Prosody patterns of feedback expressions in Hungarian spontaneous speech

Alexandra Markó[1], *Mária Gósy*[1,2], *Tilda Neuberger*[2]

[1] Department of Phonetics, Eötvös Loránd University, Budapest, Hungary
[2] Department of Phonetics, Research Institute for Linguistics of HAS, Budapest, Hungary
marko.alexandra@btk.elte.hu, gosy.maria@nytud.mta.hu, neuberger.tilda@nytud.mta.hu

10.21437/SpeechProsody.2014-84

Abstract

Speech communication incorporates non-verbal signals and semi-lexical vocal phenomena as well as words used as the listener's responses to the speaker's message. They are most common in conversation with various functions regardless of language. A specific subcategory is feedback expressions (FEs) that can be found in the listener's production as well as in the current speaker's speech production when reacting to the former speaker's message. This paper reports on the temporal and intonational characteristics of four types of FEs identified in 20 interviews and conversations from the BEA Hungarian database. Altogether 262 occurrences were categorized into four discourse functions signaling 'attention', 'comprehension', 'agreement' and 'other attitude'. Durations showed statistically significant differences across discourse functions. They were significantly longer in females than in males in all functions. The pitch range data revealed a statistically significant difference depending on discourse function and gender only in the case of the 'attention' function. The dominant frequency contour was a rise in the functions of 'attention' and 'agreement' (90%). The same contour was observed only in 75.5% of the 'comprehension' function. An integrated approach is proposed to analyze these phenomena in spontaneous speech.

Index Terms: discourse functions, prosody patterns, speaker–listener interaction

1. Introduction

Verbal communication incorporates verbal and non-verbal signals that interact both in speakers' speech production and in listeners' speech comprehension [1], [2], [3]. Semi-lexical non-verbal phenomena, short sound sequences like *ah, eh, ehm, er, erm, hmm, huh, mm, mmhm, oh, ooh, oops, phew, uh, uh-huh, um,* and even words like *yes, I see, right, okay* occur in conversations underlying cooperation between the participants of the dialogue (e.g., [4], [5]). A specific subcategory can be interpreted as 'feedback expressions' (FEs) that can be found in the listener's production as well as in the speaker's speech production when reacting to the former speaker's message. There are various terms referring to the nature of the listener's reactions like 'listener responses', 'accompaniment signals', 'continuers', 'assessments', 'acknowledgments', 'reactive tokens' (cf. [4], [6]) with meanings similar to that of the most commonly used term 'backchannel'. Backchannels have diverse definitions and descriptions (e.g., [4], [7], [8]). They are produced by one participant (the listener) in a conversation while the other one is talking. They do not cause the other speaker to cede the floor, fail to signal any intention to interrupt the speaker, and are generally non-information seeking phenomena (see [8]). The most widely identified and accepted function of backchannels is to signal attention, i.e. that the listener is attending to the speaker, reassuring the latter about his/her continuous attention [7], [9], [10]. The discourse function of a phenomenon like *uh-huh* notifying the speaker that the listener is listening was defined as early as in 1961 by Trager [11]. In a broader interpretation, however, backchannels may have various other discourse functions (such as signaling recognition, comprehension, emotional state, agreement, disagreement, attitude, support, etc., see [7], [8], [12], [13], [14], [15]), give feedbacks that make verbal communication more accurate or more continuous, and they may either support the mutual agreement between the participants or signal that some problem has arisen between them.

Backchannels signaling attention are reported to be prosodically well-defined in American English dialogues as opposed to affirmative words expressing other pragmatic functions, and the L-H% pattern was found to be characteristic of the analyzed phenomena (e.g., [7]). In addition, these backchannels were longer than affirmative words in other functions and similar to those expressing agreements.

In this paper we use the term 'feedback expression' in order to emphasize that both the listener's and the speaker's multifunctional feedback phenomena are considered in our analysis (see also [16]). In addition, FEs can occur both within turns and at turn-taking points and all of them were nonverbal phenomena.

The purpose of our study was to characterize the temporal and intonation patterns of most frequent types of FEs by measured data in Hungarian spontaneous speech. Our main question was how speakers indicate and listeners interpret the functional variations of FEs in their feedback expressions. We identified FEs as expressions that (i) responded directly to other participants' messages (irrespective of the turn of the participants), and (ii) did not require acknowledgement by the speaker. The FEs could be characterized by the following articulation gestures in most of the cases: (i) voicing emerges from the nasal cavity, potentially accompanied by an [h]-type noise component while the oral cavity is inactive, (ii) an [h]-type consonant-like (mostly voiced) sound is inserted between two low vowels. (In cases where the consonant is replaced by a very short pause, the sequence indicates negation.) In general, both versions sound as disyllabic sequences and are more or less stable in their articulation (although shorter forms might occur). The first version will be indicated by the sequence *mhm* while the second one by *uh-huh*. The articulation gestures described seem to be similar to FEs found in other languages (see 'phonetic components' in non-lexical conversational sounds in [6], [15], [17]). Former studies about some types of FEs in Hungarian supported the claim that the affirmative and interrogative functions can be correctly identified in FE extracted from spontaneous speech [18], [19], and shed light on some of their acoustic properties [20]. The prosodic structures of FEs with various functions were analyzed using experimental settings [18]. The data obtained demonstrated that the three basic types (meaning 'yes', 'no', 'question') differed in their temporal complexity and their melodic patterns. The high proportions of the listeners' correct identifications of their semantic contents confirmed the mutual

interaction between speaker(s) and listener(s). In this study, we address the acoustic patterns that disambiguate the interpretation of the three analyzed types of FEs.

Our main hypothesis was that both the durations and melody patterns of FEs are dependent on their discourse functions (see [7]). We supposed that there would be large gender differences in the various functions of FEs in all measured data.

2. Subjects, material, method

Conversations and interviews of twenty subjects (10 females and 10 males, aged between 20 and 76, mean age: 39 years) from the BEA Hungarian Spontaneous Speech Database [21] were used. All of the participants were native speakers of Hungarian from Budapest. The interviewer was always the same young female speaker (her FEs were not considered in the analysis). The total duration of the recordings was 15.2 hours; 45.7 minutes per recording, on average. All instances of FE produced by the 20 participants (irrespective of their being a listener or a speaker) were marked and labeled together with turn properties in Praat [22] by two of the authors independently of each other. Discourse function was identified by analyzing the semantic context and the speaker–listener interaction. In case of rare disagreement between the authors (below 10%), the third author was consulted. Durations, mean, minimum and maximum values of F0, pitch range (based on voice reports that were corrected manually if it was necessary) as well as the intonational structures were measured. Since a great number of FEs occurred as overlap phenomena, the melody structures could be analyzed only in 68.3% of all instances (the categorization was not problematic even in these cases). The data were subjected to various statistical analyses (one-way ANOVA, Tukey's post hoc test, Mann–Whitney U test, Kruskal–Wallis test as appropriate) using SPSS 15.0.

3. Results

3.1. Discourse functions of FEs

The majority of instances of FEs were categorized according to three main discourse functions while the fourth one contained various other attitudes that occurred in our corpus. (i) The term 'attention' will be used here when the listener signals that s/he is aware of what is being said. This is the function of notifying the speaker that the listener is listening [7]. (ii) The term 'comprehension' will be used when the listener's intention is to reassure the speaker that s/he has understood the message. (iii) The term 'agreement' will be used when the listener obviously agrees with the speaker, supports their ideas. These FEs were frequently accompanied by words like *yes, I see*. (iv) The term 'other attitude' serves as an umbrella term referring to phenomena that express attitudes or semantic content other than the former three types (such as surprise, disagreement, etc.). The examples demonstrate how FEs provide information about some hidden cognitive processes of the listener.

(i) Discourse function 'attention':
Interviewer: *én ugye még a régi rendszerben érettségiztem* ('well, I graduated in the old system')
Listening partner: *ühüm* ('mhm' meaning attentiveness)

(ii) Discourse function 'comprehension':
Interviewer: *mesélj egy kicsit arról hogy milyen szakos vagy, illetve hogy mivel akarsz majd későbbiekben foglalkozni* ('tell me about your university subject and about your plans for the future')
Listening partner: *ühüm öö hát én ugye magyar szakra járok* ('mhm /meaning I understand the task/ [*öö* = filled pause] well my main subject is Hungarian')

(iii) Discourse function 'agreement':
Interviewer: *az olvasáshoz hozzátartozik a színház is nyilván és az is ilyen ellenkezéseket szokott kiváltani a diákokból* ('reading is connected with theatre, obviously, and the latter often provokes disagreement from the students')
Listening partner: *aha igen persze ühüm* ('uh-huh, yes, sure, mhm')

(iv) Discourse function 'other attitude':
Interviewer: *az expressz járattal közlekedtem hát öt percenként indult* ('I used buses operating as express routes well that started in every 5 minutes')
Listening partner: *hm* ('hm' expressing surprise)

3.2. Occurrences of FEs

Altogether 262 instances of FEs with various discourse functions were found in our corpus (135 in the conversations, and 127 in the interviews). The mean occurrence of these phenomena was 13.1 per speaker (SD 11.74). Although the majority of the speakers used FEs less than 10 times in their spontaneous utterances (see the histogram in Fig. 1), great individual differences were found among them (from a single instance to 68). 142 of FEs were produced by female speakers while 120 by male speakers. No statistically significant difference was found depending on gender (mean occurrence for females is 14.2 per speaker, SD 11.44 and for males 12.0 per speaker, SD 11.6).

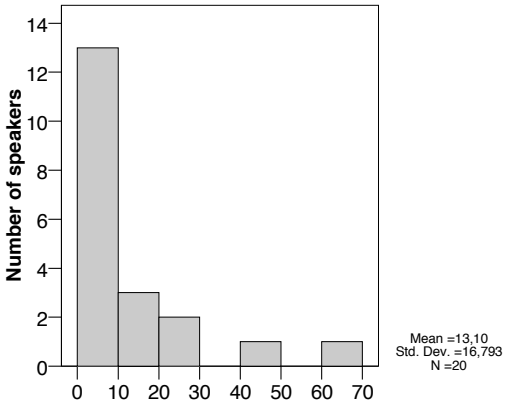

Figure 1: *The instances of tokens (FEs) across speakers (x axis = Frequency of tokens).*

The instances of FEs were heavily dependent on discourse function (Fig. 2). Females preferred FEs in the function of attention and used them more frequently than males did while males produced more instances in the discourse function of 'comprehension' than females did (Fig. 3). Since FEs occurred in the function of 'other attitude' rarely they were not included in our statistical analyses.

As expected, the majority of FEs (n = 130; 49.6% of all) occurred signaling **attention** and were produced both as in-between or overlap phenomena (see [23]), meaning that they were inserted (by the listener) during the speaker's turn in a pause period or they were uttered while the speaker was continuously speaking. The dominant form was *mhm* (90.8%; n = 118); *uh-huh* occurred in 8.5% (n = 11) of the cases and

one instance of *uh-hum* (1.7%) was also found. The latter is supposed to be the blending of the sequences *mhm* and *uh-huh*.

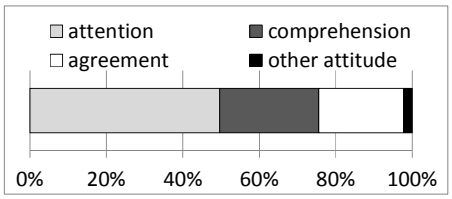

Figure 2: *The proportion of discourse functions of FEs in the corpus.*

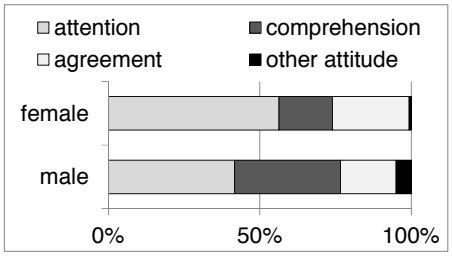

Figure 3: *The proportion of discourse functions of FEs depending on gender in the corpus.*

The discourse function signaling **comprehension** accounted for 25.6% of all instances (n = 67), and a greater variety of forms was used than in the 'attention' function. The type *mhm* was found in 70.1% (n = 47) of all instances in this function, while *uh-huh* occurred in 28.4% (n = 19), and one blended *u-um* form (1.5%) was found here, too. The instances of this function were produced during the speaker's turn in 77.6% of all cases, and 17.9% of them occurred at turn boundaries (Fig. 4).

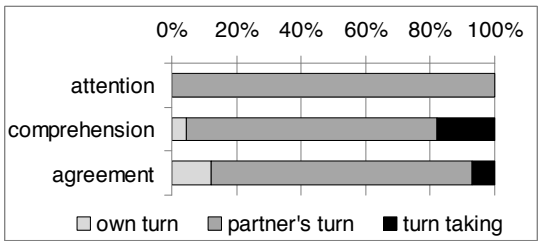

Figure 4: *The proportion of the placements of the three types of FEs.*

The discourse function **'agreement'** occurred in 22.1% of all cases (n = 58). 75.9% of them (n = 44) was identified as *mhm* and 24.1% of them (n = 14) as *uh-huh*. The great majority of the instances in this function (81.0%) were found during the speaker's turn and 6.9% of them occurred at turn boundaries, while some (12.1%) were found in the former listener's own turn.

3.3. Durations of FEs

The interrelations of the discourse functions and their durations are demonstrated in Figure 5. The forms *mhm* are longer (mean 250 ms, SD 56 ms) than those of *uh-huh* (mean 247 ms, SD 55 ms) in the function of 'attention'. The mean duration of *mhm* in the function of 'comprehension' is 289 ms (SD 70 ms) while that of *uh-huh* in the same function is 253 ms (SD 41 ms). The mean duration of *mhm* in the function of 'agreement' is 258 ms (SD 71 ms) while that of *uh-huh* in the same function is 232 ms (SD 63 ms). Instances in the function of 'comprehension' had the longest durations while no large differences were found in the cases of the other two functions. Statistical analysis confirmed a significant difference in the durations of instances depending on discourse function (one-way ANOVA: $F(5, 239) = 3.443, p = 0.005$). The Tukey's post hoc test shows, however, that the difference was significant only in the duration of *mhm* forms, and only between the functions of 'attention' and 'comprehension' ($p = 0.004$).

Durations were also analyzed in terms of gender. In this analysis only the *mhm* forms were considered due to the low number occurrences of the other forms. The mean duration of *mhm* in the function of 'attention' was 267 ms in the females (SD 40) and 200 ms in the males (SD 49). In the function 'comprehension' it was 335 ms (SD 42) in females and 262 ms (SD 69) in males, and in the function of 'agreement' it was 291 ms (SD 38) in females and 198 ms (SD 44) in males (Fig. 6). Statistical analysis including Tukey's post hoc test revealed a significant difference in durations of FEs between females and males (one-way ANOVA: $F(5, 196) = 21.011, p < 0.001$).

3.4. Pitch ranges and melody contours of FEs

Both the fundamental frequency values and the intonation patterns of all measurable instances of FEs were analyzed. In the discourse function of 'attention' 84 tokens were eligible for analysis in terms of pitch, as well as 49 instances of signals of comprehension, and 40 occurrences of the discourse function of 'agreement'. In other functions altogether 6 tokens were analyzed. The pitch range data are shown in Table 1.

Table 1. *Pitch range values depending on discourse function and gender.*

Discourse functions	Pitch range (semitone)			
	Females		Males	
	Mean	SD	Mean	SD
Attention	4.15	1.48	2.74	1.06
Comprehension	3.54	1.55	3.26	1.16
Agreement	3.48	1.35	3.80	1.67

Statistical analysis confirmed a significant difference in the pitch ranges (Kruskal–Wallis test: $\chi^2(5) = 15.813, p = 0.007$) while the Mann–Whitney U test revealed that there was significant difference in pitch ranges depending on gender only in the function of 'attention' ($Z = -4.004, p = 0.001$).

The prototypical element of the melody patterns of FEs was, in general, a final rise which was frequently preceded by a fall, a descent or monotonous contour as a preparatory one. Sometimes glottalized syllables occurred preceding the rise (e.g., [4]). 94.0% (n = 79) of all instances in the function of 'attention' ended with a rise (Fig. 7).

A similar rising contour or preparatory contour + rising was characteristic of the function of 'comprehension' in 75.5% (n = 37) of the cases. However, the descent contour was also produced in this function, in 24.5% (n = 12) of all cases. Acoustically, the steepness of these contours was diverse and various complex structures like descent + step up/rise + descent, fall + monotonous, rise + descent were also found.

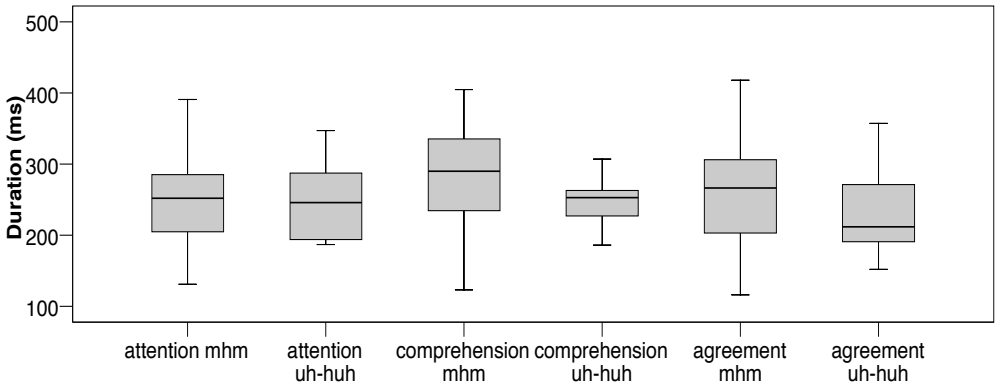

Figure 5: *Durations (medians and ranges) of various forms of FEs in the various discourse functions.*

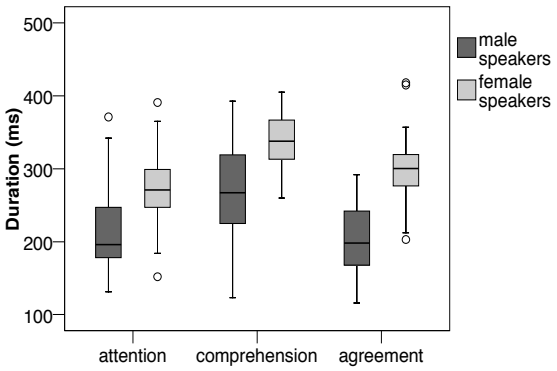

Figure 6: *Durations (medians and ranges) of* mhm *depending on the discourse function and as a function of gender.*

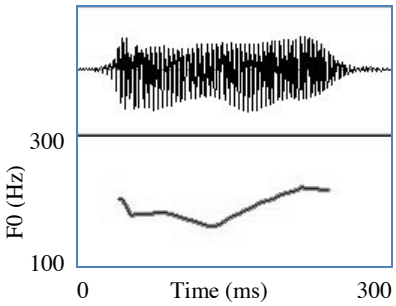

Figure 7: *Prototypical melody pattern of a* mhm *signaling attention.*

In the function of 'agreement', the dominant melody contour was also the rise (92.5%; n = 37), while a small portion (2.5.0%; n = 3) showed intonation structures with no rising contours.

Four instances of the mixed type function expressing surprise and disagreement showed descent melody contours. One token expressing a question had a rising contour while another one expressing 'no' had a monotonous contour.

4. Discussion

In this study we provided measured data on instances of FEs concerning their occurrences, temporal and melody patterns in view of the three main discourse functions they reflect. As expected, listeners felt it to be the most important to notify the speaker about their attentiveness. However, females and males behaved differently: the former used FEs in the function 'attention' more frequently than males did while the latter preferred to signal their comprehension during conversation. Males' frequent signaling of comprehension might be explained by the fact that the interviewer was a female speaker.

The main hypothesis of the research, however, was only partly confirmed. The closest interrelations were found between the durations of the various forms of FEs and their discourse functions, suggesting that speakers seem to differentiate the discourse functions by articulating them differently along the durational scale. In addition, the different forms (*mhm* and *uh-huh*) with different durations depending on function support the listeners' intention to inform the speaker about their attitudes. Females seem to express certainty and reassurance concerning the speaker's message by longer durations of FEs than males do. The non-neutral rising intonation contours of the majority of FEs reinforce the information that the discourse functions convey, while the relatively frequent fall/descent contours signal that the utterance is self-contained and finished (see [24]).

5. Conclusions

Our research findings evidenced speaker and listener interactions in conversations by measured acoustic-phonetic data of FEs with three main discourse functions. The temporal and melody patterns that the speakers produced do not seem to be incidental; however, differences in the actual acoustic patterns could be shown across languages like English, Italian, Japanese, Swedish, and Hungarian [4, 7, 10, 13, 17, 25]. However, identification of a discourse function must take into account various other factors (like gaze direction). We can conclude that analyzing feedback expression phenomena using an integrated approach considering communication situation, participants, contexts, various types of feedbacks, function, and acoustic patterns [e.g., 2, 3, 10] is crucial to fully understand spontaneous conversations.

6. Acknowledgements

We wish to thank Louise Mycock for her help concerning an earlier version of this paper.

This research was supported by OTKA project No. 108762.

7. References

[1] Schmidt, J. E., "Neue Wege der Intonationsforschung", Georg Olms Verlag, Hildesheim, Zürich, New York, 2001.

[2] Jones, S. E. and LeBaron, C. D., "Research on the Relationship between Verbal and Nonverbal Communication: Emerging Integrations", Journal of Communication, 52(3):499-521, 2002.

[3] Allwood, J., and Cerrato, L. "A study of gestural feedback expressions", In First Nordic Symposium on Multimodal Communication, Copenhagen, 7-22. 2003.

[4] Ward, N. and Tsukahara, W., "Prosodic features which cue back-channel responses in English and Japanese", Journal of Pragmatics, 32(8):1177-1207, 2000.

[5] TEI: Text Encoding Initiative: http://www.tei-c.org/index.xml, download: 17 Nov 2013.

[6] Miller, L. Verbal listening behavior in conversations between Japanese and Americans. The Pragmatics of International and Intercultural Communication. Amsterdam: John Benjamins Publishing Company, 111-130. 1991.

[7] Benus, S., Gravano, A., and Hirschberg, J. The prosody of backchannels in American English. In Proceedings of ICPhS 2007, 1065-1068. 2007.

[8] Lai, C. "Prosodic Cues for Backchannels and Short Questions: Really?" Speech Prosody Conference. 2008.

[9] Yngve, V. "On getting a word in edgewise", Papers from the Sixth Regional Meeting of the Chicago Linguistic Society: 567-577. 1970.

[10] Gravano, A., Benus, S., Chavez, H., Hirschberg, J., and Wilcox, L. On the role of context and prosody in the interpretation of okay. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 800-807. 2007.

[11] Trager, G.L., "The typology of paralanguage", Anthropological Linguistics, 3:17-21, 1961.

[12] Ward, N., "Pragmatic functions of prosodic features in non-lexical utterances." Speech Prosody 2004, International Conference, 325-328, Japan, 2004.

[13] White, S., "Backchannels across cultures: a study of Americans and Japanese", Language in Society, 18:59-76, 1989.

[14] Gardner, R., "When Listeners Talk: Response Tokens and Listener Stance", Amsterdam, J. Benjamins Publishing, 2001.

[15] Ward, N., "Non-Lexical Conversational Sounds in American English", Pragmatics and Cognition, 14:113-184, 2006.

[16] Allwood, J., Nivre, J., & Ahlsén, E. On the semantics and pragmatics of linguistic feedback. Journal of semantics, 9(1), 1-26. 1992.

[17] Cerrato, L. Some characteristics of feedback expressions in Swedish. In Proc. of Fonetik. Vol. 44, pp. 41-44. 2002.

[18] Markó, A., "Szavak nélkül. Nonverbális vokális közlések fonetikai elemzése" [Without words. A phonetic analysis of nonverbal vocal communication], in Hungarian, Magyar Nyelvőr, 129:88-104, 2005.

[19] Markó, A., "A special conversational device: humming in Hungarian", The Phonetician, 95:28-31.

[20] Neuberger, T. and Beke, A. "Automatic Laughter Detection in Spontaneous Speech Using GMM-SVM Method", in I. Habernal and V. Matousek [eds.], Text, Speech and Dialogue, 16th International Conference, TSD 2012, Pilsen, Czech Republic, Springer, 113-120, 2012.

[21] Gósy, M., "BEA – A multifunctional Hungarian spoken language database", The Phonetician, 105/106:50-61, 2012.

[22] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", Computer program, Version 5.2 retrieved 10 Sept 2010 from http://www.praat.org/

[23] Feke, M. S., "Effects of Native-language and Sex on Back-channel Behavior", in L. Sayahi [ed.], Selected Proceedings of the First Workshop on Spanish Sociolinguistics, 96-106, Somerville, MA: Cascadilla Proceedings Project, 2003.

[24] Varga, L. "Intonation and stress. Evidence from Hungarian". Houndmills, Basingstoke, Palgrave Macmillan. 2002.

[25] Cerrato, L. "Investigating communicative feedback phenomena across languages and modalities", Doctoral Thesis, Stockholm, Sweden. 2007.