Prosodic boundaries in Czech: an experiment based on delexicalized speech

Tomáš Duběda

Institute of Phonetics, Charles University in Prague, Czech Republic
dubeda@ff.cuni.cz

ABSTRACT

The present experiment attempts to determine the role of prosody in the identification of stress unit boundaries in Czech, using three types of delexicalized stimuli evaluated by native speakers. After describing the relative contribution of intonation and duration to boundary perception, an analysis of misplaced boundaries is provided. Identified patterns concern especially the relationship between tonal structure and boundary salience, the order of preference between intonation and duration, and the tendency towards perceptual filling of accent lapses.

Index terms: prosodic boundaries, Czech, delexicalization

1. INTRODUCTION

Prosodic boundaries—points of prosodic discontinuity occurring in connected speech—have undeniable relevance in lexical access, syntactic parsing and semantic analysis of speech [1, 2], and it is with respect to this fact that they are systematically realized in speech production. Boundary-related prosodic phenomena may be classified into three groups:

(i) pause as the most clear-cut boundary signal;

(ii) phenomena occurring near the boundary (tones/ contours, pitch accents, f_0 reset, final lengthening, final intensity decrease, initial strengthening);

(iii) coherence of prosodic units, which indicates the boundary in an indirect way.

The language examined in this paper is Czech, a highly inflected Western Slavonic language, spoken by approximately 12 million people. Sticking to the classification of boundary-related phenomena above, we may summarize the knowledge available for this language (mostly reposing on the study of standard read speech) as follows:

(i) As in other languages, pause is chiefly used to separate intonation units, and it may accompany hesitation or focus.

(ii) Czech has fixed stress, located on the first syllable of stressable words, which should guarantee a clear boundary signal. However, it has been shown that stress has no reliable phonotactic correlates and, due to phonological vowel length distinction, is not accompanied by lengthening [3], nor does it correspond to an intensity peak [4]. On the intonation level, the stressed syllable frequently has a low tone, followed by a rise on the next syllable. In an experiment where a neural network had to localize accents without any lexical information, relying solely on prosodic parameters, the correct identification score was around 80% [5]. As far as boundary tones are concerned, the Czech prosodic tradition includes them in the study of intonation contours, especially the nuclear ones (e. g. [3]). Evidence on f_0 resetting has been provided—though indirectly—in [6]. A certain degree of stress-unit-final lengthening has been described, as well as final intensity decrease. Potential secondary accents on odd syllables of the stress unit may be observed sometimes [3].

(iii) The coherence-oriented view of Czech stress units, proposed in [7] and developed in [8], is a possible reply to the no-prominence problem mentioned above. The category of coherence also includes intonation downtrends, investigated for Czech in [6].

These facts indicate that the need for the study of prosodic boundaries is especially felt on stress unit level, where it is correlated with the problem of perceived accents.

2. THE EXPERIMENT

The objective of the present experiment is to determine the role of prosody in speech segmentation, namely the respective contribution of intonation and duration to the perceptual existence of boundaries between successive stress units, by means of delexicalized stimuli.

Ten short declarative sentences (with 12.2 syllables on average), read by a professional speaker and recorded in a soundproof booth, were delexicalized and submitted to listeners for word segmentation. Delexicalization, especially if more sophisticated than simple spectral filtering, is a useful method of phonetic investigation [9, 10]; for Czech, the first use of this method dates back to the 1960s [11].

In our experiment, the delexicalization is roughly based on the method described in [10], but uses modified (resynthesized) natural speech instead of TTS synthesis, the latter still implying a considerable quality loss. The sentences were modified in three different ways:

(i) The *saltanaj* modification:
The *saltanaj* delexicalization [10] consists in replacing each segment by its generic representative. In the case of Czech, the recipe is: all vowels become [a]; plosives and affricates [t]; nasals [n]; fricatives (including vibrant fricatives) [s]; laterals and trills [l]; non-lateral approximants [j]. Example:

Praha byla obklopena vinicemi.
[ˈpraɦa ˈbɪla ˈʔɔpklɔpɛna ˈvɪɲɪt͡sɛmɪ]
‘Prague was surrounded by vineyards.’
tlasa tala tattlatana sanatana.
[ˈtlasa ˈtala ˈtatːlatana ˈsanatana]

In the recording phase, the speaker pronounced, after each original sentence, the corresponding *saltanaj* version, with the instruction to stick to the original prosody as much as possible. However, since imitation may introduce different artefacts resulting from the non-standard speech situation, each raw *saltanaj* version was resynthesized in the Praat program

(PSOLA) and modified to match the intonation and the duration of the original segments as closely as possible. This manipulation was carried out as follows:

– Within each segment, f_0 was measured at five equidistant points in the original sentence and transplanted to the corresponding target points in the *saltanaj* sentence (since five-point f_0 sampling includes a certain degree of smoothing, microintonation was disregarded in the manipulation; moreover, microintonation, even if transplanted, should be largely neutral to the object of the study).

– As far as duration is concerned, applying the values of the original sentence onto the *saltanaj* version is not possible because the matched sounds have different intrinsic durations. For this reason, the duration of each segment was normalized by means of the following formula (example of a generic [a] replacing an underlying [ɛ]):

$$dur_{norm}\ of\ a\ given\ generic\,[\mathrm{a}] = dur_{aver}\ of\ all\,[\mathrm{a}]\frac{dur_{raw}\ of\ the\ underlying\,[\varepsilon]}{dur_{aver}\ of\ all\,[\varepsilon]}$$

where:

dur_{norm}—normalized duration [ms]
dur_{raw}—raw duration [ms]
dur_{aver}—average duration [ms]

For the sake of statistical representativeness, average durations were measured in a larger text (more than 3000 segments) read by the same speaker. For 12 segments (11 consonants and 1 vowel), the duration prediction was perceptually inadequate, and the value had to be adapted *ad hoc*. The described method introduces several minor artefacts, which—unfortunately—cannot be easily eliminated, namely the appearance of geminates, which occur only across word or morpheme boundaries in Czech, and the emergence of structures which resemble real words (e. g. the syllable *ta* [ta] which is identical to the demonstrative pronoun 'that') – see section 3.1.

(ii) The *mamama* modification:

The *mamama* delexicalization (roughly corresponding to the *sasasa* modification proposed in [10], but allowing no consonant clusters; the nasal was preferred for its voicing and perhaps more pleasant acoustic effect, cf. [9, 12]), consists in replacing all syllables by a generic *ma* syllable with transplanted intonation and uniform duration (average duration of the [m] and [a] segments in a large speech sample recorded by the speaker, both lengthened by 7% to guarantee a speech rate comparable to the one of *saltanaj*, which has more complex syllables). The recording procedure and the f_0 transfer were similar to *saltanaj;* the matching rules for f_0 were:

– The contour of the target nucleus [a] was based on the contour of the voiced part of the source rhyme.

– The contour of the target onset [m] was based on the contour of the voiced part of the source onset; if there is no voicing in the source onset, then the values were interpolated.

An example of this transformation is:

Praha byla obklopena vinicemi.
[ˈpraɦa ˈbɪla ˈʔɔpklɔpɛna ˈvɪɲɪtsɛmɪ]
mama mama mamamama mamamama.
[ˈmama ˈmama ˈmamamama ˈmamamama]

(iii) The *flat saltanaj* modification:

The *flat saltanaj* modification is identical to the *saltanaj* version except for the fact that f_0 is set to 80 Hz (average f_0 of the speaker in the recording).

To sum up, the three delexicalized versions are:

– *saltanaj:* preserves syllabic structure, broad phonotactics, segment duration and intonation; neutralizes narrow segmental structure

– *mamama:* preserves syllabic structure and intonation; neutralizes narrow segmental structure, phonotactics and segment duration

– *flat saltanaj:* preserves syllabic structure, broad phonotactics and segment duration; neutralizes narrow segmental structure and intonation

Since all versions are based on natural imitated speech, they also contain dynamic variability, which is not controlled in this experiment. All delexicalized sentences were perceptually adequate to the originals.

The participants in the perceptual evaluation were 12 late pre-gradual or post-gradual students in linguistics (average age 27 years); a certain experience in listening and auditory analysis was a prerequisite given the difficulty of the task. These listeners were presented with the 30 items (10 different sentences, each modified in three ways), arranged in three blocks: (i) *flat saltanaj*; (ii) *mamama*; (iii) *saltanaj*. The order of the blocks corresponded to the increasing amount of available prosodic information; within each block, the sentences were ordered randomly. One item in each block was repeated to ascertain the listeners' consistency. The test was presented in an interactive application offering the possibility of repeated listening. A paper form was provided for evaluation, with the following instruction: 'You are going to hear 33 Czech sentences, modified in a way to be unintelligible. Despite that, you are requested to mark word boundaries, where you feel them, by a vertical stroke between syllables. After that, assign one of the numbers 1, 2 or 3 to each boundary, where 1 is the most distinct one (use this number no more and no less than once in each sentence), 2 is a medium boundary, and 3 is a minor boundary.' The test started with a short training task. In the form, the sentences were transcribed as a sequence of syllables. The answers may have appeared like this:

$$\mathit{tla\ \ sa} \overset{2}{\Big|} \mathit{ta\ \ la} \overset{2}{\Big|} \mathit{tat\ \ tla\ \ ta\ \ na} \overset{1}{\Big|} \mathit{sa\ \ na\ \ ta\ \ na.}$$

By asking to mark *word boundaries*, we investigated in fact the *prosodic word* boundaries, because the listeners should have no means of recognizing boundaries of unstressed words. The *real boundary location* used in the evaluation is then the location of *prosodic word* boundary, as assigned by auditory inspection. In the present paper, we only discuss the position of the boundaries obtained; their respective weight will be the object of a future analysis. Only sentence-interior boundaries are considered. Tonal transcription was based on the INTSINT system [13]; for the symbols, see legend of Table 1.

3. RESULTS AND DISCUSSION

3.1 Significance evaluation

We assume that locating prosodic boundaries requires good orientation both in the heard and written sentence. This capacity may decrease with the distance from the sentence beginning. Therefore, we measured the mean number of boundaries present at correct positions separately for the 1st—6th syllable, and for the 7th—last syllable in each sentence, to evaluate the difficulty of the task. These scores are 0.67 and 0.54 respectively, which is a significant difference (t test: p = 0.020). The difficulty seems to be evident, but the score in the second part of each sentence is by no means a reflexion of random choice. For the within-speaker consistency evaluation, we used the repeated items. In each pair of identical sentences, we marked the presence of a boundary in the correct position with 1, and its absence by 0. The mean difference between corresponding positions was 13.4%, which is—judged by the naked eye—a fairly good result. We also evaluated each syllable for its probability to be a legal word onset in Czech (which might influence the respondents' judgments): this factor turned out to be non-significant.

3.2 Global parameters of the analysis

On average, listeners assigned more boundaries than there actually were in the sentences. The ratio *average number of boundaries assigned/real number of boundaries* is: *flat_saltanaj* 1.29, *mamama* 1.26, *saltanaj* 1.39. This suggests that both richer phonotactics and richer prosody may increase the listeners' tendency towards a finer segmentation.

To evaluate the matching between identified boundaries and real boundaries, we made use of two indicators: (i) ratio *number of correctly assigned boundaries/number of real boundaries,* and (ii) ratio *number of incorrectly assigned boundaries/number of real boundaries*—cf. Figure 1.

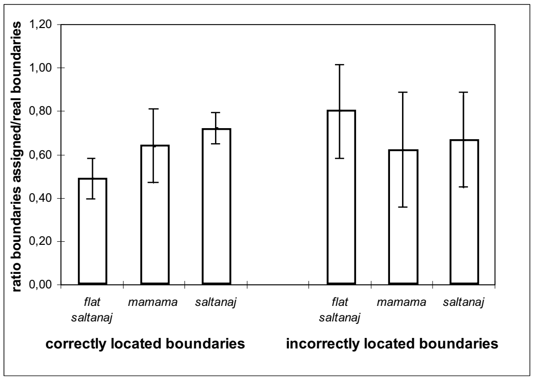

Figure 1: Correctly and incorrectly assigned boundaries in the three listening situations (mean and standard deviation). Significance evaluation (t tests):
(i) correctly assigned boundaries:
– difference *flat saltanaj—mamama*: significant (p = 0.039)
– difference *mamama—saltanaj*: non-significant (p = 0.063)
– difference *flat saltanaj—saltanaj*: significant (p < 0.001)
(ii) incorrectly assigned boundaries:
– difference *flat saltanaj—mamama*: significant (p = 0.022)
– difference *mamama—saltanaj*: non-significant (p = 0.478)
– difference *flat saltanaj—saltanaj*: significant (p = 0.011)

The results suggest that:

i) *Flat saltanaj*, showing no intonational behaviour, is the most difficult situation in boundary detection for the listeners. The number of correctly assigned boundaries is significantly lower and the number of incorrectly assigned boundaries significantly higher than in either of the remaining situations.

ii) The scores achieved in the *mamama* and *saltanaj* sentences are not significantly different. Judging by the mean, *saltanaj* shows better results in correctly assigned boundaries, and slightly worse results in incorrectly assigned boundaries than *mamama*. However, we assume that correctly located boundaries are a more reliable measure because of their generally lower standard deviation: therefore, we may say that the addition of duration and broad phonotactics has a slightly positive impact on boundary identification. The percentage of incorrectly located boundaries seems to be very high, but this is partly compensated by the fact that there are much more syllable pairs not separated by a stress unit boundary than pairs separated by such a boundary.

3.3 Analysis of boundaries with high inter-subject agreement in *saltanaj*

Describing relations between the perceptual facts and the prosodic structure of the tested sentences necessarily implies reducing the amount of data considered. We decided to exclude all boundaries that had not been identified by more than half of the listeners, and to go on with the most salient boundaries only (for an example, see Table 1). In this section, the analysis is limited to the *saltanaj* version. In some cases, we cannot avoid making conclusions from relatively small numbers of observations; it should be noted, however, that their significance is considerably increased by the low-agreement cut-off.

Considering boundaries that are perceived in natural speech, the first question is what prosodic factors distinguish those which have been identified in *saltanaj* speech from those which have not. Tonal structure does not seem to provide any clear answer, except for the fact that when the boundary is preceded by a T tone, its identification is likely to be very good (note that this fact reduces the difference in correct identification score between the first and second part of each sentence—see section 3.1). In the duration domain, pre-boundary lengthening and shortening have about the same distribution in correctly identified and omitted boundaries. The only regularity is that within correctly identified boundaries, those which have pre-boundary lengthening had a 95% agreement score (cf. syllable [lɔs] in Table 1), whereas those with pre-boundary shortening had a 81% score.

Considering boundaries added in wrong positions, their tonal structure is mostly less pronounced, but not necessarily contrary to the one of legal boundaries: while all boundaries assigned in natural speech, except for one, have a falling tonal pattern (HL, TL, ...L, HD, TD, ...D, where ... stands for absence of tone), 8 out of 12 wrongly located boundaries have a non-specified second tone. In 4 cases out of 12, none of the adjacent syllables is tonally specified, but there is pre-boundary lengthening.

To verify the presence of trochaic secondary accents, we tried to find out if there is a preference to place boundaries before these syllables. The count (6 even-placed and 7 odd-placed boundaries) does not indicate trochaic behaviour of the studied stress units.

Finally, we verified if there was a tendency towards filling longer accent lapses (cf. syllable [tʃɔ] in Table 1). We counted the number of wrongly added boundaries separately for stress units shorter than 4 syllables and longer than 3 syllables. The percentage is 17% and 56% respectively, but the difference is

Table 1: Example of boundary assignment, displaying only boundaries assigned by more than half of the listeners. INTSINT tonal symbols [13]: H—higher, L—lower, T—top, B—bottom, D—downstepped. The relative duration coefficient corresponds to the average ratio *duration of the segments within the syllable/their average duration in a large sample with almost identical speaking rate*; values < 1 correspond to shortening, values > 1 to lengthening. In the 'Judgements' lines, each value corresponds to the percentage of listeners who felt the boundary after the syllable in question. The syllable preceding real boundary location is marked by a thick frame.

V minulosti o něj nepečovali. 'In the past, they didn't take care of it.'	Analysis	intonation				T	L	H	D			D	B
		rel. duration	1.20	0.92	1.07	0.77	0.92	1.11	0.85	0.97	1.00	1.16	1.28
		syllables	ˈvmɪ	nʊ	lɔs	cɪ	ˈʔɔ	ɲɛj	ˈnɛ	pɛ	tʃɔ	va	lɪ
	Judgements	*flat saltanaj*			1.00			0.92			0.92		
		mamama				0.67		0.75					
		saltanaj			0.58	0.75		1.00			0.83		

partly due to the simple fact that longer units offer more space for misplaced boundaries; however, after normalizing by the number of syllables, the percentage is still 7% vs. 13%.

3.4 Analysis of boundaries with high inter-subject agreement in *mamama* and *flat saltanaj*

Generally, the relative number of high-agreement boundaries in *saltanaj*, *mamama* and *flat saltanaj* follows the proportions given in Figure 1. The differences in boundary assignment between *saltanaj* and *mamama* show no regular patterns. As it could be expected, incorrectly added boundaries in *saltanaj* which were triggered by pre-boundary lengthening are often replicated in *flat saltanaj* (4 cases out of 6), and are mostly missing in *mamama* (5 cases out of 6).

4. CONCLUSION

The present experiment has shown that in delexicalized speech where all prosodic parameters are preserved (*saltanaj*), boundaries between stress units can be identified in 72% of the cases (note that this percentage does not take into account misplaced boundaries). Supposing that there is no error introduced by the method, the remaining 28% would be imputable to lexical structure. When removing temporal variability, the percentage decreases, though not significantly, and when removing intonation, it goes down (significantly) to 49%.

Intonation is confirmed as being a better boundary predictor than duration. In the original (natural) sentences, all interior boundaries were within a transition from a higher to a lower tone, but despite this apparently strong marker, in delexicalized speech, listeners often felt the boundary in a different position, tonally less specified. Final lengthening seems to be relied on when there is no tonal specification. The natural tendency towards perceptual filling of accent lapses when there are not enough segmentation indices has been confirmed. On the other hand, no tendency towards trochaicity has been observed.

The present experiment completes available knowledge on Czech prosody, using a non-trivial and relatively precise delexicalization method, which could be successfully applied in other experiments as well.

Two major limits of the experiment are: (i) the size of the speech sample; (ii) the uncontrolled contribution of intensity whose variations were inherited from the imitated sentences. Also, a certain degree of error might have been introduced by the tonal transcription method and by temporal normalization.

5. ACKNOWLEDGEMENT

This research was supported by the GAAV B9101402 grant.

6. REFERENCES

[1] Cutler, A., "Exploiting prosodic probabilities in speech segmentation", *Cognitive Models of Speech Processing,* G. T. M. Altman (ed.), Cambridge MA, MIT Press, 1990.

[2] Grosjean, F. and Dommergues, J.-Y., "Les structures de performance en psycholinguistique", *L'année psychologique, 83*, 513—536, 1983.

[3] Palková, Z., *Fonetika a fonologie češtiny*, Prague, Karolinum, 1994.

[4] Duběda, T., "Intensity as a macroprosodic variable in Czech", *Speech Prosody 2006*, Dresden (CD-ROM Proceedings).

[5] Duběda, T. and Votrubec, J., "Acoustic analysis of Czech Stress. Intonation, duration and intensity revisited", *Interspeech/Eurospeech*, Lisabon, 1429—1432, 2005.

[6] Volín, J. , "F0 declination in Czech and English breath-groups", *Phonetica Pragensia X*, Palková, Z. and Veroňková, J., eds., 125—136, 2005.

[7] Palková, Z., "Einige Beziehungen zwischen prosodischen Merkmalen im Tschechischen", *XIVth Congress of Linguists*, Vol. I., Berlin, 507—510, 1987.

[8] Palková, Z. and Volín, J., "The role of f0 contours in determining foot boundaries in Czech", *15th ISPhS*, Barcelona, 1783—1786, 2003.

[9] Nooteboom, S., "The Prosody of Speech. Melody and Rhythm", *The Handbook of Phonetic Sciences*, Hardcastle, W. J. and Laver, J., eds., Blackwell, 640—673, 1997.

[10] Ramus, F. and Mehler, J., "Language identification with suprasegmental cues: A study based on speech resynthesis", *JIPA, 105*, 1999, 512—521, 1999.

[11] Janota, P., "Perception of stress by Czech listeners", *Proceedings of the 6th ICPhS*, Prague, 457—461, 1967.

[12] Rilliard, A. and Aubergé, V., "Prosody diagnostic using reiterant speech", *14th ICPhS*, 37—40, 1999.

[13] Hirst, D. and Di Cristo, A., eds., Intonation Systems. A Survey of Twenty Languages, CUP, 1998.