INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Perceptual evaluation of early versus late F0 peaks in the intonation structure of Czech question-word questions

Pavel Šturm, Jan Volín

Institute of Phonetics, Charles University, Prague, Czech Republic
pavel.sturm@ff.cuni.cz, jan.volin@ff.cuni.cz

Abstract

Question-word questions in Czech lexically mark their interrogative function in the initial position: in their standard form, they begin with an interrogative lexeme. For many linguists, this is a sufficient reason for resigning on intonation marking, so they claim that the speech melody in these questions is identical to the melody of statements. A careful observation of the current Czech speech suggests otherwise.

This paper presents a perceptual experiment in which Czech speakers evaluated two contrastive forms of the interrogative melody, specifically the one with a late peak modelled after statements (as suggested by some authors), and the one with an early peak modelled after our empirical data collected previously. Thirty-two listeners expressed a statistically significant preference for the early peak in a perception test. This outcome resonates with the sample of speech production of the questions. However, the late peak is also possible and acceptable: we assume that it might be a signal of contrastive emphasis or an implicational cue.

Index Terms: Czech, intonation, perception, wh-questions, speech melody, nucleus placement

1. Introduction

Grammarians identify various classes of interrogative sentences, but two major types are recognized quasi-universally: polar (yes-no) questions and specific (question-word) questions. In English, the latter are also known as *wh-questions* since the spelling of interrogative lexemes like *when*, *where*, *who*, *why* or *what* starts with the letters 'wh'. In Czech, the language of our concern, an analogous term would be perhaps *k-type* since most interrogative expressions contain the velar plosive spelled as 'k' (e.g., *kdy*, *kde*, *kdo*, *který*, *kolik, kam*). In this study, we will use the term question-word questions (or QWQ).

Czech is a West-Slavic language (related to Slovak or Polish) spoken primarily in Central Europe by about ten million people. With regard to question-word questions, traditional descriptions of Czech intonation suggest two possibilities, often discussed in quite a polemic manner. Some authors claim that the intonation nuclear pitch accent or nucleus (represented by a phrasal melodic peak) comes late in the sentence, thus making the melody identical to that of a statement (e.g., authoritative and often cited [1], [2], [3], [4], [5]). Others argue that the nucleus is typically anchored to the sentence-initial question word, which results in an early peak (e.g., [6], [7]). This solution would resonate with the views of some cross-language researchers (e.g. [8], [9], [10]) that the default nucleus position in QWQs is on the interrogative pronoun. Both early- and late-peak proponents, nevertheless, agree that the melodic pattern is falling, or in other words, that the boundary tone is low (cf. Figure 1 below).

According to the former solution, the question *Kde je moje svačina?* (*Where is my snack?*) would have a peak linked to the word *svačina* (*snack*), whereas the latter solution links the peak to the word *kde* (*where*). Similarly, *Kolik máte bodů?* (*How many points do you have?*) would have the late peak associated with *bodů* (*points*), while the early peak solution would anchor the peak to the word *kolik* (*how many*). Unfortunately, this controversy in Czech intonology is based solely on introspection and private diary entries of casual observations. Romportl [2] carried out some instrumental analyses of F0 tracks, but his sample was small and collected under unclear circumstances. No other material-based study has been published to our best knowledge.

It is fair to note, however, that both opposing camps admit not only the existence, but also phonological legality of the alternative forms. The core of the argument is then the typicality claim, i.e., the answer to the question which form is more frequent or archetypal (also felt as 'normal' or, in some approaches, unmarked). Unlike the previous efforts to decide the matter, we opted for ignoring our own opinion. Instead we carried out an empirical study of the problem.

Initially (i.e., before this study), we prepared a recording session for 28 native speakers of Czech who were asked to act out short dialogues in which, among other things, various QWQs were scattered. They differed in contexts, numbers of syllables and, naturally, in lexical contents. The results [11] showed that QWQs without a contrastive context are read with a melodic peak in the early part of the phrase. The peak was usually associated with the second or third post-stress syllable due to the prevalence of L*+H pitch accent in current Czech. The resulting QWQ melodies thus differed from those of statements, which are also falling but without an early peak in their typical rendering. In the group of 28 speakers, only 4 used a late peak in QWQs, which would be predicted by some of the traditional authoritative descriptions.

The question remained, however, about the perceptual aspect of the problem. Would listeners evaluate the early peak more 'normal' and the late peak as less 'normal', or vice versa? Which of the two forms is more at the forefront of their linguistic inventories, or in other words, less in need of special contextual justifications? Our null hypothesis would predict that there is no difference between early and late melodic peaks in terms of their evaluation by listeners. Yet the long-lasting dispute in Czech linguistics and our own informal observations suggest that the two forms do not have the same use and the same impact. Therefore, alternative hypotheses would state that subjects will favour one of the solutions proposed in the above-cited literature, preferring the early or the late peak in QWQs.

http://dx.doi.org/10.21437/Interspeech.2019-2082

2. Method

2.1. Material

As the key point was to compare early and late F0 peaks, it was obvious that we could not use two natural productions, since the speakers might introduce other effects into their speech apart from the differences in peak position. Furthermore, we could not simply use one natural recording (e.g., with an early peak) and manipulate the F0 contour into the other condition (late peak) because this would create a bias against the manipulated version. Therefore, both conditions had to be created from a single recording *and* from a mid-way (monotonous) F0 pattern that would not favour any condition.

2.1.1. Target items

Eight native speakers of Czech (4 males, 4 females) were recorded in a sound-attenuated booth at the Institute of Phonetics, Prague (16-bit 48-kHz uncompressed wav files). They read eight QWQ phrases, differing in lexical content, rhythmical structure and the number of syllables. For example: /ˈgdɛbɪx_ˈnaʃɛltɪ ˈbanaːnɪ/ *Where would I find the bananas?* or /ˈprot͡ʃvaːmto ˈnɛvɪʃlo/ *Why didn't it work out?* The speakers were instructed to use monotonous intonation and to speak at a comfortable tempo (the articulation rate ranged between 4.8 and 7.0 syllables per second, mean = 5.9). They generated several repetitions of each utterance to train natural production in the monotonous mode. The final productions sounded fairly monotonous, but – not surprisingly – some acoustic variation in F0 was always present. One phrase from each speaker was selected so that at the end of the selection process, each target sentence for the perception test was produced by a different voice.

Instead of fabricating the F0 contours from scratch, the manipulation was based on real productions from our previous study [11]. The original F0 contours – both early and late – were measured in ST in Praat [12], smoothed and transplanted to the target monotonous sentences using pitch-synchronous overlap and add method (PSOLA) to create two manipulated versions, differing only in the position of the main F0 peak (Fig. 1). A group of 6 judges (incl. the authors) listened to the material in order to identify any problems with the resynthesized stimuli. As a result, several files had to be further processed to remove artefacts (due to specific segments or features, like glottalization) and local F0 movements were more precisely synchronized with syllable nuclei when the alignment sounded artificial (in the syllables outside of the main peak). The sixteen finalized resynthesized recordings were saved as 16-bit 32-kHz wav files.

2.1.2. Distractors and training items

It would have negative impact on the results if listeners noticed that two versions of QWQ sentences were tested, differing systematically in melody. Therefore, two types of distractor items were added: statements and yes-no questions. These recordings were extracted from various corpora of the Czech language (read dialogues, a variety of voices). The recording conditions and the structure, tempo and speaking style of the sentences were comparable to that of the target sentences. There were 16 statements and 6 yes-no questions. All of them were resynthesized (without any change), and 8 statements and 2 yes-no questions were resynthesized also with a slight change in the F0 contour. In total, the material included 16 target items and 32 distractor items.

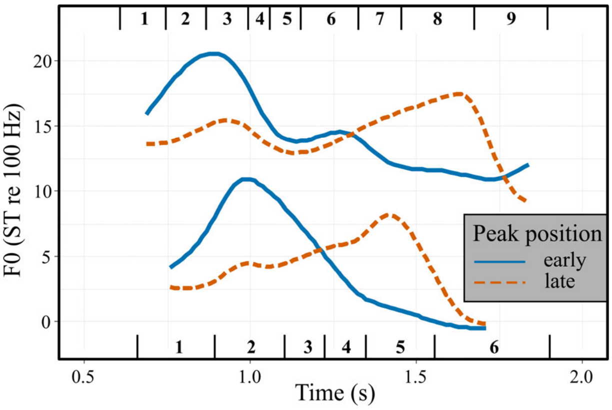

Figure 1: *Manipulated contours for the early vs. late peak condition for two sentences/speakers (female above, male below). Syllable boundaries are indicated by vertical ticks on the upper and lower axes.*

Similarly, six items were created for a training session. There was one QWQ with an early peak, one QWQ with a late peak (but produced by a different speaker), three statements (two identical sentences, produced by a male and a female speaker) and one yes-no question.

2.2. Listeners

A group of 32 listeners participated in the experiment (17 females and 15 males, with median age of 28.5). Most of the listeners came from the Bohemian region of the Czech Republic. They were all native speakers of Czech and were not phonetically trained. Naive listeners were preferred because we aimed for eliciting *general/holistic impressions* rather than evaluations in the *analytical (metalinguistic) mode* that professional phoneticians might more readily resort to.

2.3. Experimental design and procedure

The experiment was created and administered in Praat [12] via the Demo Window environment, which allows for complete control over the experiment. In order to tie the perceptual evaluation to a real-life concept, the participants were told that they were going to evaluate recordings of different speakers reading a few lines from a screenplay. The actors supposedly produced several takes of the scripted text, so the listeners should not worry about hearing identical texts by various voices. The listeners were asked to imagine that there was an audition (casting) for a role, and that they were the judges deciding who is going to win. The criterion was whether the performance sounded natural (like from a real conversation) or strained (obviously acted, or read from a page). These were used as definitions of the extremes of a 7-point scale (1 = strained, 7 = natural), which was in fact continuous: the slider could be positioned anywhere on the scale, limited only by the screen's pixel resolution. Context was provided on the screen, comprising a transcript of the stimulus plus a lead-in sentence. For instance, the QWQ *Where would I find the bananas?* was preceded by the text *I'll prepare it, then* or the QWQ *Why didn't it work out*? was preceded by the yes-no question *Did you investigate the causes?*

Whereas the training session was presented in a fixed order, the test session was administered in a randomized order with the following restrictions. To prevent the two manipulated conditions from appearing in close proximity, the stimuli were split into two groups: eight target items (4 early, 4 late) appeared in the first block, eight (the opposite versions) in the second block. In addition, each block included the same number of distractor items. The order of the two blocks was balanced across participants (half received A-B order, half B-A). Items within the two blocks were randomized for each listener. There was a pause after every 12 items, creating four groups of items. The duration of the whole experiment was approximately 15 minutes (it depended, for example, on the fact that the participants had the option to replay each sentence once more).

The perceptual test was distributed by email. The participants followed instructions self-contained in the test. They were instructed to use headphones and to ensure a tranquil environment (specifically, they were asked to participate at a moment of leisure and to eliminate all manners of disturbance, such as an animal in the room, talkative family members, etc.). After completing the experiment, an output file with the coded results was sent back to the experimenters. In addition to slider position (ranging between 0 and 1), total response time for each item was collected.

2.4. Analysis and statistics

The statistical analysis was conducted through the fitting of linear mixed-effects (LME) regression models via the *lme4* [13] package in *R* [14], using *ggplot2* [15] and *effects* [16] for visualization. Mixed-effects models allow for both fixed and random effects. Fixed effects included PEAK POSITION (early × late), BLOCKING (AB × BA), ORDER (specific order of items in the test), listener SEX and AGE and SPEAKER SEX. Random effects included random intercepts for LISTENER and PHRASE and random slopes for PEAK POSITION (for the importance of including random slopes, see e.g. [17]). A random slope in the statistical model allows for the situation in which each listener/phrase may show different sensitivity to the fixed effect (in our case the peak condition). Thus, if the effect of peak position is verified by the analysis even after accounting for this random variation, the effect may be interpreted as generalizable to other listeners/phrases. Rather than using stepwise regression to eliminate insignificant variables (see [18] for arguments against this practice), the model comprised all variables. Significance testing was done by comparing the full model to a reduced model (without the effect in question) by the *anova* function using likelihood ratio tests.

3. Results

3.1. Fixed effects

The significance of individual variables without interaction was tested first. PEAK POSITION significantly improved the model ($\chi^2(1) = 7.9$, $p < 0.01$), as did ORDER ($\chi^2(1) = 5.5$, $p < 0.05$). None of the other predictors reached significance: BLOCKING ($\chi^2(1) = 0.01$, $p = 0.91$), LISTENER AGE ($\chi^2(1) = 0.4$, $p = 0.52$), LISTENER SEX ($\chi^2(1) = 1.7$, $p = 0.19$), SPEAKER SEX ($\chi^2(1) = 3.0$, $p = 0.09$). The parameters of the model are given in Table 1. The late peak condition lowered the evaluation score by 8% of the scale with an SD of 2%, and later trials were evaluated more positively than earlier trials. Figure 2 shows the key effect of interest, peak position.

Further, the effects of possible interactions were explored. The crucial interaction was between PEAK POSITION and BLOCKING, as blocking might have affected the size of the effect. However, the interaction between the two factors was insignificant ($\chi^2(1) = 1.9$, $p = 0.17$). The other pairwise interactions with PEAK POSITION were not expected to be relevant a priori, but were checked anyway. The interactions did not reach significance (ORDER: $\chi^2(1) = 0.1$, $p = 0.80$; AGE: $\chi^2(1) = 0.02$, $p = 0.88$; SEX: $\chi^2(1) = 1.6$, $p = 0.20$; SPEAKER SEX: $\chi^2(1) = 0.5$, $p = 0.50$). Therefore, only the non-interaction terms are included in the final model.

Table 1: *Regression coefficients of fixed effects in the LME model (for random effects see Figures 3 and 4).*

Fixed effect	Estimate	SE	t-value
Intercept	0.4734	0.0910	---
PEAK POSITION (late)	-0.0768	0.0241	-3.18
BLOCKING (BA)	0.0046	0.0402	0.12
ORDER	0.0048	0.0020	2.35
AGE	0.0014	0.0021	0.64
LISTENER SEX (male)	-0.0568	0.0418	-1.36
SPEAKER SEX (male)	0.1237	0.0635	1.95

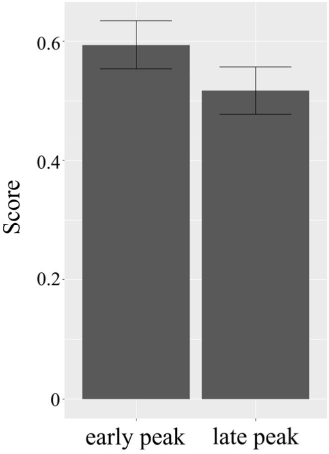

Figure 2: *The effect of* PEAK POSITION *on naturalness score (0 = strained, 1 = natural) in the LME model.*

3.2. Random effects

For proper generalizability, the random effect structure should be maximal [17]. It is especially important to include random slopes for the key fixed effects.

The coefficients for PHRASE revealed that individual phrases were evaluated differently, but the effect of PEAK POSITION was comparable in all eight phrases (Fig. 3). In contrast, there were both differences among listeners (some were generally stricter, some more lenient) and among the strength of the manipulated condition on the listeners (Fig 4). Three listeners even yielded a minor effect in the opposite direction.

3.3. Response time

Response time (RT) as measured here only approximates the time taken by the listeners to respond (unlike in standard RT experiments, there was no time pressure on the participants and, moreover, longer response times might have also occurred due to the fact that the participants were allowed to replay the stimulus once). LME analyses with LOGARITHMIC RT as the dependent variable and the same structure of fixed and random effects as above showed a highly significant effect of ORDER ($\chi^2(1) = 21.6$, $p < 0.001$); the response time substantially decreased during the course of the experiment. The rest of the predictors and all interactions were not significant. However, the effect of PEAK POSITION bordered on significance ($\chi^2(1) = 3.7$, $p = 0.05$). In comparison with early peaks, the response time was somewhat larger in the late peak condition (i.e., listeners responded more slowly, or needed more replays when listening to a late F0 peak).

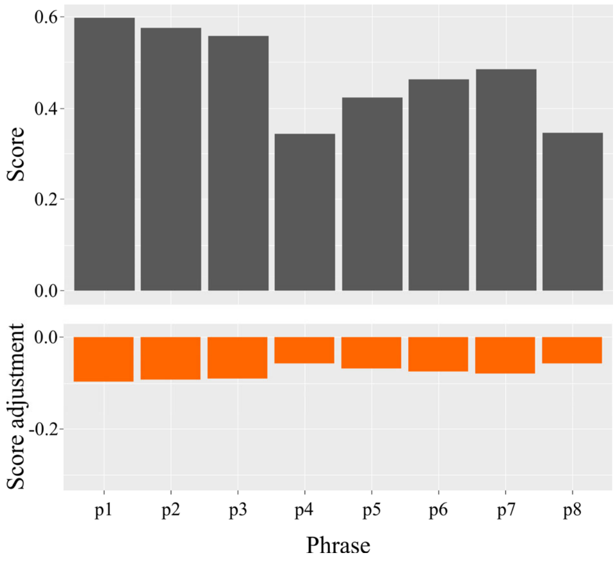

Figure 3: *Random effect coefficients for* PHRASE *(phrase intercepts above, slopes indicating late peak adjustment below; score 0 = strained, 1 = natural).*

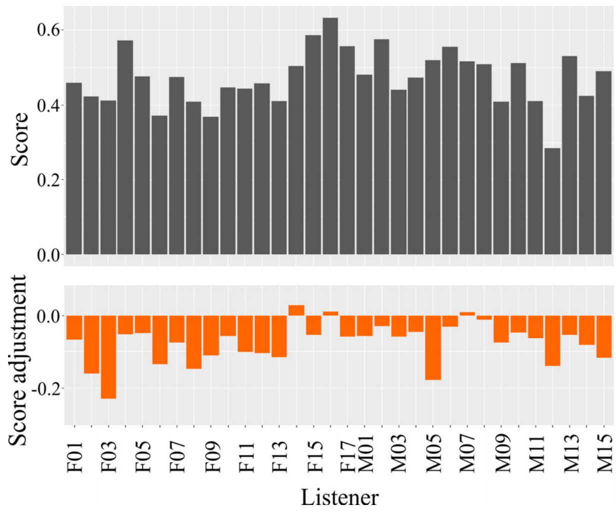

Figure 4: *Random effect coefficients for* LISTENER *(listener intercepts above, slopes indicating late peak adjustment below; score 0 = strained, 1 = natural).*

4. Discussion

The longstanding uncertainty in Czech linguistics concerning the 'unmarked' or typical form of question word questions (QWQs) used to be tackled on the basis of personal opinion or experience. Our experiment was carried out to investigate the matter empirically, in line with current research conventions. The disparate accounts could be attributed to a difference in the researchers' approach, or, one might argue, to language change. However, our results from the perception task, where the stimuli with an early peak were evaluated as more natural, as well as from the preceding production experiment [11] resonate with [6] from the 1930's and [7] from the 1960's, while being in opposition to [1] and [2] from the 1950's (sources [3] and [4] only seem to rely on the earlier studies). Linear language change is therefore an unlikely cause.

Although we intended to uncover the general lay listeners' preferences, we have to warn against over-generalization. Despite a preference for the early peak, the induced context allowed – sometimes to a greater, sometimes to a smaller extent – for the late peak as well. It was usually not evaluated as unnatural. This is because the late peak is an apparent choice in the case of narrow focus, or less commonly, but also possibly, expressive emphasis. Furthermore, the bottom graph in Figure 3 demonstrates that individual phrases behaved somewhat differently, even if in the same direction. The meaning of an utterance is always unique and it comes into being when the utterance interacts with the context. The context provided with each QWQ necessarily interacted with the listener's idiosyncratic experience. Therefore, the greater acceptability of the early peak is quite clear, but in individual cases, the late peak may be appropriate.

Similarly, individual listeners assessed the utterances differently. For instance, listener F3 and M5 (Fig. 4) were much more sensitive to the manipulation condition than most of the others. Moreover, three listeners out of 32 exhibited the opposite direction of the effect, even if to a negligible extent. This again proves that a speaker's experience and different cultural norms are part of the evaluation process.

It is useful to notice that our listeners were less strict towards the end of the experiment (but note that items were presented in random order to each listener), while, at the same time, RT decreased. This clearly manifests adaptation to the experimental task (and can be accounted for in the model).

Future research should experiment with more contexts and explore the contrastive and affective effects of not only peak positions in the phrase, but also peak pitch ranges and peak alignments (synchronizations) with the segmental chain.

5. Conclusions

This research addressed a long-lasting debate among Czech intonologists as to whether the typical or unmarked QWQs should contain an early or a late melodic peak. Users of the Czech language found the early peak more natural sounding. This is in agreement with production data, as QWQs differ from unmarked statements by the presence of an early peak.

6. Acknowledgements

This work was supported by 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 the Charles University project Progres Q10, *Language in the shiftings of time, space, and culture*. We are also grateful to the students of a phonetics seminar *Research in Speech Perception* for their help with experiment administration.

7. References

[1] F. Daneš, *Intonace a věta ve spisovné češtině*. Praha: ČSAV, 1957.

[2] M. Romportl, *K tónovému průběhu v mluvené češtině*. Praha: Královská česká společnost nauk, 1951.

[3] M. Romportl, *Základy fonetiky*. Praha: Karolinum, 1982.

[4] Z. Palková, *Fonetika a fonologie češtiny*. Praha: Karolinum, 1994.

[5] 5 J. Petr, M. Dokulil, K. Horálek, J. Hůrková-Novotná, and M. Knappová, *Mluvnice češtiny 1.* Praha: Academia, 1986.

[6] 6 V. Mathesius, "Řeč a sloh," in *Čtení o jazyce a poesii*, B. Havránek and J. Mukařovský, Eds. Praha: Družstevní práce, 1942, pp. 13–104.

[7] 7 B. Hála, *Uvedení do fonetiky češtiny na obecně fonetickém základě*. Praha: ČSAV, 1962.

[8] 8 S. Beck, "Intervention effects follow from focus interpretation," *Natural Language Semantics,* vol. 14, pp. 1–56, 2006.

[9] 9 A. Haida, "The indefiniteness and focusing of wh-words," in *Proceedings of SALT 18*, T. Friedman and S. Ito, Eds. New York: Cornell University, 2008, pp. 376–393.

[10] 10 H. Truckenbrodt, "On the prosody of German wh-questions," in *Prosody and Meaning*, G. Elordieta and P. Prieto, Eds. Berlin: de Gruyter, 2012, pp. 73–118.

[11] 11 J. Volín and P. Šturm, "Fundamental frequency tracks of question-word questions in natural and synthetic speech," *Akustické listy*, in print.

[12] 12 P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.0.47, retrieved 8 February 2019, http://www.praat.org/, 2019.

[13] 13 D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[14] 14 R Core Team, "R: A language and environment for statistical computing [Computer program]," R Foundation for Statistical Computing, Vienna, https://www.R-project.org/, 2018.

[15] 15 H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer, 2009.

[16] 16 J. Fox, "Effect displays in R for generalised linear models," *Journal of Statistical Software*, vol. 8, no. 15, pp. 1–27, 2003.

[17] 17 D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of Memory and Language*, vol. 68, no. 3, pp. 255–278, 2013.

[18] 18 F. E. Harrell, *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer, 2001.