THE EFFECT OF DURATIONAL CUES ON THE REASSIGNMENT OF A SYLLABLE IN THE METRICAL STRUCTURE OF CZECH SENTENCES

Radek Skarnitzl & Jan Volín [1]

Institute of Phonetics, Faculty of Arts, Charles University, Prague
radek.skarnitzl@ff.cuni.cz, jan.volin@ff.cuni.cz

ABSTRACT

Examinations of the acoustic correlates of lexical stress in Czech revealed final lengthening operating at the level of individual words: a word-final (non-nuclear and phrase-internal) vowel was shown to be significantly longer than word-internal vowels. This study aimed to find whether listeners are sensitive to these temporal details: in ambiguous word–pseudoword sequences read by four speakers, the duration of the first vowel of the pseudoword was lengthened to see whether this would result in the reassignment of the given syllable as the final syllable of the preceding word. A perception test was individually administered to 45 respondents. The results show that listeners tend to reassign a syllable even in some sentences without temporal changes, although less frequently. Moreover, the perception strongly depends on the particular sentence, with only some behaving according to expectations. The results are discussed especially with respect to cue weighting in speech perception.

Keywords: lexical stress, temporal structure, speech perception, cue weighting, Czech

1. INTRODUCTION

Czech is a language in which lexical stress is fixed to the first syllable of a prosodic word (stress group). In such languages, the function of stress is not nested in an isolated word but in a prosodic phrase, where it provides for correct parsing of the whole unit. The acoustic correlates of lexical stress have recently been examined by [10] (*cf.* also [11]). The results of the study confirmed earlier methodologically more scattered findings, namely that the stressed syllable in Czech does not manifest any of the typical signs of prominence, in the sense of positive deviations of acoustic parameters: vowels in stressed syllables are not marked by higher sound pressure level, by higher or more variable F0, longer duration, shallower spectral slope, nor more peripheral vocalic articulation, as reflected in formant values.

Earlier studies [5], [9] suggested that word stress in Czech is not determined by the acoustic prominence of the stressed syllable, but rather by the prosodic configuration and cohesion of the entire stress group and by prosodic discontinuities between neighbouring stress groups. More specifically, Czech stress groups seem to be characterized by a post-stress F0 rise (L*+H), with the second syllable in a stress group typically lying higher than the stressed one [7]. More specifically, three-syllabic and longer stress groups produced by Czech Radio newsreaders were found to display a rising-falling pattern of F0 [12], and this post-stress rise seems to be characteristic of most speech styles in Czech.

While vowels in stressed syllables were not found to be more prominent than those in unstressed syllables in the traditional treatment of prominence [10], the study uncovered a previously unknown tendency for vowels in the last syllable of a prosodic word to be significantly longer than in the preceding syllables. It must be pointed out that only phrase-internal and non-nuclear stress prosodic words were analyzed, so that this result is not affected by the well-known phrase-final lengthening (or deceleration), or by higher-level prosodic prominence (see the review in [4]). In other words, the study found lengthening at the level of individual words, in phrase reading as well as in spontaneous speech, and the extent of this lengthening was, on average, 30%.

Such phrase-internal word-final lengthening has been documented before in carefully controlled speech involving minimal pairs: for instance, Beckman and Edwards [1] found the *schwa* in the first word of sentence a. below longer than that at the beginning of sentence b.

a. *Poppa posed the question strongly...*
b. *Pop opposed the question strongly...*

Similarly, Cutler and Butterfield [3] compared vowel durations in target pairs like *inquires* vs. *in choirs* or *lettuce* vs. *let us* and found a small effect of duration. In [13], the authors carried out quite an extensive study on news bulletins read out by professional news readers with an objective to examine polysyllabic shortening in words, interstress intervals and word rhymes. They did not confirm the existence of independent polysyllabic shortening in either of the units, but they found evidence of final lengthening in all three domains. (Phrase-initial and phrase-final units were excluded from their analyses.) To the best of our knowledge, phrase-internal word-final lengthening as in [10] has not been identified in less controlled or even spontaneous speech before.

The magnitude of the observed word-level lengthening effect found in [10] was relatively large, especially for a language like Czech in which vowel quantity is phonologically distinctive. It is thus conceivable that word-final lengthening within prosodic phrases may function as a perceptual boundary cue. This hypothesis was indirectly supported by a study of synthetic Czech speech using unit-selection [6]. The authors found that placing word-final but phrase-internal vowels from the source database (or rather, diphones pertaining to those word-final vowels) into non-word-final contexts significantly deteriorated the subjective quality of the synthesized speech, presumably due to the disruption of local timing relations.

Since F0 configurations throughout prosodic words and F0 discontinuities appear to aid the segmentation into stress groups [9], [12], and since duration represents a clear acoustic discontinuity between neighbouring stress groups [10], the objective of this study is to examine the role of duration in cuing stress group boundaries. In other words, we are interested in seeing whether listeners will be sensitive to temporal manipulations to such an extent that an ambiguous syllable will be reassigned from one stress group to another.

2. METHOD

2.1. Material

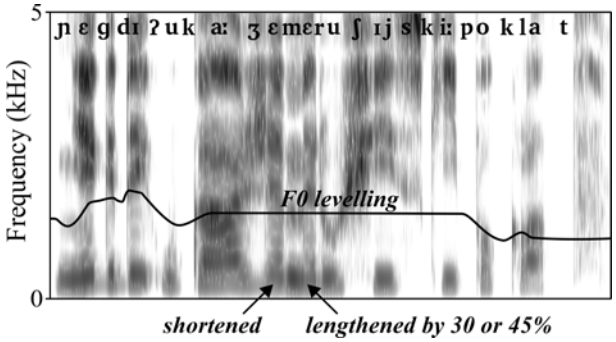
Four native speakers of Czech (two female, two male) read 24 four-word phrases as in (1) below, where the second word has a different lexical or grammatical meaning when the first syllable of the third word is placed at the end as in (2), and the third unit is a pseudoword in both sentences but is always phonotactically valid in Czech. The speakers also read version (2) for the purpose of comparison (see below).

(1)	Někdy	ukáže	merušijský	poklad.
	/ ˈɲɛgdɪ	ˈʔukaːʒɛ	ˈmɛruʃɪjskiː	ˈpoklat /
	Sometimes	*[s/he] shows*	*“merushiy”*	*treasure.*
(2)	Někdy	ukážeme	rušijský	poklad.
	/ ˈɲɛgdɪ	ˈʔukaːʒɛmɛ	ˈruʃɪjskiː	ˈpoklat /
	Sometimes	*[we] show*	*“rushiy”*	*treasure.*

The manipulations were performed using PSOLA [8] as implemented in Praat [2]. To eliminate the effect of melodic changes, we first monotonized F0 from the second syllable of the second word (in this way, we introduced the characteristic post-stress rise in F0 mentioned above) until the end of the third word. This monotonized-only version, henceforth labelled stimulus A, served as reference; it was expected to yield interpretations corresponding to the original reading (1). Subsequently, the duration in the first vowel of the third word (1) was lengthened in two steps, by 30 and 45%; these are henceforth labelled stimulus B and C, respectively. As mentioned above, 30% was the mean duration difference between the final and non-final vowels [10]; however, it was desirable to include a larger lengthening to be able to compare the responses of the listeners in stimuli which, upon informal listening, clearly triggered the reassignment of the target syllable.

Performing only the lengthening resulted in two consecutive “lengthened” syllables: one naturally ([ʒɛ] in the example above) and one artificially ([mɛ]), which considerably disrupted the local timing. That is why the vowel which was originally word-final ([ɛ] in the word *ukáže* above) had to be shortened. The degree of shortening was determined from the corresponding version (2) of the given sentence in which the respective vowels appeared word-internally (*ukážeme*). The entire procedure is schematically shown in Figure 1.

Figure 1: Schematic representation of the manipulations performed on sentence (1) to induce perception corresponding to sentence (2); see text.



The resulting phrases were used to compile a 2AFC perception test whose aim was to determine whether the implemented duration changes would induce a shift in interpretation from (1) to (2), and whether the lengthening of 30 and 45% would yield significantly different responses.

2.2. Perception test

A pilot test showed that the repetition of identical phrases with different degrees of lengthening, albeit pronounced by different speakers, was not engaging enough for the listeners to produce significant results. They seemed to be unable to change their interpretation of a given phrase once they formed an opinion on the wording. That is why we redesigned the experiment so that each phrase appeared only twice, always read once by a male and once by a female speaker, yielding the total of 48 stimuli. Of these, 10 were type A stimuli (i.e., only with F0 monotonized) and 19 were type B and type C stimuli each.

Four different versions of the perception test were created. Since it was crucial that the same underlying phrases (pronounced by a male and a female speaker) appeared as far from each other as possible, the stimuli were presented in a fixed order within each test version. Care was taken that phrases similar in terms of their syntactic and/or morphological structure did not appear close to each other.

In order to further reduce the possibility of listeners fixating on one of the two interpretations, the entire test was preceded by a brief exposure (40 seconds) to all the 48 words in the form of a list on two subsequent PowerPoint screens. The words were ordered alphabetically and counterbalanced with the experiment in that the first slide always showed the version that would appear second in the first half of the listening experiment and vice versa.

The perception test was administered using the ExperimentMFC tool in Praat to 45 respondents, all students of philological majors at the university in Prague (38 females, 7 males). The two possible interpretations of the phrase (with their order counterbalanced across the two appearances within one test, as well as across the four test versions) were shown on the screen. After a 2500-ms silence, during which they could read the interpretations, the listeners were instructed to click on the phrase which they regarded as a more probable version of what the speaker was reading. Listeners were allowed to replay each stimulus three times. Three trial items, which used different speakers, preceded the test itself; a short break was included after every eight items to reduce fatigue.

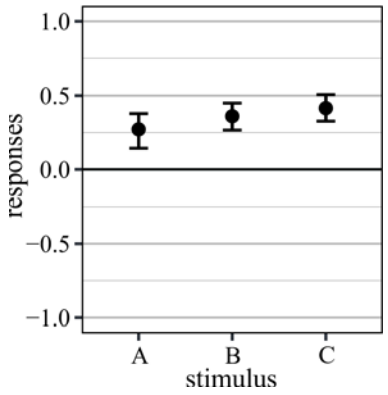
2.3. Analysis

The listeners' responses were associated with values as follows: 1 corresponds to version (2) above (i.e., with the initial syllable of the pseudoword perceived as the final syllable of the preceding word); –1 corresponds to version (1) above (perception without syllable reassignment). For each group we calculated the mean value and estimated confidence intervals using the bootstrap method with a significance level of 0.05 (Bonferroni-corrected for multiple testing). This means that a null hypothesis of no significant difference between the perception of versions (1) and (2) cannot be rejected if the confidence interval in the charts includes the value of 0.

3. RESULTS

It is obvious already from the global results shown in Figure 2 that the hypothesis has not been confirmed unequivocally: the first syllable of the pseudoword has been perceptually reassigned as the last syllable of the preceding word significantly more often than not in all three types of stimuli. In other words, these results indicate that even in type A stimuli, where no lengthening was introduced, the mere monotonization of F0 was sufficient to induce a change in perception from version (1) (e.g., [ˈʔukaːʒɛ ˈmɛruʃɪjskiː]) to version (2) (e.g., [ˈʔukaːʒɛmɛ ˈruʃɪjskiː]). This perceptual reassignment is only slightly more salient in stimuli of the B and C type, which involve 30% and 45% lengthening of the pseudoword's first syllable nucleus, respectively.

Figure 2: Responses to stimuli A (F0 levelling only), B (30% lengthening), and C (45% lengthening); see text for more detail.

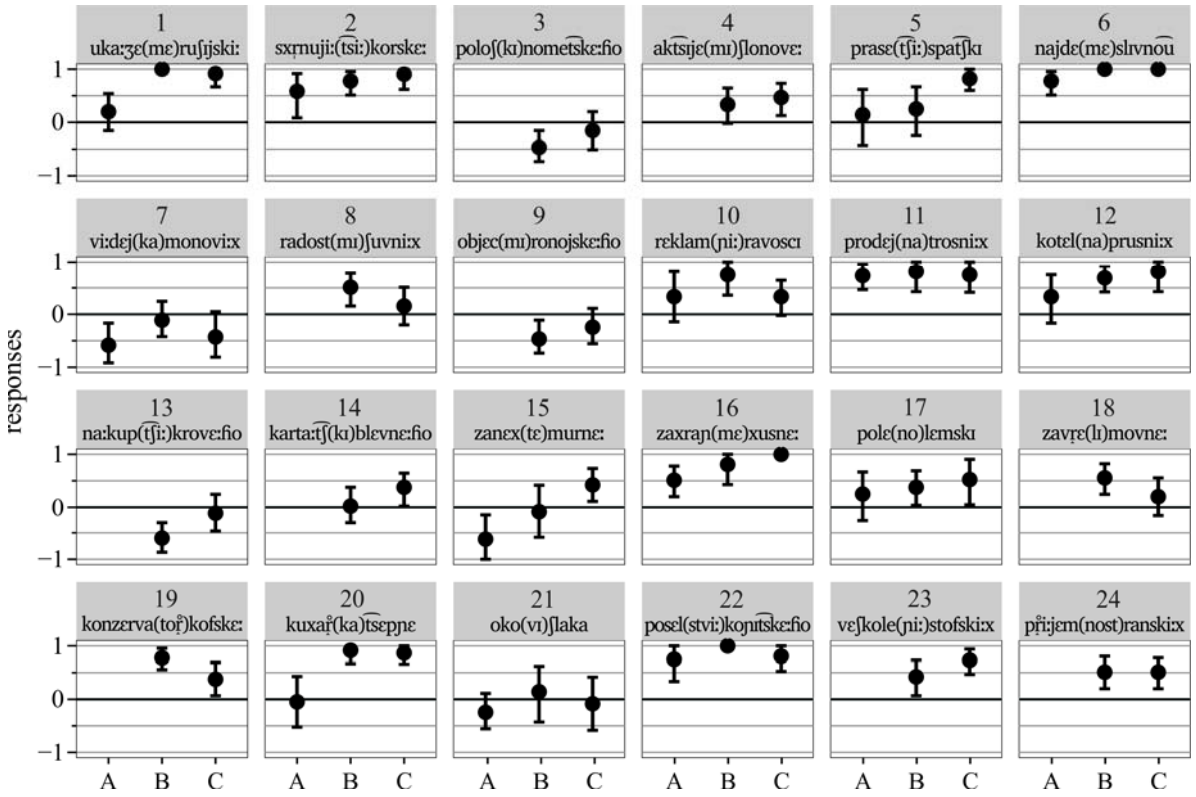


It should be clear, however, that the global results conceal a lot of variability. In Figure 3 overleaf, we plot the responses for all 24 phrases (note that not all stimuli of type A are represented; *cf.* section 2.2). The results show one phrase which may be regarded as exemplary from the perspective of our original hypothesis: in phrase 15, stimulus A with only F0 levelling was perceived by the respondents as corresponding to the original reading [ˈzanɛx ˈtɛmurnɛː], stimulus C with the largest lengthening of the vowel in syllable [tɛ] induced the most significant response in favour of the opposite reading, [ˈzanɛxtɛ ˈmurnɛː], with stimulus B evaluated as not significantly in favour of either reading.

There are several other phrases which deviate from this pattern only to a small degree. In phrases 1, 12, 17 and 20, stimuli B and C were perceived according to the expectations, with the target syllable perceptually reassigned significantly more often than not. In stimuli A, there was no statistically significant difference between the responses in favour of either of the two readings. In phrase 5, it is also stimulus B, with the smaller degree of lengthening, where the responses did not significantly differ.

Four phrases – numbers 2, 6, 11, and 16 in Figure 3 – manifest a rising or at least level tendency towards the reassigned interpretation with progressive lengthening but the tendency to reassignment is, in fact, significant already in stimuli A which involve no lengthening.

Figure 3: Responses to stimuli A (only F0 monotonization), B (30% lengthening), and C (45% lengthening) in the 24 phrases. The target syllable is shown in brackets.



The responses to the remaining phrases where all three stimuli (A, B, C) are represented (i.e., 7, 10, 21, 22), but also to many of those with stimuli B and C only, are more difficult to explain. In some of these, the 30% lengthening of stimulus B induced a greater change in the direction of the reassigned interpretation than the 45% lengthening of stimulus C. In others, none of the manipulations induced the expected significant shift in perception.

4. DISCUSSION

The results of the perception test do not speak unanimously for or against the hypothesis of word-final lengthening as a cue of the word boundary. Apparently, the phenomenon is not robust enough to outweigh the influence of other cues which contribute to the correct word boundary detection by the listener. Our test items provide competitors of different strength. The strongest phonetic cue is quite probably the melodic contour, and depriving the listener of melodic cues seems to have resulted in a much more ambiguous segmentation than hypothesized.

Moreover, by monotonizing F0 we actually did not monotonize pitch; the perceptual impression was likely not one of melodically monotonous sequence of syllables where vowels of different height met around the target syllable. However, controlling for vowel height would have resulted in highly unnatural phrases, i.e., phrases with perhaps statistically stronger results but without ecological validity.

The levelling of F0 may also have strengthened the effect of other, especially temporal cues which are presumably not as important for segmentation in normal speech. For instance, it is conceivable that [ʃ] in the pseudoword *vyšlaka* (phrase 21) will be shorter than in *šlaka* (but *cf.* [13]), or that the closure duration in [k] at the beginning of the pseudoword *kamonových* (phrase 7) will be longer than within the word *výdejka* where it only marks an ordinary grammatical suffix.

One last, potentially strong competitor to the word-final lengthening, might be the level of activation of the target words due to their frequency of occurrence in the listeners' cumulative input. For instance, one can imagine that *výdej* and *výdejka* (again phrase 7) will not be perceived as equally likely by individual respondents.

Be that as it may, the overall trend that emerged from 2160 judgements produced by our 45 listeners suggests that word-final lengthening is not irrelevant to word boundary perception in Czech.

[1] This study was supported from the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734) and by CUNI project Progres 4, "Language in the shiftings of time, space, and culture".

7. REFERENCES

[1] Beckman, M. E. & Edwards, J. 1990. Lengthenings, shortenings, and the nature of prosodic constituency. In: Kingston, J., Beckman, M. E. (eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge: Cambridge University Press, 152–178.

[2] Boersma, P., Weenink, D. 2018. Praat: Doing phonetics by computer, version 6.0.43. Computer program, retrieved from www.praat.org.

[3] Cutler, A., Butterfield, S. 1990. Syllabic lengthening as a word boundary cue. In: Seidl, R. (ed.), *Proc. 3rd Australian International Conference on Speech Science and Technology* Canberra, 324–328.

[4] Fletcher, J. 2010. The prosody of speech: Timing and rhythm. In: Hardcastle, W. J., Laver, J., Gibbon, F. E. (eds.), *The Handbook of Phonetic Sciences*. Oxford: Blackwell Publishing, 523–602.

[5] Janota, P., Palková, Z. 1974. The auditory evaluation of stress under the influence of context. *Acta Universitatis Carolinae – Philologica, Phonetica Pragensia* IV, 29–59.

[6] Jůzová, M., Tihelka, D., Skarnitzl, R. 2017. Last syllable unit penalization in unit selection TTS. In: Ekštein, K., Matoušek, V. (eds.), *Proc. 20th International Conference on Text, Speech and Dialogue* Cham, 317–325.

[7] Jůzová, M., Volín, J. 2018. F0 post-stress rise trends consideration in unit selection TTS. In: Sojka, P., Horák, A., Kopeček, V., Pala, K. (eds.), *Proc. 21st International Conference on Text, Speech and Dialogue* Cham, pp. 360-368.

[8] Moulines, E., Charpentier, F. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Comm.* 9, 453–467.

[9] Palková, Z., Volín, J. 2003. The role of F0 contours in determining foot boundaries in Czech. *Proc. 15th ICPhS* Barcelona, 1783–1786.

[10] Skarnitzl, R. 2018. Phonetic realization of lexical stress in longer words in Czech. *Slovo a slovesnost* 79, 199–216. [in Czech]

[11] Skarnitzl, R., Eriksson, A. 2017. The acoustics of word stress in Czech as a function of speaking style. *Proc. Interspeech* Stockholm, 3221–3225.

[12] Volín, J. 2008. On intonation in newsreading: the pitch of the first syllable of a prosodic word. *Čeština doma a ve světě* 1-2/2008, 89–96. [in Czech]

[13] Windmann, A., Šimko, J., Wagner, P. 2015. Polysyllabic shortening and word-final lengthening in English. *Proc. INTERSPEECH Dresden, 36–40.*