Speech Prosody 2016
31 May - 3 June 2016, Boston, USA

The prosody of backchannels in Slovak

Štefan Beňuš
Constantine the Philosopher University in Nitra, Slovakia
Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
sbenus@ukf.sk

Abstract

This paper explores the prosodic realization of single affirmative cue words functioning as backchannels in cooperative task-oriented dialogues. It focuses on the comparison between four major lexical items signaling this meaning (*mhm*, *no*, *uhhuh, áno*) and compares the results with the realization of backchannels in a similar corpus of American English. Additionally, the entrainment properties of backchannels in terms of their similarity to the end of the preceding utterance in Slovak are examined. The results suggest that backchannels in Slovak are realized in a largely similar way to those in American English but with differences in intensity and duration when compared to acknowledgements and agreements. Slovak backchannels are also similar in slope and curvature of f0 contours to the ends of the turns preceding them.

Index Terms: backchannels, Slovak, entrainment, computational paralinguistics

1. Introduction

Backchannels are feedback items such as *okay* and *mhm* that acknowledge that the speaker is attending to the interlocutor and signal the interlocutor to continue speaking since the speaker producing a backchannel does not wish to take the floor. They typically belong among the most frequent lexical items in corpora of conversational speech. However, these items are also commonly functionally overloaded in that they signal multiple pragmatic, discourse, and interactional meaning. For example, [1] identified ten such meanings for *okay* in American English. Some of these functions can be disambiguated based on the prosodic characteristics of backchannels. For example, [2] reported higher pitch, intensity and pitch slope for backchannels in American English as compared to agreements and other discourse functions, and also found a tendency for backchannels to follow intonational phrases (IPs) with rising pitch preceded by a low tone around 0.5s before the end of the IP.

In addition to providing ample material for studying the relationship between the prosody and meaning, backchannels are also interesting in a broader cognitive sense since they show a link between the mental state(s) of the speaker and her prosody [3], and might be examined for similarities and differences in cross-linguistic research to identify general and language-specific aspects of human conversation. Furthermore, they signal multiple emotional characteristics, e.g. [4], or participate in negotiating inter-personal relationships such as dominance [5].

Another feature characterizing backchannels is their relationship to the preceding context. Reference [6] measured the similarity between the final 500ms of a turn and the initial material of the following turn in collaborative games corpus of American English and found that backchannels were significantly more similar in mean pitch to interlocutor's speech preceding them than they were to subsequent speech (which was lower) and also more similar than other turn types were to their prior turns (they were higher). The authors suggested that this entrainment of backchannels is related to their unobtrusive realizations when they participate in creating common ground.

Working with the same corpus, [7] investigated the potential effect of two confounding factors on the backchannel entrainment. First, backchannels typically follow preceding speech with shorter latencies than other non-overlapping turn types. Hence, the relevant representations of the interlocutor's prosody is more strongly activated in the speaker's cognition and thus it is more likely for the speaker to prosodically continue where the interlocutor has ended. Second, backchannels also tend to be shorter than other turn types since the speaker does not intend to take the floor and continue speaking, and his or her pitch and intensity may therefore be lower in the backchannel than in these other turns. After matching backchannels with other turn types regarding their latency and length, [7] reported that backchannels were still more similar to preceding turns than other turn types.

Using the same corpus as the one analyzed in the current paper, two recent studies provide a first look at the global and local characteristics of prosodic entrainment in Slovak [8, 9]. Despite evidence of both global and local entrainment, especially in intensity, the propensity of the observed entrainment in general can be characterized as lower than in English.

Given the discussion above, the primary goal of the current paper is twofold. First, investigate the relationship between prosody and function for Slovak backchannels and compare with known characteristics of English backchannels. It is believed that researching cross-linguistic properties of prosody-function relationship in backchanneling contributes to our understanding of cognitive prosodic system.

Second, given the robust entrainment of English backchannels and somewhat weaker general entrainment observed in Slovak, the other goal is to test if the strong entrainment properties of backchannels are language specific or might be considered universal. Despite the ubiquity of backchannels in speech, relatively little attention has been devoted to prosodic cues inviting them; [10] described a region of low pitch lasting at least 110 ms before the backchannel, and [11] described a rising pitch, and high levels of pitch and intensity among the cues increasing the likelihood of the occurrence of a backchannel. But the similarity of f0 contours as a potential inviting cue for backchanneling has not been investigated.

Finally, the motivation for this research is related to the fact that the disambiguation of the discourse/pragmatic functions as

10.21437/SpeechProsody.2016-85

well as understanding their entrainment characteristics poses a challenging task for spoken dialogue systems. However, both recognizing these functions in human speech by an automated system and synthesizing them in appropriate contexts with appropriate prosody promises improvements in the naturalness and efficiency of such dialogue systems.

2. Methods

Data come from SK-Games a corpus of interactional task-based dialogues adapted with slight modifications from the object games of the Columbia Games Corpus, described e.g. in [11]. Eleven native speakers of Slovak were paired to play nine sessions so that 7 of those speakers played twice while the remaining 4 played once. Each session included 14 tasks of placing a given object on one player's (The Placer) screen to match as closely as possible the position of this object on the other player's (The Describer) screen. The players did not see each other and thus had to reach the agreement on the placements collaboratively through spoken interactions. The roles in describing and placing the objects changed repeatedly and were balanced between the two players.

After recording, the signal was semi-automatically segmented to inter-pausal units (IPUs, threshold of 150ms) and speech within these units was manually transcribed and automatically aligned with the signal using the SPHINX toolkit adjusted for Slovak [12]. This forced alignment was then manually corrected. The current corpus comprises roughly six hours of speech, and consists of 35,758 words.

The list of affirmative cue words (ACWs) was determined by selecting the most frequent words (N > 100) that can signal feedback with positive polarity. The threshold was chosen to balance the coverage of both the types and conversational functions. The list includes six such cue words {*áno*, *dobre*, *hej*, *no*, *mhm*, *uhhuh*} with the first three roughly corresponding to 'yes', 'good/well', 'yes' respectively, and *no* is a shortened form of *áno* but with much more functional ambiguity as will be shown below.

Following [13], the potential functions of these cue words were summarized in a labeling scheme shown in Table 1. The majority of functions are similar to various functions of ACWs in American English, and the highlighted three (RP, R, S) serve as the core for the current paper. However, some new functions were identified for Slovak ACWs, most notably those labeled as Z, RZ, or J. A single experienced annotator (author) labeled all occurrences of ACWs in the corpus using Praat [14] interface trying first to determine the function from the transcription and listen to the speech only if transcripts were ambiguous. While the comparison with other labelers and inter-annotator agreement will be performed in future, this version of subjective labeling provides internal consistency for the initial exploration of backchanneling in Slovak.

To investigate the relationship between the functions and the prosodic realization of ACWs, standard features of median, maximum, minimum of f0 and Intensity were extracted from each ACW and z-score normalized for the current speaker and session. The same was performed for the final 500ms of the turn preceding all turn-initial ACWs (or less if the turn was shorter than 500ms) provided that the ACW was judged as a direct feedback to the material in the preceding turn, which was manually determined. Latency, i.e. the interval between the preceding turn-end and the onset of these turn-initial ACWs, was also extracted.

Table 1. *Labeling scheme for identifying discourse and pragmatic functions of Slovak affirmative cue words.*

	Meaning
D	Encourage some action, go on, do something
E	I want to repair/redo something I've just said or did
H	Hesitation, I am stalling for time
J	Softening of what is to follow, a hedge
K	Signal the end of the current topic or discourse segment
L	Literal modifier
N	I want to start a new topic or a new discourse segment
PH	Express assessment of something that just happened, usually on receiving a score
Q	Asking a question, checking
R	I acknowledge that I understand, I got it
RN	I acknowledge that I understand, and I want to start a new discourse segment
RP	I acknowledge that I understand, and please continue (Backchannel)
RZ	I acknowledge that I understand, but I want to add something or express mild disagreement
S	I agree, also as an answer to a questions, usually meaning yes
Z	I want to express an idea opposite to implied before, usually meaning but or well
X	None of the labels correspond to the perceived meaning

Additionally, discrete cosine transformation was employed on the interpolated and time-normalized f0 curves after converting them to semitones. The second and third coefficients (referred to here as dct1 and dct2) corresponding to the slope and curvature of the f0 curve respectively were calculated [15]. This was done both for the f0 curves of ACWs as well as for the last 500ms of the turn preceding the ACWs. Finally, to assess the degree of prosodic similarity of ACWs to the preceding material, negated absolute difference between the value of feature *f* for the ACW and that for the preceding material, was also calculated; e.g. [16].

$$ENT = -|\,spea\ker^{1}{}_{f} - spea\ker^{2}{}_{f}\,| \quad (1)$$

For statistical testing of the effect of discrete factors (here functional labels and lexical items) on the continuous prosodic features linear mixed effects models implemented in R's *lmerTest* package were used. For the significance of the differences t-tests were used with Satterthwaite approximations to degrees of freedom for the calculation of adjusted p-values with the *glht* function for general linear hypotheses testing of R's *multcomp* package [17]. This allows for testing of hypotheses with correcting for the post-hoc comparisons. When building these models, random intercepts and slopes varying with the factor under investigation for subjects were always used to ensure as much anti-conservative design as possible. For the sake of readability, *t* and *p* values are reported only for the most crucial results but all results reported in text as significant reached significance in these models at $p < 0.05$.

3. Results

3.1. Descriptive observations

Table 2 below shows the distribution of lexical items and the functions listed in Table 1 for the entire corpus.

Table 2. *Distribution of the functions and lexical tokens for Slovak affirmative cue words.*

	ano	*dobre*	*hej*	*mhm*	*no*	*uhhuh*	∑
D	0	0	0	0	40	0	**40**
E	0	0	0	0	43	0	**43**
H	0	0	0	0	22	0	**22**
J	0	0	0	0	16	0	**16**
K	0	117	2	2	52	0	**173**
L	0	28	0	0	0	0	**28**
N	1	11	1	0	114	1	**128**
PH	0	27	2	5	49	0	**83**
Q	14	10	84	0	0	0	**108**
R	62	46	81	167	112	46	**514**
RN	2	15	0	13	53	2	**85**
RP	55	4	6	431	266	79	**841**
RZ	5	1	5	12	86	0	**109**
S	120	1	166	63	118	5	**473**
Z	0	0	0	0	56	1	**57**
X	0	2	5	6	34	2	**49**
∑	**259**	**262**	**352**	**699**	**1061**	**136**	**2769**

The table offers several observations. First, not surprisingly, backchannels (RP) are the most frequent ACWs. Acknowledgements (R) and agreements (S) are also very common, and other functions are much less frequent. Second, of these three most common functions, acknowledgements have a very balanced distribution in terms of lexical items while backchannels tend to disprefer *dobre* and *hej* and agreements *dobre* and *uhhuh*. Third, *no* is not only the most frequent, but also the most functionally loaded ACW since it can convey any meaning except question/check and literal modifier. In this sense, *no* is similar to *okay* in American English. On the other hand, many functions can be expressed only by *no* (D, E, H, J).

Finally, exploring speaker variability in the choice of lexical items for backchannels, all subjects produced *mhm* (between 3 and 89 tokens per speaker), all but one produced *no* (1-51), but *uhhuh* and *áno* were not so well represented: 5 speakers produced one or zero *uhhuh* and 6 speakers produced one or zero *áno*.

3.2. Prosodic differences among backchannels

We now take the four most frequent backchannels {*áno*, *no*, *mhm*, *uhhuh*} and examine prosodic differences among them. In the temporal features, not surprisingly, *no* was significantly shorter than the other three backchannels, but this is primarily due to *no* being monosyllabic and the other three backchannels being disyllabic. Interestingly, *áno* was longer than *no* ($t = 4.6$, $p < 0.001$) but shorter than *mhm*/*uhhuh* ($t = 3.2$, $p = 0.005$) Latency was not affected by lexical choice much, only *áno* had significantly shorter latencies than the other backchannels (e.g. $t = 3.1$, $p = 0.01$ for *áno*-*mhm* comparison).

Regarding mean pitch and intensity, the lexical items did not differ, only *mhm* was significantly less loud than the other backchannels (e.g. $t = 5.1$, $p < 0.001$ for *no*-*mhm*), which is plausibly related to the lower sonority of the nasal realization of *mhm* as compared to vowels in the other three backchannels. The f0 contours of the four backchannel types are illustrated in Figure 1. Despite the general similarity of the shapes, the top two backchannels (*mhm*, *uhhuh*) had greater slope and curvature, measured with the dct1 and dct2 coefficients respectively, than the bottom two (*no*, *áno*) with no significant differences within these two pairs; $t = 4.5$, $p < 0.001$, $t = 3.9$, $p < 0.001$ for the *mhm*-*no* comparison in dct1 and dct2 respectively.

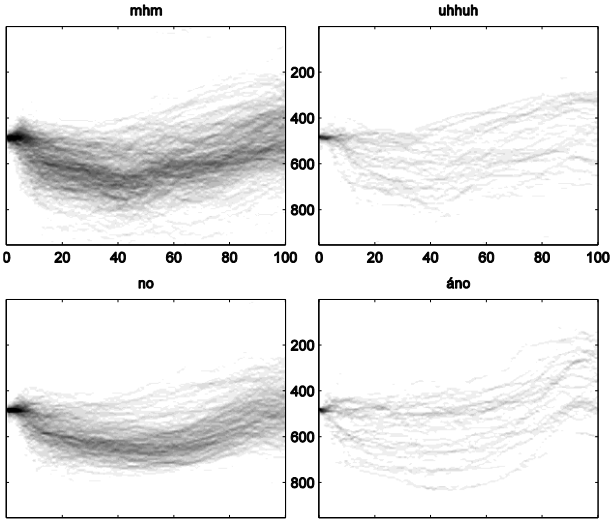

Figure 1: *f0 contours for four most frequent backchannels using a bit-map method [18]; f0 contours were normalized to 100 points and transformed to semitones; x-axis is time normalized to 100 units and y-axis corresponds to f0 height but the units are meaningless.*

3.3. Triplet Bc-Ack-Agr

The three most frequent functions of Slovak ACWs are backchannel (RP), acknowledgment (R), and agreement (S). These also turn out to be particularly difficult to disambiguate. We test the prosodic realization of these functions using linear mixed models with a 4-level grouping factor FUNCTION (RP, R, S, and XX that includes all the other functions except literal modifier), SPEAKER having random slope and intercept, and LEXICALITEM having random intercept. In the temporal features, RPs were significantly longer than all other three categories ($t = 4.0$, $p < 0.001$ for the closest RP-R comparison) and RPs had significantly shorter latencies than XX ($t = 2.6$, $p = 0.03$). No other significant differences were observed.

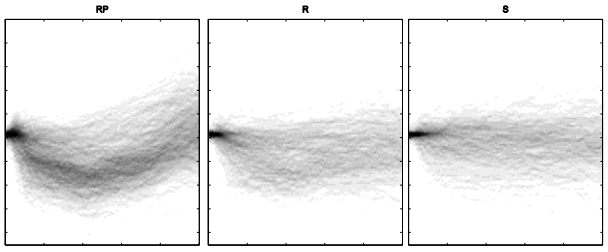

Figure 2: *f0 curves of the backchannels (RP), acknowledgments (R), and agreements (S); same techniques as in Figure 1.*

For pitch mean, our test showed the only significant difference between RPs and Rs in that the former were higher than the latter ($t = 2.6$, $p = 0.035$). In intensity, agreements (S) were significantly louder than the other functions; $t = 3.8$, $p < 0.001$ for S-RP comparison

Regarding the shape of the f0 contours, shown in Figure 2, backchannels (RP) had significantly greater slope and curvature than all the other functions; e.g. $t = 6.0$, $p < 0.001$ and $t = 7.8$, $p < 0.001$ respectively when compared to acknowledgments (R). Tendencies for greater curvature of acknowledgments when comparing and other functions were reported; $t = 2.3$, $p = 0.08$ for agreements and $t = 2.5$, $p = 0.051$ for the collapsed XX category. Variances in slope and curvature were significantly greater in RPs/Rs than Ss suggesting more varied realization of backchannels and acknowledgments than agreements.

3.4. Similarity to preceding turn

As discussed in Introduction, backchannels were found to be more similar (entraining) to the final prosody of the preceding turn than turns identified as smooth switches. We were interested, if this unobtrusiveness feature of backchannels can be also identified when backchannels are compared to other affirmative cue words. Hence, the entrainment of backchannels is tested here in even more constrained material than the one analyzed in [6] and [7] where backchannels were compared to all smooth switches.

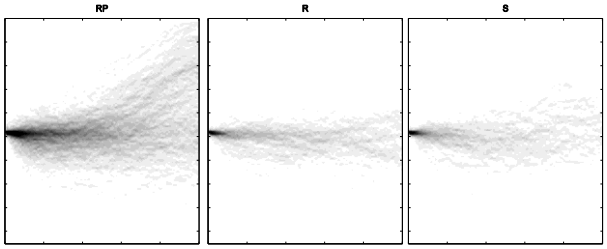

Figure 3: *f0 curves of the last 500ms of the turn preceding the backchannels (RP), acknowledgments (R), and agreements (S); same techniques as in Figure 1.*

The analysis with the measure of entrainment defined in (1) above did not yield any significant difference between the backchannels and other functions. Hence, when compared to other similar lexical items with non-backchannel function, backchannels do not seem to display much entrainment.

However, consider Figure 3 that shows f0 contours of the last 500ms preceding the three most common functions RP, R, and S. Turn endings that precede backchannels RP have robustly greater pitch slope than material preceding all the other functions; e.g. $t = 4.5$, $p < 0.001$ comparing RPs and the most similar slightly rising material preceding agreements (Ss). Given the same relationship reported for the slope of f0 curves for ACWs in Figure 2 above, we might say that backchannels entrain with the preceding material on the (rising) slope of the f0 contours. A similar, albeit less robust, result obtains for f0 curvature since the material preceding backchannels had greater curvature than that preceding acknowledgments ($t = 2.7$, $p = 0.027$). Hence, the 'scoopy rising' f0 shape of Slovak backchannels is similar to the f0 contour of the preceding material, and this similarity tends to be greater than for other common functions of affirmative cue words.

Evidence for complementary entrainment can be observed in the tendency for rising f0 contours preceding Ss and falling contours of the agreements themselves. This is plausibly related to the observations in [19] who suggest that in task-collaborative interactions, the demands for constructive engagement such as providing an answer to a question or a response after a request, which are commonly incongruent prosodically, take precedence over the tendency for the interlocutors to entrain.

4. Discussion & Conclusions

This paper set out to compare the prosodic properties of backchannels in Slovak and Standard American English (SAE). The labeling of discourse/pragmatic functions identified that all six affirmative cue words might function as backchannels but two {*dobre, hej*} were very infrequent and thus not included in the quantitative analysis. Similarly to English, Slovak backchannels have rising pitch and follow turns ending in rising pitch. The examination of f0 shapes with discrete cosine transformation provides a novel way of characterizing these shapes with the slope and curvature characteristics. These firmly establish a 'scoopy' nature of backchannel f0 contours

Several features of Slovak backchannels are different from the English ones when compared to the findings of [2]. First, Slovak backchannels were longer than acknowledgments, which was not observed for SAE. Additionally, Slovak agreements were louder than backchannels while the opposite relationship was reported for SAE. It is thus plausible that the unobtrusiveness of backchannels in Slovak is signaled by overall lower intensity while general entrainment is more prominent in SAE for conveying this unobtrusiveness.

As concerns the entrainment characteristics, the most direct way of comparing similarity of f0 and intensity of a backchannel to that of the end of the preceding turn did not yield any evidence of entrainment. This might suggest that the general tendency for lower entrainment in Slovak than in English also extends to backchannels. However, this would be a pre-mature conclusion. First, note that [7] operationalized backchannel entrainment in English as a significant difference between the similarity between the interlocutors (partners) vs. the mean similarity between the pairs of subjects that did not interact with each other (non-partners). Second, we did observe entrainment on the shape of f0 contours for backchannels that the 'scoopy rising' contours of backchannels tend to be 'invited' by similar turn-endings preceding them.

Future work includes the assessment of entrainment using the same methodology of partner vs. non-partner differences as in [7] and the research into the prosodic characteristics of other functions of Slovak ACWs. For example, the new Z & RZ functions expressing (mild) disagreements show interesting differences in temporal latencies predicted from EEG studies regarding the difference between expected/unmarked and unexpected/marked turn continuations. Also, the data suggest prosodic gradation regarding the acknowledgments (R), their combination with additional function (RP, RN, RZ), and plain expression of this other function (N, Z).

5. Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055.

6. References

[1] A. Gravano, Š. Beňuš, J. Hirshberg, S. Mitchell, and I. Vovsha, "Classification of Discourse Functions of Affirmative Words in Spoken Dialogue," *Proceedings of 10th Eurospeech-Interspeech Conference*, 2007.

[2] Š. Beňuš, A. Gravano, and J. Hirschberg, "Prosody of backchannels in American English," Proceedings of 16th International Congress of Phonetic Sciences, pp. 1065-1068, 2007.

[3] H.H. Clark, *Using language*. Cambridge University Press, Cambridge, 1996.

[4] T. Stocksmeier, S. Kopp, and D. Gibbon, "Synthesis of prosodic attitudinal variants in German backchannel ja," *Proceedings of Interspeech*, pp. 1290-1293, 2007.

[5] Š. Beňuš, A. Gravano, and J. Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, 43 no. 12, 3001-3027, 2011.

[6] M. Heldner, J. Edlund, and J. Hirschberg, "Pitch similarity in the vicinity of backchannels. *Proceedings of Interspeech*, 2010.

[7] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, pp. 578-583, 2015.

[8] Š.Beňuš, R. Levitan, J. Hirschberg, A. Gravano, and S. Darjaa, "Entrainment in Slovak collaborative dialogues," *Proceedings of the 5th IEEE Conference on Cognitive Infocommunications*, pp. 309-313, 2014.

[9] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: A cross-linguistic comparison. *Proceedings of 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 325-334, 2015.

[10] N. Ward and W. Tsukahara, "Prosodic features which cue backchannel responses in English and Japanese," Journal of Pragmatics 32, no. 8, pp. 1177–1207, 2000.

[11] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language* 25, no. 3, pp. 601-634, 2011.

[12] S. Darjaa, M. Cerňak, M. Trnka, M. Rusko, and R. Sabo, "Effective triphone mapping for acoustic modeling in speech recognition," *Proceedings of Interspeech*, 2011.

[13] A. Gravano, J. Hirschberg, and Š. Beňuš, "Affirmative cue words in task-oriented dialogue," Computational Linguistics 38(1), pp. 1-39, 20012.

[14] P. Boersma and D. Weenink, Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

[15] J. Harrington, *Phonetic Analysis of Speech Corpora*. Willey-Blackwell, Oxford, 2010.

[16] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-Prosodic Entrainment and Social Behavior," *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012.

[17] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," Biometrical Journal 50, no. 3, pp. 346–363, 2010.

[18] M. Heldner, J. Edlund, K. Laskowski, and A. Pelcé, "Prosodic features in the vicinity of pauses, gaps and overlaps," *Nordic Prosody – Proceedings of the Xth Conference*, pp. 95 – 106, 2009.

[19] P. Healey, M. Purver, and C. Howes, "Divergence in dialogue. PloS one, 9 no. 6, e98598, 2014.