Prosodic forms and pragmatic meanings: the case of the discourse marker ‘*no*’ in Slovak

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

Abstract—**This paper provides a first analysis of the Slovak word [no] that is extremely ambiguous in terms of pragmatic and discourse functions and can correspond to ‘okay’, ‘yes’, ‘well’, ‘mhm’, ‘so’, and others. We report that the function of backchannel/continuer is the most easily disambiguated by the pitch contour, duration, and other features. Other functions, while also frequent in our corpus, require more sophisticated multi-factor analyses for identifying best disambiguating features.**

I. Introduction

If spoken human-machine interactions are to be perceived by humans as natural, they most probably should be based on the systematic aspects of human-human spoken interactions. Some of the most common communicative social signals in human-human spoken interactions are feedback vocalizations. These backchannels, acknowledgment tokens and other cue words are important for building common ground understanding, space-time orientation, and other crucial aspects of successful communication. Moreover, they are indispensible if the listeners are not to be perceived as passive receivers of information but as supportive, cooperative agents required for any natural and meaningful conversation. Furthermore, these vocalizations offer insights into the real-time responsiveness within human-human interactions. They are also typically very short, prosodically well-defined by surrounding silent pauses, and they are thus relatively easy to study using the recognition/synthesis tools of natural speech processing (such as analysis-by-synthesis). Finally, they are typically syntactically detachable from a sentence, commonly used in initial positions within utterances, and have a range of prosodic contours.

The problem, however, is that these vocalizations are commonly ambiguous potentially cuing several pragmatic functions. Humans are very good at disambiguating these multi-faceted relationships between the acoustic/prosodic realizations of these feedback vocalizations and their socio-pragmatic meanings in a given conversational situation. Machines, on the other hand, lack behind in being able to decode, understand, and act upon these communicative social signals. Partially for these reasons, human-machine interactions are commonly perceived by humans as stilted, very limited, and machines are ultimately taken as sub-par “communicators”.

Feedback vocalizations or affirmative cue words have received relatively great attention especially in common languages like English or Swedish; e.g. [1], [2], [3] among many others. One of the fundamental findings well documented in many studies is that giving feedback is not considered anymore as an activity of only one conversational partner (usually the listener), but that it is a part of the collaborative process in which both speaker and listener coordinate their contributions to ensure mutual understanding [4].

Additionally, the relationship between the form and the meaning of these phrases is very complex. For example, some of these vocalizations are extremely ambiguous, for example [5] identified as many as ten pragmatic functions of *okay* in collaborative tasks in American English. On the other hand, the list of potentially pragmatically meaningful vocalizations and grunts is rather extensive [6]. Hence, the relationship between multiple meanings and multiple forms of these conversational tokens is a fruitful research area in quest for better understanding of the rapport of human-human conversations.

Another fundamental observation is that feedback is primarily used for conversational grounding by which speaker and listener, in a collaborative and coordinated way, establish mutual understanding both of information presented in the course of exchange as well as information assumed to be shared even prior to the exchange [7], [8].

Yet another domain in which feedback vocalizations have been studied is discourse structure. As reviewed in [9], cue phrases may provide so called *contextual coordinates* for an utterance in the discourse, by which these phrases display the discourse structure and relations within this structure [10]. This approach was later extended in [11] who proposes that cue phrases bring to the listener’s attention a particular kind of relationship that the upcoming speech has with the immediate discourse context. Hence, such sue phrases, or discourse markers, thus not only diplay the discourse structure but play a prominent role in creating it.

In this paper we present the first investigation of discourse markers in Slovak. The advantage of our approach is two-fold.

First, we concentrate on the Slovak word *no*. As it represents a shortening of *áno*, which means ‘yes’ in Slovak, it typically can signal many functions of *okay* identified in [5] such as backchannel, acknowledgment, beginning of a new discourse segment, or agreement. Moreover, Slovak *no* also functions as a conjunction in the meaning of ‘but’ and when functioning as a discourse marker it can take similar meanings to ‘but’ or ‘well’. Hence, in addition to affirmative meaning, it can also signal non-commitment and mild disagreement. Slovak thus allows extending the group of typically affirmative meanings associated with feedback and cue words to the dimension of polarity.

The second advantage is that Slovak data, that are recorded in a similar manner to [1] and [4], provide a wonderful test case for cross-linguistic research. Cross-linguistic and cross-cultural understanding of the human cognitive system of interactional communication is crucial for identifying key aspects of human-human interactions that should have a priority for implementation into human-machine voice-activated communication system. For example, in English, *yes* is both minimally ambiguous and more formal. *Okay* is more ambiguous than *yes* but in terms of the degree of formality not very different from *yes*. On the contrary, in a similar opposition in Slovak, *áno* is more transparent and formal but *no* is less transparent (i.e. more ambiguous) but also clearly more colloquial and informal.

TABLE I.
RATES AND RANKS OF AFFIRMATIVE CUE WORDS IN THE CORPUS

Cue word	Gloss	Frequency (%)	Rank
no	okay/yes/well/ mmhm/but	2.98	2
mhm	mmhm	2.0	7
hej	yes	1.13	11
dobre	good/well/okay	0.94	17
uhhuh	uhhuh	0.49	30
áno	yes	0.43	40

II. METHODOLOGY

A. Corpus

Data for this paper come from a newly recorded corpus of task-oriented dyadic collaborative conversations in Slovak. Speech was recorded while participants played a computer game designed to elicit conversation that was adapted from the OBJECT Games described in [12], [13], or [1]. Briefly, interlocutors were seated in a sound-treated quiet room with computer screens and keyboards so that they did not have any visual contact but could hear each other. One player described the position of a target object with respect to other objects on her screen, and the other tried to move the same object to the same position on his own screen. Subjects were encouraged to match the positions perfectly and points were awarded on a 100-point scale based on how closely the pixel-positions of the two object matched. Each game included 14 tasks of object placing and the roles of the person describing the position and the role of placing an object were equally divided between the two players.

The current corpus consists of 6 games played by 7 subjects (3 females 4 males) so that 5 subjects played the game twice (with a different partner) and 2 male subjects played only one game. The corpus contains almost four hours of speech (3h, 54m), there are 21773 words in total, and 2371 unique words.

Table I shows the rates and ranks of the most common affirmative cue words (ACW) in the corpus. We see that in total, roughly 7% of all words function as ACWs and individually, ACWs belong among the most common words in this corpus like *je* ‘is‘ (Rank 1), *a* ‘and‘ (Rank 3), *uh* (Rank 6), or *tam* ‘there‘ (Rank 13).

TABLE II.
LABELING SCHEME FOR ANNOTATING THE FUNCTIONS OF “NO”

Label	Meaning
R	I acknowledge that I understand, I got it
RP	I acknowledge that I understand, and please continue
RZ	I acknowledge that I understand, but I want to add something or express mild disagreement
RN	I acknowledge that I understand, and I want to start a new topic or a new discourse segment
N	I want to start a new topic or a new discourse segment
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*
H	Hesitation, I am stalling for time
E	I want to repair/redo something I’ve just said or did
PH	Express assessment of something that just happened, usually on receiving a score
J	Softening of what is to follow, a hedge
K	Signal the end of the current topic or discourse segment
D	Encourage some action, go on, do something
?	None of the labels correspond to the meaning I perceive

B. Labeling

The recorded speech was manually transcribed and inter-pausal units (IPUs) were determined in such a way that each roughly 150ms pause not associated with producing plosives were identified. The transcribed speech was then manually aligned to these IPUs. The alignment of individual words with the acoustic signal is currently under way.

As can be seen from Table I, word *no* is very common, and at the same time, it presents a rather extreme case of pragmatic ambiguity since it can stand for multiple functions. Analyzing a subset of the corpus, we designed a labeling scheme for the most common discourse and pragmatic functions that the word *no* can express. This scheme is shown in Table II. The majority of functions are similar to various functions of *okay* in American English as analyzed in [4] or [14]. However, several functions are different, most notably those labeled as Z, H, J, or RZ. We found that one of the most challenging functions is J that is commonly prosodically aligned into a greater intonational unit with the following material and in which *no* represent a prosodic clitic.

At the current stage, a single experienced annotator labeled all occurrences of *no* in the corpus using Praat graphical user interface [15]. The annotator could listen to the context of any length and replay speech as many times as required.

C. *Feature extraction*

Praat was also used for extracting most common features such as duration, mean, maximum, minimum of F0 and Intensity, as well as voice quality features such as jitter, shimmer, harmonics-to-noise ratio or spectral tilt. These features were then normalized using Z-score; each value minus mean divided by standard deviation while means and standard deviations were based on the no-tokens in each session and speaker in the corpus.

III. Results

A. *Descriptive observations*

There are 644 tokens of *no* in the corpus, hence on average 107 for each session. Speakers varied in their token frequencies between 17 for the lowest and 159 for the highest rates. Fig. 1 shows the frequencies of the functions listed in Table II.

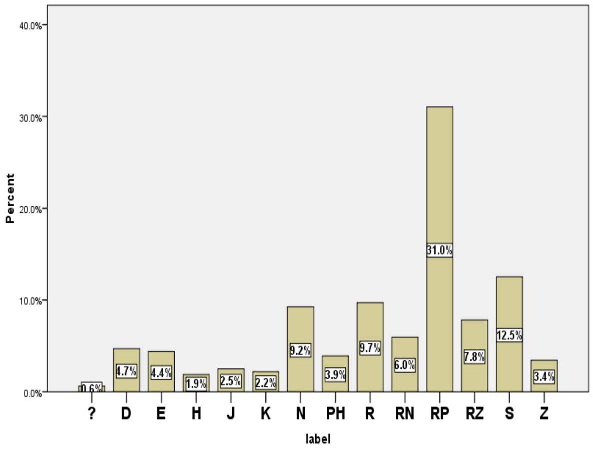

Figure 1 Distribution of all functions on *no* from Table II in the corpus of Slovak collaborative tasks

We see that the backchannel/continuer function is by far the most frequent with 31% followed by agreement, acknowledgement functions and cue beginning a new discourse segment. Token associated with slightly negative polarity (RZ and Z) are less frequent but their rates are comparable to other functions.

Fig. 2 provides exploratory information on the F0 contours associated with the four most frequent *no* tokens using and adaptation of the technique introduced in [16]. In this bitmap technique, contours of F0 from all tokens were extracted and plotted with partially transparent dots. Hence, the more dots occur at a particular point, the darker it is and the accumulation of these dots from multiple instances forms a density cloud. F0 contours were normalized to 100 points and transformed to semitones.

Figure shows that the backchannel/continuer function is signaled by a clear fall-rise contour. This is similar to findings for American English in e.g. [14]. Interestingly, the four functions seem to produce a continuum between clearly falling-rising contour of backchannels/continuers through more plateau contours of the middle two panels (agreements and acknowledgments) towards somewhat falling contours of cues for new discourse segment. Moreover, agreements (S) seem to be more similar to backchannel/continuers (RP) than plain acknowledgments (R) do. This is slightly surprising since RP function includes partially R functions while S is mostly independent from RP.

A similar plot for less frequent functions is not very informative since most of the functions seem to have a plateau contour and pitch contours thus seem to provide minimal cues for disambiguating these functions. This includes both functions with slight negative polarity labeled as Z and RZ. A very problematic function J is typically produced with a falling contour.

B. *Initial quantitative findings*

Given the skewed distribution of the functions shown in Fig.1 we decided to test if the most frequent function of backchannel/continuer differs in prosodic or voice quality features from other functions. Fig. 3 illustrates the most robust findings. Welch Two Sample t-tests showed that tokens of *no* that function as backchannels/continuers are significantly longer, slightly lower, less loud and produced with greater spectral tilt than the tokens corresponding to other functions.

Ref [4] also found that prosodic features were very useful for identification of the cue beginning function of *okay*. Hence, we tested if, in addition to the F0 contour difference identified in Fig. 2, there are other prosodic or voice quality features signaling this function in Slovak. Interestingly, none of the analyzed features could

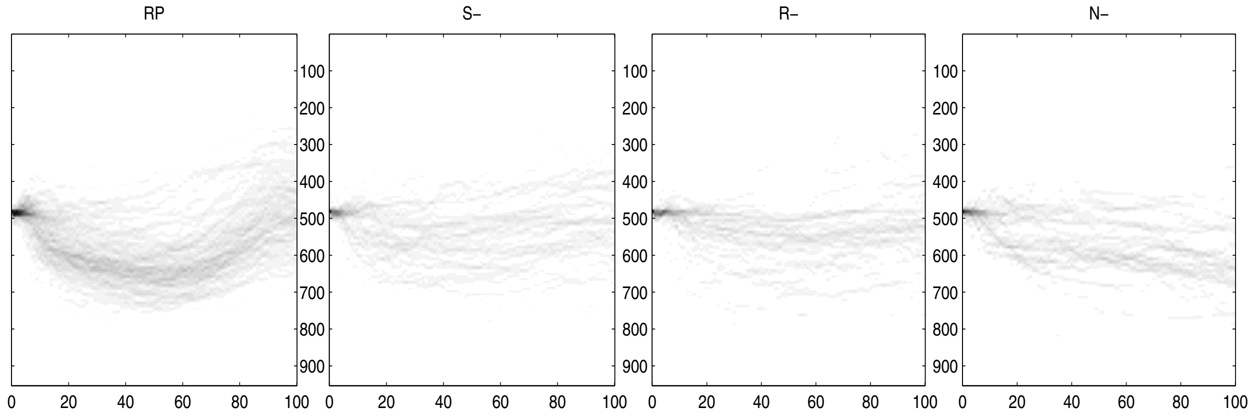

Figure 2. Density clouds for normalized F0 contours in four most frequent functions of no (see explanation of labels RP, S, R, and N in Table II). The units on the y axis are not meaningful

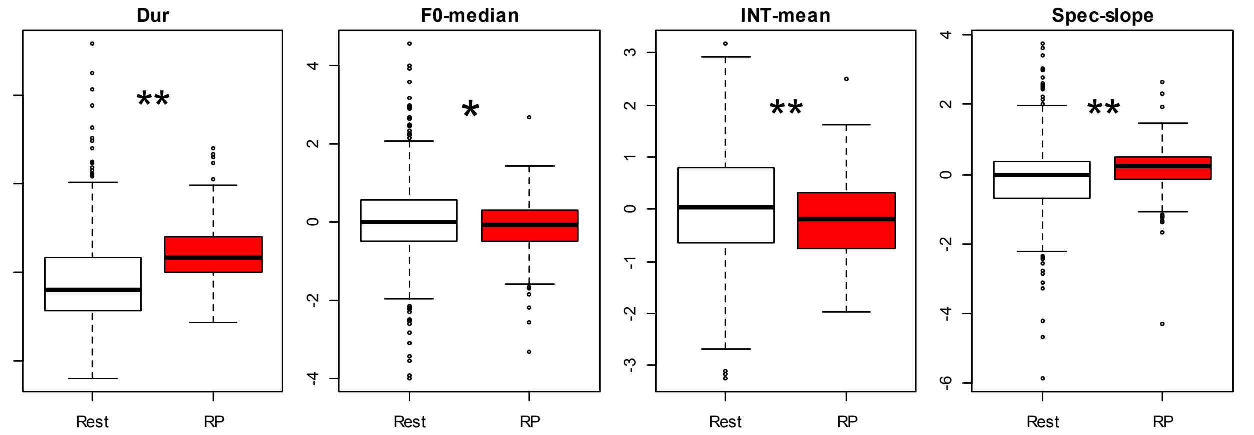

Figure 3 Z-score normalized values (from the left) for the duration, F0 median, Intensity mean and spectral slope of no-tokens divided to backchannel/continuer function (RP) vs. all other functions (Rest). "*" refers to t-test significant at $p < 0.05$ and "**" significance at $p < 0.01$

discriminate between this cue beginning function and the pooled remaining functions.

Finally, let us look at the peculiarity of Slovak *no* tokens in comparison to English *okay* in terms of the presence of a discourse/pragmatic meaning with slight negative polarity. We found that median F0 was higher for these tokens than for the rest of the functions and that all intensity measures also showed significantly higher values. Fig. 4 illustrates these findings with the most robust measures of F0 and intensity.

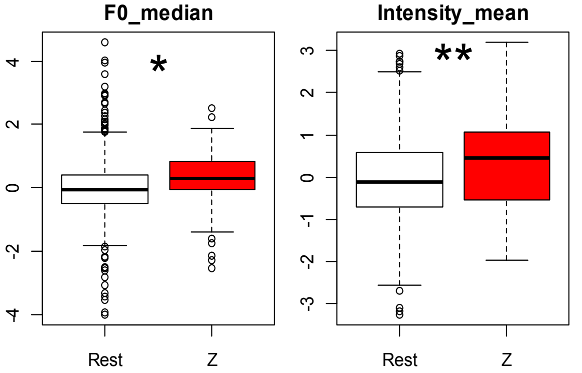

Figure 4 Z-score normalized values for F0 median (left) and Intensity mean (right) of no-tokens divided to Z/RZ functions (Table II) vs. all other functions (Rest). "*" refers to t-test significant at $p < 0.05$ and "**" significance at $p < 0.01$

IV. Discussion and Conclusion

We provided a first preliminary look at the relationship between the prosody and discourse/pragmatic meanings of Slovak feedback vocalization corresponding to the word *no*. We identify relatively salient features for the identification of the backchannel/continuer function related to duration, pitch, intensity, and spectral slope. Somewhat less salient but still significant finding was related to identifying functions with mild negative polarity.

However, more sophisticated clustering or multi-factor techniques will be required for determining if other prosodic and voice quality features are useful for the disambiguation of other functions. Furthermore, we plan to provide a more objective function labeling by involving more annotators and work with majority votes as well as collecting more data. Ultimately, we hope to provide a systematic description of the relationship between the discourse/pragmatic meaning and prosodic realization of all major feedback vocalization in Slovak so that this systematic knowledge might be used in applications utilizing human-machine communication, for example in dialog systems or management of crisis situations.

To conclude, the variability of Slovak *no* tokens provides a fruitful field of research both in terms of cross-linguistic and cross-cultural comparison among languages (e.g. between Slovak and American English) as well as in terms of deeper understanding of feedback vocalizations as one of the most fundamental communicative social signals in spoken conversations.

Acknowledgment

This work results from project implementation: Research and development of new information technologies for forecasting and mitigation of crisis situations and safety, ITMS 26240220060 supported by the Research & Development Operational Programme funded by the ERDF.

References

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

[2] D. Jurafsky, E.Shriberg, B. Fox, and T. Curl, "Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL/COLING, Workshop on Discourse Relations and Discourse Markers*, pp. 114–120, 1998.

[3] J. Allwood, J. Nivre, and E. Ahlsen, "On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1), pp. 1–30, 1992.

[4] J. Bavelas, P. De Jong,H. Korman, S. Smock Jordan, " Beyond Back-channels: A Three-step Model of Grounding in Face-to-face Dialogue," Proceedings of Interdisciplinary Workshop on Feedback Behaviors in Dialog, pp. 5-6, 2012.

[5] A. Gravano, Š. Beňuš, J. Hirschberg, S. Mitchell, I. Vovsha, "Classification of Discourse Functions of Affirmative Words in

Spoken Dialogue,“ Proceedings of 10th Eurospeech-Interspeech Conference, 2007.

[6] N. Ward, “Non-Lexical Conversational Sounds in American English,“ *Pragmatics and Cognition*, 14 (1), pp. 113–184, 2006.

[7] H. H. Clark, and E. F. Schaefer, “Collaborating on contributions to conversations,” *Language and Cognitive Processes*, 2(1), pp. 19-41, 1987.

[8] H. H. Clark, and E. F. Schaefer, “Contributing to discourse,” *Cognitive Science*, 13, pp. 259-254, 1989.

[9] B. Fraser, “What are discourse markers?“ Journal of Pragmatics, 31(7), pp. 931– 952, 1999.

[10] D. Schiffrin, “*Discourse Markers*,“ Cambridge University Press, Cambridge, 1987.

[11] G. Redeker, “Review article: Linguistic markers of linguistic structure, “ *Linguistics*, 29(6), pp. 1139–1172, 1991

[12] A. Gravano, “Turn-taking and affirmative cue words in task-oriented Dialogue,“ PhD thesis, Columbia University, 2009.

[13] A. Gravano, and J. Hirschberg, “Turn-taking cues in task-oriented dialogue”, *Computer Speech and Language*, vol. 25, pp. 601–634, 2011.

[14] Š. Beňuš, A. Gravano, J. Hirschberg, “Prosody of backchannels in American English,“ Proceedings of 16th International Congress of Phonetic Sciences, 2007.

[15] P. Boersma, D. Weenink, “Praat: Doing phonetics by computer”, http://www.praat.org, 2001.

[16] M. Heldner, J. Edlund, K. Laskowski, A. Pelcé, “Prosodic features in the vicinity of pauses, gaps and overlaps,“ In Vainio, M., Aulanko, R., Aaltonen, O. (Eds.), Nordic Prosody – Proceedings of the Xth Conference, pp. 95 – 106, Frankfurt am Main: Peter Lang, 2009.