Between Recognition and Resignation – The Prosodic Forms and Communicative Functions of the Czech Confirmation Tag "jasně"

Jan Volín,[1] *Lenka Weingartová*,[1] *Oliver Niebuhr*[2]

[1]Institute of Phonetics, Charles University in Prague, Czech Republic
[2]Department of General Linguistics, ISFAS, Christian-Albrecht-University of Kiel, Germany
{jan.volin|lenka.weingartova}@ff.cuni.cz, niebuhr@isfas.uni-kiel.de

Abstract

Like question tags, confirmation tags such as the Czech affirmative particle *jasně* can be used with various prosodic characteristics that augment, reverse or otherwise modify their relatively unspecific lexical meaning. We extracted 172 instances of *jasně* from several dialogues and assessed their discourse function. 36 prosodic correlates in temporal, amplitude and fundamental frequency domains were measured and used in three computational classifiers: linear discriminant analysis, classification trees and artificial neural networks. All three methods significantly reflected the functional assessments and additionally indicated the relative importance of individual predictors in a mutually consistent manner.

Index Terms: affirmative particle, confirmation tag, Czech, discourse, intonation, pragmatics.

1. Introduction

It is obvious and has been repeatedly shown for many languages that 'question tags' like *isn't it* in English, *nicht wahr* in German, or *verdad* in Spanish are very rich both phonetically and functionally, cf. [1,2,3,4,5,6,7]. They are used by speakers to keep the interaction going and/or promote the flow of information. At the same time, the lexical semantics of question tags is fairly unspecific. Taken together, this allows question tags to occur in very different communicative contexts and in combination with all kinds of prosodically expressed speaker attitudes and emphatic intensifications.

The same also applies to 'confirmation tags' like *of course* in English, *alles klar* in German, or *todo bien* in Spanish. Confirmation tags, which, if they are single words, can also be called affirmative particles, are moreover everything but rare in conversation. Their high frequency in combination with their flexible application and constant segmental basis make confirmation tags – just like the better investigated but probably rarer question tags – an ideal research subject for studying the prosodic forms of a language and their respective communicative functions. In this context, the present study deals with the disyllabic Czech confirmation tag *jasně* ['jas.ɲe] whose closest English equivalents would be *sure*, *agreed*, *of course*, or *fair enough*. Speakers ordinarily insert *jasně* at the beginning of their utterances, typically as a separate prosodic phrase, in order to react to a preceding turn of the interlocutor.

Our major aim is to determine, describe, and systematize the prosodic and functional variation that can occur on *jasně*, in this way also advancing our understanding the prosodic system of Czech in general. While particularly the phonological factors and prosodic correlates of lexical stress as well as the related phenomena of phrasing and rhythm have been intensively analyzed for Czech in the last decades (cf. [8,9,10,11,12,13]), only relatively little is known about the use of intonation and emphasis patterns in Czech. Unlike research on the prosody of emphatic expressions in Czech, which does virtually not exist, there are some studies on Czech intonation. However, the corresponding research has so far often been descriptively oriented in the sense that intonational forms and functions have been characterised and contrasted on the basis of exemplars, impressions and experience, or with the primary aim to develop annotation inventories and enhance speech technology applications, cf. [14,15,16]. Analyses that aimed at a detailed, empirically based understanding of intonational forms, functions, and their linkages have only just come up in the last few years, cf. [17]. Our paper follows this more recent, empirical line of research.

This paper will summarize the production part of our study. The production data come from a large corpus of enacted (i.e., text-based) dialogues conducted by 30 native speakers of Czech. The data were acoustically analyzed in terms of a number of different duration, F0, and intensity measures. The production part will soon be complemented by a perception part, serving to cross-validate the form-function links that emerge on Czech *jasně*. Both the communicative functions and the acoustic-prosodic parameters on which our analysis of *jasně* is based were inspired by German whose intonational and emphatic categories and structures have been thoroughly explored in the last decade, cf. [18,19,20,21,22]; and it is probably not exaggerated to state that intonation and emphasis structures in German are already fairly well understood.

Against this background, the specific questions that we address here are the following:

(1) Do we find systematic prosodic variation on Czech *jasně*?

(2) If the answer to (1) is positive, is this variation functionally motivated, i.e. meaningful? Or is the variation just contextually motivated and due to speaking rate, phrase structure, or speaker-specific effects?

(3) If the answer to (1) is positive, is this variation multiparametric or rather dominated by a single prosodic parameter?

The answers to these questions will later allow to put the issue into a cross-linguistic perspective. For example, the absolutely strict lexical stress position in Czech reduces the corresponding functional load of duration and/or intensity so that these parameters could even play a more important role in signalling emphasis categories than in German. If this is the case, will the respective prosodic patterns be still associated with the same communicative functions as in German?

Three classifiers will be used to gain a cross-evidenced view of the variables, of which some map very similar properties as the others differing only in conceptual detail (see below). This will provide a methodological advantage for further research.

10.21437/SpeechProsody.2014-11

2. Method

A total number of 180 *jasně* tokens from the Prague Phonetic Corpus [23] were used. The target word occurred in six different contexts in the corpus and each was uttered by 30 native Czech non-professional speakers, 24 female, 6 male, aged 20-25 years. Scripted texts were used to elicit short dialogues from pairs of speakers. The speakers were explicitly encouraged to familiarize themselves with the dialogues and then act them out as convincingly as possible. The participants got along with the task very well, taking various affective approaches. Nonetheless, two trained phoneticians, who controlled the recording process, asked for new trials when dysfluencies or unnatural renderings occurred.

The recordings were made digitally at a sampling rate of 32 kHz and with a 16-bit quantization in the sound-treated studio of the Institute of Phonetics in Prague, using an IMG ECM2000 microphone and a SB Audigy 2ZS soundcard.

2.1. Perceptual categories

First, all 180 target word tokens were surveyed on an auditory basis by three Czech trained phoneticians (two of whom were authors of this paper). This auditory survey in combination with the phoneticians' native-speaker intuitions led to setting up eight functional categories:

Type 1: neutral acceptance

Type 2: eager agreement

Type 3: impatience

Type 4: indifference, patronizing

Type 5: wonder, surprise

Type 6: recognition, realizing

Type 7: resignation

Type 8: reassurance, sympathy

Having set up these categories, each target word was listened to and assigned to one of the categories. The assignment procedure was conducted independently by the three phoneticians. In the case of disagreement the respective token was discussed and the majority vote was taken. In the end, eight tokens had to be discarded due to disagreement of all three listeners, so that 172 words were left for further analysis.

2.2. Acoustic measurements

Acoustic analyses of the target words were carried out in *Praat* [24], individual segment boundaries were manually labelled. The following parameters were measured:

Temporal:

- word duration (in ms)
- relative segment duration (in % of word duration, and in % of syllable duration)
- relative syllable duration (in % of word duration)
- difference between the duration of syllable nuclei (in ms, [a]-[e])
- difference between the duration of syllable onsets (in ms, [j]-[ɲ])
- durational profiles: the outcome of a cluster analysis (4 clusters, k-means) where the individual cases (renderings of the word) were clustered according to their segment durations (in % of word duration)

F0:

- first and second extreme of the F0 contour (i.e., maximum and minimum or vice versa) normalized to speaker range (in %) and speaker average (in ST)
- the difference of the first and second extreme in the F0 contour (in ST re 100 Hz)
- the difference between vowels, i.e., between the F0 mean values taken in the middle third of each vowel (in ST)

Speaker range and average values for normalization were taken from all six utterances in which the target word occurred, rather than just from the target word itself. Errors in F0 extraction were manually corrected, and portions of the signal with creaky voice were excluded, so that we obtained an estimate of the speaker's modal range. Values of the minima and maxima in the target word were measured manually, F0 micro-perturbation was disregarded. Creaky voice in the target words was subsequently set to be at 0 % of the speaker's range rather than at negative values, since this has improved the discriminatory power of the variable in preliminary analyses and most probably reflects the speaker's intention of hitting 'ultimate low' rather than a specific frequency target.

Energy:

- maximum SPL value in the target word, normalized by average utterance SPL (in dB; pauses and silences were excluded)
- location of the SPL maximum (in % of word duration, and in the corresponding segment)
- SPL in the middle of each segment, normalized by average utterance SPL (in dB)

Apart from these acoustic measurements, the intonation contour was also annotated by the third author with labels adapted from the Kiel Intonation Model [20,25]:

- prominence strength of each syllable in three levels (0 = no prominence, 1 = weak prominence, 2 = strong prominence)
- synchronization of the pitch-accent peak (early, medial, late) in weakly or strongly prominent syllables
- final boundary tone (0 = flat, 1 = moderately descending, 2 = falling to the lower end of the speaker's range)

Afterwards, the position of the (more) prominent syllable in the disyllabic target word and its and pitch-accent synchronization were merged into a single contour-descriptor label (e.g., 'FA' = early peak on the first syllable; 'MB' = middle peak on the second syllable).

All 36 parameters listed above were then used as variables in subsequent statistical analyses.

The discriminative strength of each variable was explored through one-way ANOVAs. For classifying the data into the eight perceptual categories, linear discriminant analysis (LDA), classification and regression trees (CART) and artificial neural nets (ANN) were used. The advantage of using CARTs and ANNs is the possibility to employ both continuous (e.g., temporal or F0 parameters) and categorical (e.g., duration profiles, intonation labels) variables as predictors. Moreover, CART can use one and the same variable repeatedly at different split decisions. All the classifiers used were from the STATISTICA software package [26].

3. Results

3.1. Counts of Tokens in Functional Categories

Out of the 172 investigated tokens, 43 cases were assigned to Type 2 (eager agreement) and 42 cases to Type 1 (neutral acceptance). Type 6 (recognition, realizing) was represented by 29 cases, Types 4 (patronizing) and 5 (wonder) both by 18 cases, while for Types 7 (resignation), 8 (reassurance), and 3 (impatience) only 9, 8 and 5 cases, respectively, were found.

3.2. Discriminant Analysis

Linear discriminant analysis was performed after searching for continuous variables that do not correlate too highly with each other and differentiate well among the individual functional categories. Unrestrained analysis (i.e., mapping the structure of the dataset with rather loose tolerance levels) resulted in a success rate of 57.6 % and identified the following variables as best reflecting the assumed functional categories: duration of the vowel /e/ in the second syllable of *jasně* relative to the word duration, vowel /e/ duration relative to syllable duration, consonant /s/ duration relative to syllable duration, duration of first syllable relative to word duration, durational difference between vowels, durational difference between syllable onsets, and F0 difference between the first and second extreme.

Several further analyses were performed with more stringent tolerance levels. Although the success rate for the best outcome dropped to 52.3 %, the results can be considered more generalizable. Only five variables were ultimately used: word duration, durational difference between vowels, durational difference between syllable onsets, F0 difference between vowels and normalized value of the first F0 extreme. Table 1 displays the ensuing confusion matrix. It is apparent that under-represented functional categories (Type 3, 7 and 8, i.e., impatience, resignation and reassurance respectively) were not recognized in this more rigorous setting of the LDA.

LDA		Observed Types							
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Predicted Types	*1*	22	11	0	4	1	2	1	1
	2	15	28	2	2	0	7	2	1
	3	0	0	0	0	0	0	0	0
	4	2	1	1	9	5	0	0	4
	5	2	0	0	1	10	0	0	1
	6	0	3	2	2	0	20	6	0
	7	0	0	0	0	1	0	0	0
	8	1	0	0	0	1	0	0	1
Corr. %		**52**	**65**	**0**	**50**	**56**	**69**	**0**	**12**

Table 1. *Confusion matrix resulting from the most successful discriminant analysis. Figures represent individual cases, except in the last line with percentages of correctly recognized cases within a category.*

The success rate for other categories was 50 % and more. The highest numbers off the diagonal can be found for Types 1 and 2 (neutral acceptance and eager agreement). They seem to be highly confusable, although the correctly identified cases in these two abundant types still prevail. The most distinct functional category seems to be Type 6 (recognizing) with nearly 70 % of the cases correctly separated from other categories and with errors towards Types 1 and 2 again. As stated above, the lowest success was achieved for the smallest groups, of which Type 3 was represented by 5 instances only.

3.3. Classification and Regression Trees

The algorithm used in STATISTICA calculates automatically the usefulness of all the input variables and ranks them according to their effectiveness in the classification process. The most successful tree achieved the success rate of 65.7 %, which is by about 10 % more than in our earlier discriminant analyses. The best tree had 9 splits and was based on word duration (ranked as the most important predictor), F0 difference between vowels, durational difference between syllable onsets, relative duration of a syllable within the word, normalized F0 value of the first extreme and normalized intensity of the first vowel. Other intensity measures and categorical labels of intonation and temporal profile were found unimportant, while the word duration and F0 difference between vowels were used twice, i.e. for two different splitting decisions. The ensuing confusion matrix is shown in Table 2. Further splitting could still be ordered, but only at the expense of generalizability, hence we did not proceed with it. The number of confusions between Types 1 and 2 is lower than in previous analyses, but a considerable number of Type 2 cases were misclassified as Type 6 (see below, Table 2, the second numbered column).

CART		Observed Types							
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Predicted Types	*1*	21	4	0	0	0	0	0	0
	2	12	28	2	0	0	0	1	0
	3	0	0	0	0	0	0	0	0
	4	1	0	0	11	0	0	0	1
	5	3	0	1	0	18	0	0	0
	6	2	11	2	4	0	29	8	1
	7	0	0	0	0	0	0	0	0
	8	3	0	0	3	0	0	0	6
Corr. %		**50**	**65**	**0**	**61**	**100**	**100**	**0**	**75**

Table 2. *Confusion matrix resulting from the most successful CART analysis. Figures represent individual cases, except in the last line with percentages of correctly recognized cases within a category.*

The best classification of categories was achieved for Types 5 and 6 (wonder and recognizing) whose all instances were correctly found. However, some other types were also mistakenly added to these categories. From this point of view, Type 5 seems to be better as only four improper cases were added to it. The rare Types 3 and 7 (impatience and resignation) were not recognized at all, but Type 8 (reassurance), which was represented by 8 cases in our dataset was classified relatively successfully.

3.4. Artificial Neural Nets

Thirty different architectures and settings were tried always with eight output neurons (corresponding to eight functional categories). The initial analyses used all the variables available as the input with the aim to evaluate of their individual usefulness. Sub-sequent analyses only utilized the most effective predictors. It turned out that Multilayer Perceptron Neural Networks outperformed other available types (RBF, LNN).

ANN		Observed Types							
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Predicted Types	*1*	30	11	1	2	1	3	1	1
	2	6	27	1	1	2	3	0	1
	3	1	0	2	1	1	1	0	0
	4	2	1	0	9	0	1	0	2
	5	2	1	0	3	14	0	2	2
	6	0	3	1	2	0	20	2	0
	7	0	0	0	0	0	0	4	1
	8	1	0	0	0	0	1	0	1
Corr. %		**71**	**63**	**40**	**50**	**78**	**69**	**44**	**12**

Table 3. *Confusion matrix resulting from the most successful ANN classification. Figures represent individual cases, except in the last line with percentages of correctly recognized cases within a category.*

As in the previous analyses, the most efficient variable was the duration of the word. Rather surprisingly, the second best was the durational profile of the word (a categorical variable), followed by the prominence strength on the first syllable (categorical), durational difference between syllable onsets, F0 difference between vowels, synchronization of the F0 peak (categorical), and normalized intensity of the first vowel. The automated neural networks that we used weigh some of the input variables by zero and make them ineffective. As a result, they do not suffer from dimensionality problems. The unrestrained model achieved a success rate of 66.3 %. When only the nine best variables were used in a three-layer perceptron architecture, the success rate dropped to 62.2 %, but probably with the advantage of better generalizability. Confusion matrix of the final analysis is displayed in Table 3.

Unlike in CART analyses, there are no 0 % or 100 % success rates in mirroring the functional categories. Type 5 (wondering) and Type 1 (neutral acceptance) were the best recognized with the success rates over 70 %. Some correct classification occurred even in the small groups of Type 3, 7 and 8 which were previously found difficult to capture (apart from Type 8 in CART analysis).

4. Discussion

Three classifiers performed their analyses with comparable levels of success. However, the lowest success rate in the case of the conventional discriminant analysis suggests that continuous linear relationships do not model prosodic dependencies best. As noticed in the past, acoustic correlates of prosodic features are used in various combinations, and the same features can be used for different communicative functions, in this way creating discrete 'islands' in a multidimensional space. If this is true, more advanced classifiers should be advantageous. More specifically, the best recognized categories, Types 5 and 6 (wondering and recognizing), were each found by CART at two different endpoints of the classification tree. This supports the idea that the same pragmatic or discourse effect can be achieved through different prosodic means. One way or another, our results allow for positive answers to the first two questions from the introduction: the prosodic variation in our data set appears to be systematic and functionally motivated.

The third question concerned the variables responsible for prosodic profiling of the individual functional categories. The word duration as an expression of the articulation rate was identified as a useful discriminatory element in all analyses performed. It seems that the rapidity (or slowness) with which the word *jasně* is pronounced is a reliable marker of the appended functions. Various other durational characteristics kept reoccurring as well, of which the most important one was the difference in duration of the consonantal onsets of the syllables. Interestingly, auditory inspections turned our attention to the duration of the word-initial consonant, which was markedly longer for some functional categories than for others, but the duration of this consonant relative to the duration of the word was computationally less useful than the same duration relative to the duration of the second syllable onset. Local durations of consonants thus might function in speech by being contrasted against each other rather than by being compared with the carrier unit as a whole.

As to the melodic correlates, the one repeatedly occurring as effective was the difference between F0 means measured in the middle thirds of the vowels (in ST). This variable could be sometimes replaced with the difference in F0 extremes within the word with a few percent shifts in the success rates. The relative pitch of the first syllable – expressed as either the F0 value within the speaker's range or as the annotated labels adapted from the Kiel Intonation Model – were also found relevant by the computational classifiers, although they were not eventually utilized in the most successful models.

The profiles of functional categories to be further examined in perceptual tests appeared to be as follows. Types 1 and 2 were spoken significantly faster than all the other types. It seems plausible to find neutral and eager stances brisk, whereas patronizing, wonder, realizing, resignation, and reassurance spoken more slowly. The major discriminator between Types 1 and 2 was then the difference in duration of the syllable onsets. Type 2 (eagerness) has significantly longer the word-initial consonant. A similar relationship is found between patronizing (short word-initial consonant) on the one hand, and wondering and realizing on the other hand (longer word-initial consonant). Melodically, Type 5 (wonder) was the only one with clearly rising F0 contour. Patronizing (Type 4) and reassurance (Type 8) were spoken with flat contour, while the rest of the types had falling melodies. As to energy, Types 1, 2 and 6 exhibited high SPL in first (i.e, stressed) vowel, whereas Types 4, 5 and 8 low SPL.

Similarly to English [27] or German [21], Czech functional categories seem to rely to a great extent on temporal and melodic cues, although intensity plays its role, too. An experiment is currently in progress, testing the perceptual response of German and Czech listeners to the exemplars from our current study.

The under-represented categories 3, 7 and 8 were difficult to classify. Although this can be due to the computational safeguards (not generalizing for small samples), our intuitive evaluation suggests that these categories are not just relatively rare, but also internally disparate. The pragmatic messages they signal (impatience, resignation, or reassurance, respectively) may be expressed by various means and, therefore, be less well-defined than their more frequent counterparts. Further research in this respect is needed, but prior verification of these categories by larger listener groups is crucial.

5. Acknowledgements

The 1st & 2nd author were supported by the Programme of Scientific Areas Development at Charles University in Prague, Subsect. 10, Linguistics: Social Group Variation. We also thank Hana Bartůňková for her help with the categorization.

6. References

[1] Cattell, R. “Negative transportation and tag questions”, Language 49(3): 612–39, 1973.

[2] Millar, M. and Brown, K., “Tag questions in Edinburgh speech”, Linguistische Berichte 60: 24–45, 1979.

[3] Cruz-Ferreira, M., “Tag Questions in Portuguese: Gramnnar and Intonation”, Phonetica 38: 341–352, 1981.

[4] Bald, W.-D., “English tag-questions and intonation”, in K. Schuhmann [Ed.], Anglistentag 1979: Vorträge und Protokolle, 263–91, Berlin: Technische Universität Berlin, 1980.

[5] Tottie, G. and Hoffmann, S., “Tag questions in British and American English”, Journal of English Linguistics 34(4): 283–311, 2006.

[6] Dehé, N. and Braun, B., “The prosody of question tags in English”, English Lang. & Linguistics 17.1: 129–156, 2013.

[7] Reese, B. and Asher, N., “Prosody and the interpretation of tag questions”, Proc. Sinn und Bedeutung 11: 448–462, Barcelona: Universitat Pompeu Fabra, 2006.

[8] Janota, P., “Personal Characteristics of Speech”. Praha: Academia, 1967.

[9] Janota, P. and Palková, Z., “Auditory evaluation of stress under the influence of context”, AUC Philologica 2/1974, Phonetica Pragensia, 4: 29–59, 1974.

[10] Volín, J., “Z intonace čtených zpravodajství: výška první slabiky v taktu”, Čeština doma a ve světě 1–2: 89–96, 2008.

[11] Volín, J. and Weingartová, L., “Idiosyncrasies in local articulation rate trajectories in Czech”, Proceedings of Perspectives on Rhythm and Timing, 67, Glasgow: UG, 2012.

[12] Romportl, J., “Statistical Evaluation of Prosodic Phrases in the Czech Language”, Proceedings of the Speech Prosody 2008 Conference, 755–758, Campinas, Brazil, 2008.

[13] Dankovičová, J., “Articulation rate variation within the intonation phrase in Czech and English”. Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, 1999.

[14] Kolář, J., Romportl, J. and Psutka, J. “The Czech speech and prosody database both for ASR and TTS purposes”, Proceedings of Eurospeech 2003, 1577–1580, Geneva: ISCA, 2003.

[15] Bartošek, J. and Hanžl, V., “Intonation Based Sentence Modality Classifier for Czech Using Artificial Neural Network”, Proc. NOLISP 2011, 162–169, 2011.

[16] Duběda, T. and Raab, J., “Pitch Accents, Boundary Tones and Contours: Automatic Learning of Czech Intonation”, Lecture Notes in Computer Science 5246: 293–301, 2008.

[17] Duběda, T., “Towards an inventory of pitch accents for read Czech”, Slovo a slovesnost 72: 3–12, 2011.

[18] Dombrowski, E., “Semantic Features of Accent Contours: Effects of F0 Peak Position and F0 Time Shape”, Proceedings 15th ICPhS, 1217–1220, Barcelona, 2003.

[19] Kohler, K., “Timing and communicative functions of pitch contours”, Phonetica, 62(2–4): 88–105, 2005.

[20] Niebuhr, O., “Perzeption und kognitive Verarbeitung der Sprechmelodie, Theoretische Grundlagen und empirische Untersuchungen”, Language, Context, and Cognition, Vol. VII, Berlin/New York: deGruyter, 2007.

[21] Niebuhr, O., “On the phonetics of intensifying emphasis in German”, Phonetica 67: 170–198, 2010.

[22] Niebuhr, O. and Zellers, M., “Late pitch accents in hat and dip intonation patterns”, in Niebuhr, O. and Pfitzinger, H. R. [Eds.], Prosdies: context, function, and communication, Berlin/New York: de Gruyter, 2012.

[23] Skarnitzl, R., “Prague Phonetic Corpus: status report”, AUC Philologica 1/2009, Phonetica Pragensia, XII: 65–67, 2010.

[24] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer” [computer program], version 5.3.35. Online: http://www.praat.org/.

[25] Kohler, K. J., “Modelling prosody in spontaneous speech”, in Y. Sagisaka, N. Campbell, and N. Higuchi [Eds.], Computing Prosody, Computational Models for Processing Spontaneous Speech, New York: Springer: 187–210, 1997.

[26] StatSoft, Inc. (2004). STATISTICA [computer program], ver. 7.

[27] Beňuš, Š., Gravano, A. and Hirschberg, J.: “Prosody, emotions, and... 'whatever'”, Proceedings of Interspeech 2007: 2629–2632, 2007.