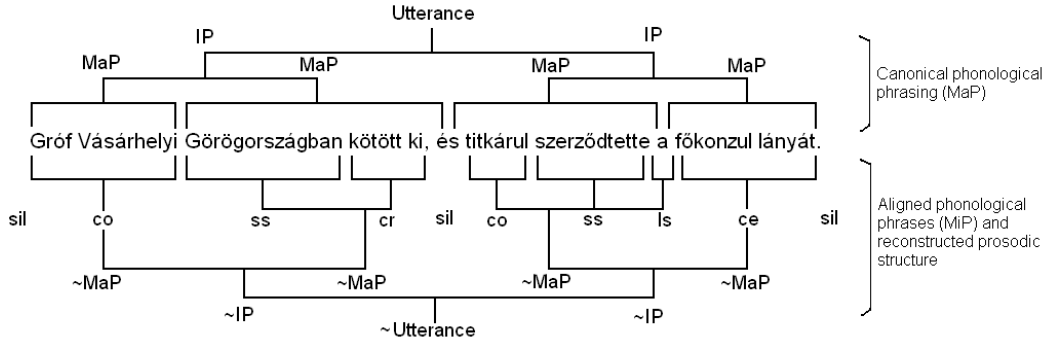

Figure 4: Reconstruction of the prosodic layering based on phonological phrase alignment for the Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*".

sodic hierarchy becomes at least partly recoverable based only on the output of phonological phrase sequence aligned to an utterance. The layering of prosody will be represented in the different types of the phonological phrases, e.g. the beginning of a *clause onset (co)* phonological phrase represents the beginning of an intonational phrase and might represent the beginning of a new utterance. In a similar way, the end of a *continuation rise (cr)* phonological phrase is also the end of the embedding intonational phrase. Or, the end of a *low clause ending (ce)* phonological phrase is also the end of the embedding intonational phrase (one level higher) and the utterance itself two levels higher which embeds the intonational phrase. The prosodic segmentation of the utterance shown in Fig. 3 yields a prosodic layering (hierarchy) as shown in Fig. 4 (which is quite close to the one presented in Fig 2).

It is important to notice that the "deepness" of the prosodic segmentation based on phonological phrase alignment is speaker- and speaking style-dependent (Vicsi-Szaszák, 2010). If the speaker uses a rich prosody in her/his utterances, the deepness of the segmentation can reveal a high ratio of minor phonological phrases or even more, the boundaries of distinct words in speech (see subsection 3.6), or even separate constituents of compound words in some cases. However, if the speaker uses a "flat" prosody, the deepness of the prosodic segmentation will degrade, in extreme cases the intonational phrase cannot be divided further for phonological phrases as prosodic cues are missed or missing (deleted). The very flexible time warping capability of Hid-

den Markov models (composed of up to 11 states) explains why parts of speech of such variable length can be processed using the same phonological phrase model set and approach, i.e. an utterance composed of 6 words corresponds to one no more dividable intonational phrase in an utterance, whilst the same 6 words with the same meaning can be divided into 3 or more phonological phrases in another realization (utterance).

The prosodic segmentation is based on suprasegmental prosodic features (fundamental frequency and energy) described in subsection 3.3. During prosodic segmentation, a sophisticated prosodic phrase sequential model can be used, which constrains which phonological phrase can follow a given other one. This model has an identical role to a language model in speech recognition. The model incorporates the prosodic structure presented in subsection 3.1 and presumes that the utterance is composed of phonological phrases, which are influenced by higher level (intonational phrase or utterance level) constraints. The basic topology is as follows: each utterance begins with a clause onset phonological phrase (*co*) and ends with either a low clause ending phonological phrase (*ce*) or a high ending phonological phrase (*cr*). The utterance can embed several further intonational phrases starting with strongly stressed phonological phrases (*ss*) and ending with continuation rise (*cr*), if they are not utterance-final IPs/PPs. Stressed phonological phrases (*ss* and *ms*) are allowed to appear anywhere within the utterance. The *ss* symbol refers to a stronger accent expected to be placed at the focus of the utterance. The low-stress phonological phrase (*ls*) is allowed optionally, but only immediately after a high ending (continuation rise, *cr*). Between all utterances, a silence (*sil*) is supposed. However, this relatively strict utterance model is supposed to fit read or moderately spontaneous speech. If speech is spontaneous, a better choice can be using an utterance model which simply allows every phonological phrase entity to be aligned with no respect to its context (Szaszák-Nagy-Beke, 2011).

3.3 *Acoustic-prosodic pre-processing*

The acoustic-prosodic features used in phonological phrase models of the prosodic segmenter rely on fundamental frequency and energy. Fundamental frequency (F0) is extracted by ESPS method using a 25 ms long window. Intensity is computed with a window of 150 ms.

The frame rate for both variables is set to 10 ms. The obtained F0 contour is first filtered with an anti-octave jump tool in order to eliminate or at least reduce pitch tracking errors. This is followed by a smoothing with a 5 point mean filter. In order to ensure a relatively continuous F0 contour, a linear interpolation is carried out in logarithmic domain. However, the interpolation is omitted for voiceless sections which are longer than 150 ms or for sections where the F0 difference between the two neighbouring voiced parts shows a rise reaching at least 110% after an unvoiced part. Delta and acceleration coefficients are also appended to both F0 and intensity streams.

3.4 *Training of the prosodic segmenter*

According to Vicsi-Szaszák (2010), the training of the acoustic-prosodic models of the prosodic segmenter was performed on a part of the Hungarian database BABEL (Roach et al., 1996), hand-labelled initially for phonological phrases based primarily on the F0 contour, but also on the annotators' perception of phonological-phrase-initial stress after listening to the utterance. 1600 utterances from 32 speakers were used to train 11 state left-to-right Hidden Markov Models for each phonological phrase + silence presented in Table 1. The reason for training 11 state models (iteratively optimized during validation) is the suprasegmental nature of prosody: phonological phrases usually correspond to longer sections of speech compared, for example, to phoneme models in automatic speech recognition, which are 1- to 5-state long, but most often 3-state long).

3.5 *Initial testing of the prosodic segmenter*

Initial testing of the prosodic segmenter was carried out using 10-fold cross-validation. This means that after randomly ordering the utterances, they were divided into 10 equal subsets (160 utterances each). Training and testing was then performed 10 times by using each subset as a test set and the remaining 9 as the train set. In 10-fold cross-validation each utterance is tested in one of the cycles, but it is guaranteed that once an utterance is placed into the test set, it is excluded from the train set.

For utterances under test, a phonological phrase alignment was generated which was compared to its hand-labelling used as reference. Once these alignments were available for all utterances, four

performance indicators were measured: the recall and precision of the phonological phrase boundary recovery, the average time deviation between detected and reference phonological phrase boundaries and the accuracy of the classification regarding the type of phonological phrases.

The recall is measured with the following formula:

$$\text{Recall} = \frac{tp}{tp + fn}, \tag{1}$$

where tp stands for true positives, that is, the number of phonological phrase boundaries correctly found within 150 ms of the original one in the reference; fn stands for false negatives, that is, the number of missed phonological phrase boundaries (present in reference but not detected).

Precision is measured as:

$$\text{Precision} = \frac{tp}{tp + fp}, \tag{2}$$

where fp stands for false positives: phonological phrase boundaries detected where they should not be according to the reference, or more than 150 ms apart from reference phonological phrase boundary.

The recall of phonological phrase alignment-based prosodic segmentation was 82.1%, the precision was 77.7%.

The average time deviation (σ_t) of segmentation for phonological phrases was measured for true positives as:

$$\sigma_t = \frac{1}{tp} \sum_{i=1}^{tp} \left| t_i - t_i^{ref} \right|, \tag{3}$$

where tp stands again for the number of phonological phrase boundaries correctly found within 150 ms vicinity of the reference boundary. t_i is the detection time of the i^{th} phonological phrase boundary, t_i^{ref} is the location of the corresponding reference boundary. For the above tests, average time deviation was found to be: $\sigma_t = 50.4$ ms.

Finally, classification accuracy is measured as the ratio of correctly classified phonological phrase boundaries (tp_{cc}) versus all true positive phonological phrase boundaries (tp):

$$\text{Acc} = \frac{tp_{cc}}{tp}. \tag{4}$$

Classification accuracy was found to equal overall 73.1%.

3.6 *Prosodic segmentation vs. word boundaries*

Vicsi and Szaszák used a similar prosodic segmentation for phonological phrases to partially recover word boundaries in Hungarian and Finnish languages (Vicsi-Szaszák, 2010), (Vicsi-Szaszák, 2005). Of course not all phonological phrase boundaries coincide with word boundaries, the authors also underline that for Hungarian, a word boundary detector in the strict sense cannot be implemented in contrast to the mentioned Japanese (Hirose et al., 2001). However, they trained the prosodic-acoustic models of phonological phrases on samples in which phonological phrase boundaries coincided with word boundaries. Highly relying on the first syllable fixed stress of Hungarian, word boundaries were predicted in the vicinity of phonological phrase boundaries. Analysis of word boundary detection rates based on phonological phrase alignment showed 77.3% precision and 57.2% recall rate for Hungarian (on BABEL speech database), 69.2% precision and 76.8% recall rate for Finnish allowing a maximum of ± 100–150 ms deviation between phonological phrase and word boundary markers (Vicsi-Szaszák, 2005). The goal of the experiments described in present paper can be related to this issue, namely, to prove or to disclaim the conjecture that the detected word boundaries correlate well with syntactic phrase boundaries, while missed word boundaries are more likely to be embedded within a syntactic phrase, and therefore tend to form a union both prosodically and syntactically.

4 ANALYSING THE PROSODY-TO-SYNTAX MAPPING

The main goal of the paper is to present a detailed analysis regarding the prosody-to-syntax automatic mapping possibilities in spoken language. This implies the comparison between the prosodic and syntactic structures, obtained based on analyses presented so far both for prosody and syntax. The syntactic phrasing will be used as reference, and hence – although it was primarily obtained in automatic way – it has to be checked and disambiguated by human experts. The automatically obtained prosodic phrasing on the other hand is left intact as it is produced by the prosodic segmenter tool. The reason for this is that this approach will permit to evaluate the usability of the pro-

posed algorithm in real conditions, where the automatically obtained prosodic phrasing may contain errors.

4.1 *Material and method*

The material used for current experiments was taken from the BABEL Hungarian language speech database (Roach et al., 1996). BABEL is a read speech database, involving 60 non-trained speakers' data. The speech material covers paragraphs composed of at least 6 sentences (contextually linked), numbers, isolated digits, and CVC items. A subcorpus taken from the paragraphs was used, containing 155 different sentences uttered by a total of 60 speakers. Most of the sentences occurred at least two times, hence a total set of 330 sentences was used.

The utterances were segmented on word level, obtained by performing automatic forced alignment on word-level transcriptions with a Hungarian language ASR. Indeed, an automatic phoneme segmentation was performed, which was traced back to word-level alignment. This means that time positions of word boundaries were known. This will be necessary for temporal word boundary information, as syntactic phrase boundaries themselves are located always on word boundaries.

In order to obtain the syntactic analysis, sentences (transcriptions of the speech utterances) were fed into the Hunpars syntactic analyser. Where the syntactic analyser yielded unresolved ambiguity, a human expert intervened in order to leave one and only one syntactic parsing for every sentence contained in the speech utterances (as minimal pair sentences were not included in the material, there was always only one canonically correct syntactic parsing candidate for each sentence).

In parallel, speech utterances were fed into the prosodic segmenter tool too. This produced the phonological phrase alignment of the utterances. As the prosodic segmenter tool was also trained on the BABEL database (and a testing of 10-fold cross-validation was also done for it as described in subsection 3.5), special attention was paid to ensure that the utterance currently under analysis is excluded from the training set of the prosodic segmenter.

Hereafter, the correspondence of the automatically detected phonological phrase boundaries and syntactic phrase boundaries was investigated. Correspondence was assessed separately on each syntactic level (layer) of the syntactic hierarchy, to a depth of 5 levels (top-

down: 0, −1, −2, −3, −4): sentences are divided into clauses (level 0). Clauses consist of first level syntactic phrases (level −1), which can contain daughter phrases (level −2) and so on down to level −4. While levels 0, −1 and −2 are quite common and occur in most of the sentences, deeper embedding (level −3 and especially level −4) is quite rare. The numbering of the syntactic layers is included in Fig. 1 for evidence.

The main interest is to see whether syntactic phrase boundaries can be detected based on phonological phrase alignment, and whether the syntactic hierarchy (layering) can be reconstructed based on the aligned phonological phrases (or prosodic layering, as the differentiation among phonological phrases allow for the reconstruction of the latter, as explained in subsection 3.2 and in Fig. 4).

Syntactic and phonological phrase boundaries were considered to meet if they occurred within 150 ms time interval. This value was chosen based on the following considerations:

- the time interval should allow some deviation in a range of a length of a demi-syllable, because reference word boundaries (necessary for the identification of the onset and ending times of the syntactic phrases) are segmented automatically, and
- the prosodic segmenter also displays some uncertainty (for example, if an utterance ends with an unvoiced sound, it is often inevitably chopped).
- phonological phrases aligned by the prosodic segmenter are much longer than 150 ms (average phonological phrase length is 618 ms for the test corpus with a standard deviation of 211 ms).

Syntactic phrase (XP) onsets were always aligned to phonological phrase (PP) onsets, syntactic phrase endings were always aligned to phonological phrase endings.

4.2 *Recovering syntactic phrase boundaries*

In the first experiment, phonological phrase segmentation is used to recover syntactic phrases automatically. The performance is evaluated using the *recall* measure defined in equation (1), but now tp stands for correctly recovered syntactic phrase boundaries (true positives) and fn stands for missed syntactic boundaries (false negatives). The type of the aligned phonological phrase is irrelevant in this exper-

Table 2: Recall of syntactic phrase boundaries by phonological phrase boundaries summarized for all phonological phrase types. 1B/L = one (highest level) syntactic boundary kept per word; MB/W = multiple syntactic boundaries allowed per word

Syntactic	Onset		Ending		# of occ.
Level	**1B/W**	MB/W	**1B/W**	MB/W	(MB/W)
0	**0.85**	0.85	**0.79**	0.79	3124
−1	**0.45**	0.70	**0.48**	0.68	10339
−2	**0.42**	0.70	**0.48**	0.69	5763
−3	**0.44**	0.74	**0.45**	0.65	814
−4	**0.48**	0.70	**0.50**	0.67	187
All	**0.54**	0.72	**0.55**	0.69	20227

imental setup, currently the only interest is to see what portion of syntactic phrase boundaries are detectable on the different syntactic layers, based on the phonological phrase alignment.

Results are shown in Table 2 separated for phrase onsets and phrase endings. As multiple-level syntactic embeddings are possible, several syntactic boundaries can occur at the same place. In one scenario, only the highest-level syntactic boundary is counted in case of multiple level occurrences (referred to as 1B/W), while in the other one, all different level syntactic boundaries are counted (MB/W). This means that a word preceded by level 0, −1 and −2 syntactic boundaries yields one 0 level hit (true positive) in 1B/W if detected, but gives a total of 3 hits, one for each level 0, −1 and −2 in MB/W if detected (and, of course, gives false negatives for all the 3 levels if remains undetected).

Average recall rate was 71% (in MB/W) or 55% (in 1B/W), which is considerably higher in the case of clauses: 85% (onsets) and 79% (endings). A total of around 70% of the syntactic phrase boundaries can be detected on each syntactic layer. Deeper syntactic embedding did not seem to degrade detection rates. For statistical evidence Kruskal-Wallis tests were performed on the obtained data. These also confirm that phonological and syntactic phrases are correlated ($\chi^2 = 6430.606; p < 0.000$).

Pairing up the corresponding syntactic phrase onset and syntactic phrase ending boundaries on each level and comparing recall rates by Mann-Whitney and Wilcoxon W tests show that on clause level, onsets are significantly better detected ($Z = -7.807; p < 0.000$). However,

on levels -1 and -2, there is no significant difference in recall rates counted for onsets and endings (level 1: $Z = -0.407, p > 0.1$; level 2: $Z = -0.016; p > 0.1$). There were no significant differences either on levels -3 and -4.

Non-clause (< 0) syntactic levels do not yield different recall rates (either in 1B/W or in MB/W settings), this means that lower level syntactic phrases are not less intensively marked by prosody: clauses can be identified by a higher recall rate, but there is no significant difference between syntactic phrases depending on the syntactic layer ($\chi^2 = 0.224; p > 0.1$). Each syntactic phrase seems to behave as an independent entity in terms of detectable prosodic features, independently of its position in the syntactic hierarchy. These findings may also support theories supposing a recursive nature of speech prosody (cf. Wagner, 2005).

4.3 *Reliability of the syntactic phrase recovery*

In the next step, the reliability of the segmentation was analysed, separated for all phonological phrases (except for silence (*sil*)). The reliability of the phonological phrase alignment (i.e if a phrase boundary is detected based on prosody, to what extent it is sure that there is a real syntactic boundary there or that the hit is not a false one) is measured with precision according to equation (2), but now tp stands for the number of phonological phrase boundaries which coincide with syntactic phrase boundaries, and fp stands for inserted phonological phrase boundaries which do not coincide with syntactic phrase boundaries within 150 ms (false positives). Precision measures are shown in Fig. 5 for phrase onsets and endings, separated for each phonological phrase type used. As it can be seen, the onset of a *co* phrase yields a good precision rate for syntactic phrase onsets, in parallel with the hypothesis that the beginning of a *ce* phrase should refer to a level 0 syntactic phrase, which is prosodically better marked than deeper syntactic phrases. The ending of the *ce* phrase is associated more often with deeper and hence less prosodically marked syntactic phrases (see later Table 3). Phonological phrase endings of *ce* and *cr* phrases give high precision for syntactic phrase endings: again, this can refer to corresponding level 0 syntactic phrase endings which are prosodically better marked. (see later Table 4) These hypotheses are addressed in the next subsection (subsection 4.4).

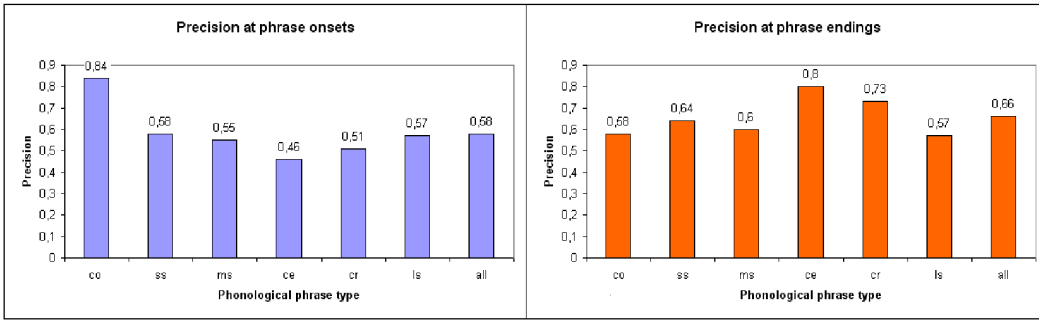

Figure 5: Precision of syntactic phrase recovery based on phonological phrase boundary detection (within 150 ms) for phrase onsets (left) and endings (right)

4.4 *Towards a reconstruction of syntactic layering*

As presented in subsection 3.2, the prosodic layering can be – at least partially – reconstructed based on the type of the phonological phrases. The next analysed point is whether there can be found some interconnection between the type of phonological phrase and the position in the hierarchy of the syntactic phrase they refer to. This could also explain differences in precision seen in Fig. 5 and justify the hypotheses raised. This would mean that not only the syntactic phrase boundaries, but also the syntactic structure in terms of its layering may become recoverable based on phonological phrase alignment.

The distribution of the aligned phonological phrases was hence examined on each syntactic layer, separately, in order to see whether some types of phonological phrases can be associated with specific syntactic layers or not. Tables 3 (for phrase onsets) and 4 (for phrase endings) show relative frequencies of the layer position of the recovered syntactic phrase (to which layer it belongs to in the syntactic hierarchy) depending on the type of the phonological phrase.

Based on the results in Table 3, a detected *co* type phonological phrase onset corresponds to a clause onset with 86% relative frequency. This means that this type of phonological phrase is a good indicator of a clause onset. Level −1 syntactic phrase onsets are well predictable if the phonological phrase type is *ss*, *ms*, *ce*, or, to a lesser extent, *cr*. Phonological phrase type *ls* onset is ambiguous, it can sign both a clause onset (50% rel. frequency) and a first level syntactic phrase onset (41%). Down from syntactic level −2, all phonologi-

Phonological phrase type	Distribution of XP levels 0	−1	−2	−3	# of occ.
co	**0.86**	0.07	0.04	0.02	1736
ss	0.12	**0.78**	0.07	0.02	2517
ms	0.09	**0.83**	0.06	0.01	1399
ce	0.14	**0.80**	0.04	0.02	2094
cr	**0.22**	**0.72**	0.04	0.01	1326
ls	**0.50**	**0.41**	0.07	0.02	1467
all	0.36	0.56	0.05	0.02	10539

Table 3: Distribution of syntactic phrase (XP) levels (or layers) based on phonological phrase type (phonological phrase onsets compared to syntactic phrase onsets)

Phonological phrase type	Distribution of XP levels 0	−1	−2	−3	# of occ.
co	0.05	**0.74**	0.11	0.08	1736
ss	0.09	**0.68**	**0.20**	0.03	2517
ms	0.08	**0.68**	**0.18**	0.04	1399
ce	**0.83**	0.11	0.04	0.02	2094
cr	**0.60**	**0.28**	0.09	0.03	1326
ls	0.13	**0.64**	**0.17**	0.06	1467
all	0.34	0.49	0.13	0.04	10539

Table 4: Distribution of syntactic phrase (XP) levels (or layers) based on phonological phrase type (phonological phrase endings compared to syntactic phrase endings)

cal phrase types are distributed uniformly, the aligned phonological phrase type cannot be used to predict syntactic level. Results prove that intonational phrases and clauses are very closely related, and that clauses can be automatically well separated from lower-level syntactic phrases. This means that two layers of the syntactic hierarchy can be accurately recovered: level 0 and lower levels, which cannot be further separated (but levels under level −1 occur much more rarely then level −1 phrases and hence, the major skeleton (the top) of the syntactic structure can be recoverable).

The detected *ce* phonological phrase ending mostly corresponds to a clause ending, this is approved by the 83% frequency (Table 4). The ending of a phonological phrase of type *cr* signs often a clause ending (60%), although it can also correspond to a level −1 syntactic phrase ending with a relatively high frequency (28%). Ending of phonological

phrases of types *co* predict a level −1 syntactic phrase ending with 74% frequency, endings of phonological phrases *ss*, *ms* and *ls* can refer to level −1 and −2 syntactic phrase endings. Levels −1 and lower levels cannot be separated further based on the comparison of endings of phonological phrases and syntactic phrases.

4.5 *Head classification of the syntactic phrase*

Relations between the types of phonological (co, ss, ms ce, cr, ls) and syntactic phrases (NP, AdjP, AdvP, NumP, VV, VV-Inf, PostpP) were also investigated. It was found that there is no significant difference in the phonological phrase type depending on the type of syntactic phrase in the Hungarian language ($\chi^2 = 0.349; p > 0.1$). This result is not surprising, especially with regard to the relatively free word order of Hungarian. In other languages, where the position of syntactic constituents (words or phrases) refers to grammatical relations, syntactic phrase classification (based on the head) might be possible, as clause onsets and endings are known, however, this issue would need further experimental examination and evaluation. For morphologically rich languages characterized by a more free word ordering, morphological analysis seems to be unavoidable for such purposes. In speech-based systems, this involves the use of automatic speech recognition, the output of which could be segmented to phonological phrases, and then fed into a syntactic parser and morphological analyser.

4.6 *Robust intonational phrase – clause recovery*

Precision and recall of phonological phrase segmentation were also analysed with a reduced phonological phrase set: *ms* and *ls* phonological phrase types were discarded during the phonological phrase alignment, as *ss* was expected to replace *ms*. Phonological phrase type *ls* was discarded because it yielded ambiguous results in syntactic phrase onset detection regarding the identification of the syntactic level. Results for phonological phrase/syntactic phrase onsets show (see table 5 for phrase onsets and table 6 for phrase endings) significantly higher precision (overall 64% for onsets, 65% for endings, see Fig. 6) and lower recall (overall 48% at onsets, 47% at endings, 1B/W) rates, while precision of phonological phrase/syntactic phrase ending detection is not significantly different, but recall rates are worse. The lower recall can be explained by the fact that phonological phrases with less charac-

Phonological phrase type	Distribution of XP levels 0	−1	−2	−3	# of occ.
co	0.88	0.07	0.02	0.02	1835
ss	0.13	0.77	0.07	0.02	3455
ce	0.26	0.67	0.04	0.02	1914
cr	0.37	0.58	0.04	0.01	1782
all	0.42	0.51	0.05	0.02	8986
Recall	0.80	0.39	0.34	0.37	

Table 5:
Distribution of syntactic phrase (XP) layers based on phonological phrase type with reduced phonological phrase set (onsets compared to onsets), 1B/W recall is also shown in the last line

Phonological phrase type	Distribution of XP levels 0	−1	−2	−3	# of occ.
co	0.03	0.43	0.07	0.04	1835
ss	0.05	0.45	0.12	0.02	3455
ce	0.50	0.12	0.03	0.01	1914
cr	0.41	0.19	0.06	0.03	1782
all	0.21	0.32	0.08	0.02	8986
Recall	0.67	0.41	0.39	0.39	

Table 6:
Distribution of syntactic phrase (XP) layers based on phonological phrase type with reduced phonological phrase set (endings compared to endings), 1B/W recall is also shown in the last line

teristic stress (*ms* and *ls*) are sometimes identified as *ss* but may also remain undetected (phonological phrase *ss* cannot replace all of their occurrences). Interpreting these in a prosodic hierarchy point of view, this approach seems to operate on major phonological phrase layer and not on the minor one. As it allows for more precise clause onset detection (see Table 5), it can be used individually or combined to the minor phonological phrase alignment based recovery approach (subsection 4.4) if higher precision is required in the reconstruction of the top layer of the prosodic hierarchy.

5 CONCLUSIONS

In the paper, automatic recovery of the syntactic structure was addressed based on prosody. The output of a phonological phrase level segmentation tool was used to predict syntactic phrase boundaries. Up to 85% of the clause boundaries and about 50% of further non-

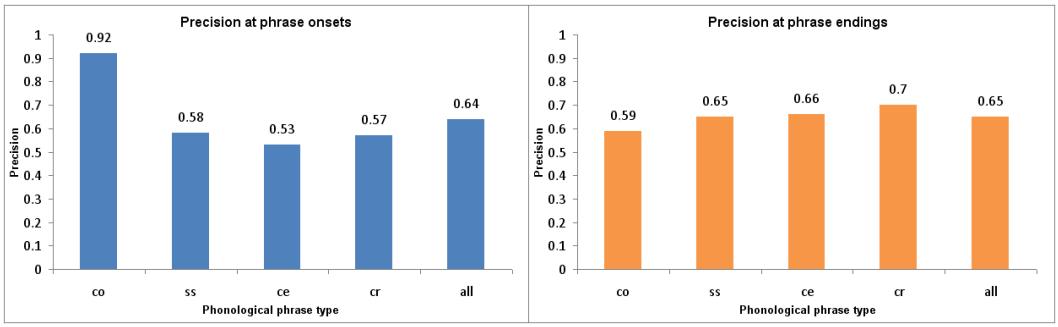

Figure 6: Precision of syntactic phrase recovery based on phonological phrase boundary detection (within 150 ms) for phrase onsets (left) and endings (right), using the reduced phonological phrase set

coinciding lower-level syntactic phrase boundaries could be automatically recalled. Precision of clause boundary detection (i.e. when intonational phrase boundaries met clause boundaries) was 84% (even 92% with a reduced phonological phrase model set), precision of lower level syntactic boundary detection (i.e. syntactic phrase boundary met by phonological phrase boundary) ranged between 46% and 58%, allowing at most 150 ms deviation between the phonological and the syntactic boundary markers. Clause level and underlying syntactic phrase level could be well separated based on the type of the aligned phonological phrase.

No relation was found between the type of the syntactic phrase and the type of the phonological phrase. This is not surprising, as the investigated language was the Hungarian, characterized by free word order. Based on the results presented, prosody seems to have a synchronizing and signalling function in terms of identification of the underlying syntactic units, but does not seem to reflect the finer relations among these units in lower syntactic layers. These results also raise some evidence of the recursive nature of speech prosody: syntactic boundaries are well signalled by prosody irrespective of the syntactic layer by same recall rates for each layer with no significant difference among them), but after a point in the hierarchy (layers down from layer -1 in the syntactic structure and layers of minor phonological phrases in the prosodic structure), layering information disappears from prosody, but layer boundaries remain detectable with the same

accuracy. This can suggest a hypothesis that at this point semantics takes over the layering role from prosody, however, this issue needs further investigation.

Although the results shown in the paper were obtained for the Hungarian language, no language-specific knowledge was used during the experiments, per se, the syntactic analyser and prosodic segmenter modules are language specific. The prosodic segmenter module, on the other hand, has already been successfully used for Finnish and German languages (Vicsi-Szaszák, 2010). The presented results can highly contribute to support automatic speech understanding. Possible application areas of the results can be speech segmentation based on prosody for supporting meaning extraction, surface syntactic structure analysis based on speech, support for text-based syntactic analysis, topic-comment separation, keyword spotting where the keyword is stressed, etc.

ACKNOWLEDGEMENTS

Authors express their gratitude to Alexandra Markó, ELTE University, Budapest, Hungary and to former MSc. student Katalin Nagy at the Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics for their help in the experiments presented in this paper.

Authors would like to thank the CESAR (http://cesar-project.net/) project, funded under the ICT-PSP (Grant Agreement No. 271022), a partner of META-NET (http://meta-net.eu), for its support for the work done on the BABEL corpus.

REFERENCES

A. Babarczy, G. Bálint, G. Hamp, A. Rung (2005), Hunpars: a rule-based sentence parser for Hungarian, *Proc. of the 6th International Symposium on Computational Intelligence*, Budapest, Hungary.

A. Batliner , B. Möbius , G. Möhler , A. Schweitzer and E. Nöth (2006), Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground, *Proc. Eurospeech 2001, Vol. 4.*, Aalborg, Denmark, pp. 2285–2288.

S. Becker, M. Schröder, W.J Barry (2006), Rule-based Prosody Prediction for German Text-to-Speech Synthesis, *Speech prosody*, Dresden, Germany, p. 31

J. BUTZBERGER, H. MURVEIT, E. SHRIBERG and P. PRICE (1992), Spontaneous speech effects in large vocabulary speech recognition applications, *Proceedings of the 1992 DARPA Speech and Natural Language Workshop*, pp. 339–343.

E. CHANG, J.-L. ZHOU, S. DI, C. HUANG and K.-F. LEE (2000), Large vocabulary Mandarin speech recognition with different approaches in modeling tones, *International Conference on Spoken Language Processing*.

A. CHRISTOPHE, S. PEPERKAMP, C. PALLIER, E. BLOCK, amd J. MEHLER (2004), Phonological phrase boundaries constrain lexical access: I. Adult data. *Journal of Memory and Language*, Vol. 51, pp. 523–547.

F. GALLWITZ, H. NIEMANN, E. NÖTH, W. WARNKE (2002), Integrated recognition of words and prosodic phrase boundaries. Speech Communication, Vol. 36. pp. 81–95.

G. GAZDAR, E.H. KLEIN, G.K. PULLUM and I.A. SAG (1985), *Generalized Phrase Structure Grammar*, Oxford: Blackwell, and Cambridge, MA: Harvard University Press.

K. HIROSE, N. MINEMATSU, Y. HASHIMOTO and K. IWANO (2001), Continuous Speech Recognition of Japanese Using Prosodic Word Boundaries Detected by Mora Transition Modeling of Fundamental Frequency Contours, *Proceedings of ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, USA, pp.61–66.

J. HIRSCHBERG (1993), Pitch Accent in Contex: Predicting Intonation and Prominence from Text, *Arftificial Intelligence*, Vol. 63., No. 1–2.

J. ITO and A. MESTER (2008), *Rhythmic and interface categories in prosody* Ms., UC Santa Cruz. Presented at PRIG (Prosody Interest Group), UCSC.

K. IWANO (1999), Prosodic Word Boundary Detection Using Mora Transition Modeling of Fundamental Frequency Contours – Speaker Independent Experiments. *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 99)*, Budapest, Hungary, vol.1, pp.231–234.

E.M. KAISSE (1985), Connected Speech: The Interaction of Syntax and Phonology, *Academic Press*, San Diego.

K. É KISS (2002), The syntax of Hungarian. *Cambridge University Press*, UK.

I. KOUTNY, G. OLASZY, P. OLASZI (2000), Prosody prediction from text in Hungarian and its realisation in TTS conversion, *International Journal of Speech Technology*, Vol. 3–4, pp. 187–200.

X. LI, Y. YANG, Y. LU (2010), How and when prosodic boundaries influence syntactic parsing under different discourse contexts: An ERP study *Biological Psychology*, Volume 83, Issue 3, March 2010, pp. 250–259.

E. NÖTH, A. BATLINER, A. KIESSLING, R. KOMPE, and H. NIEMANN (2000), Verbmobil: the use of prosody in the linguistic components of a speech understanding system, *IEEE Trans*, ASSP, Vol. 8, pp. 519–532.

C. POLLARD, I.A. SAG (1994), Head-Driven Phrase Structure Grammar. *University of Chicago Press*.

P.J. PRICE, M. OSTENDORF,S. SHATTUCK-HUFNAGEL, C. FONG (1991), The use of prosody for syntactic disambiguation, *Journal of the Acoustical Society of America* Vol. 90 No. 6, pp. 2956–2970.

P. ROACH et al. (1996), BABEL: An Eastern European multi-language database, *Proc. of the* 4^{th} *International Conference on Speech and Language Processing*, Philadelphia, USA, Vol 3. pp. 1892–1893.

E. SELKIRK (2001), *The Syntax-Phonology Interface*, in N.J. SMELSER and P.B. BALTES (Eds), International Encyclopaedia of the Social and Behavioural Sciences, Oxford: Pergamon, pp. 15407–15412.

E. SHRIBERG, A. STOLCKE, Direct modeling of prosody: An overview of applications in automatic speech processing, Proc. ISCA International Conference on Speech Prosody, 2004.

E. SHRIBERG, A. STOLCKE, D. HAKKANI-TÜR, G. TÜR (2000), Prosody-Based Automatic Segmentation of Speech into Sentences and Topics, Speech Communication 32(1–2), 127–154.

K. SILVERMAN (1993), On costumizing prosody in speech synthesis: names and addresses as a case in point, *Proc. ARPA Workshop on Human Language Technology*, pp. 317–322.

K.N. STRELNIKOV, V.A. VOROBYEV, T.V,CHERNIGOVSKAYA, S.V. MEDVEDEV (2006), Prosodic clues to syntactic processing – a PET and ERP study, *NeuroImage* Volume 29, Issue 4, pp. 1127–1134.

M. SZARVAS, T. FEGYÓ, P. MIHAJLIK, P. TATAI (2000), Automatic Recognition of Hungarian: Theory and Practice. International Journal of Speech Technology 3:(3–4) pp. 237–251.

Gy. SZASZÁK, K. NAGY and A. BEKE (2011), Analysing the correspondence between automatic prosodic segmentation and syntactic structure, Proc of Interspeech 2011, Florence, Italy, pp. 1057–1061.

V. TRÓN, L. NÉMETH , P. HALÁCSY, A. KORNAI, Gy. GYEPESI, D. VARGA (2005), Hunmorph: Open source word analysis. Proceedings of the ACL 2005 Workshop on Software, Ann Arbor, MI, pp. 77–85.

N.M. VEILLEUX, M. OSTENDORF (1993), Prosody/parse scoring and its application in ATIS. Proc. ARPA Human Language Technology Workshop ’93. pp 335–40.

K. VICSI and Gy. SZASZÁK (2005), Automatic Segmentation of Continuous Speech on Word Level Based on Supra-segmental Features, *International Journal of Speech Technology*, Vol. 8 No. 4, pp. 363–370.

K. VICSI, Gy. SZASZÁK (2005), Using prosody to improve automatic speech recognition, *Speech Communication* Vol. 52, No. 5, pp. 413–426.

M. WAGNER (2005), *Prosody and recursion*, Ph.D. dissertation, MIT.