Multimodal aspects of positive and negative responses in Polish task-oriented dialogues

Zofia Malisz [1,2], Maciej Karpiński [1,3]

[1] Centre for Speech and Language Processing, Adam Mickiewicz University, Poznań, Poland
[2] Bielefeld University, Germany
[3] Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland
zofia.malisz@uni-bielefeld.de, maciej.karpinski@amu.edu.pl

Abstract

The paper reports on a multimodal analysis of short positive and negative responses in a corpus of Polish task-oriented dialogues. In the present task, responses most frequently functioned as realisations of affirmative and negative feedback and confirmation/disconfirmation. We describe prosodic and visual cues, namely intonation, head movement and smile that may distinguish between the main dialogue act functions in the studied corpus. Integrating our findings, we present a preliminary multimodal picture of this class of dialogue contributions in Polish.
Index Terms: dialogue, feedback, Polish, multimodal

1.Introduction

In asymmetric task-oriented dialogues, where one speaker is assuming the leading role in the task, it is usually the instruction giver that tends to employ forward communicative functions such as instructions and offers, most frequently [1]. However, in most cases, the successful completion of the task is also dependent on the appropriate usage of "backward looking" feedback and answering acts. Such acts are most often realised by responses, i.e. relatively short words such as *yeah*, *okay*, *exactly*, *no* and quasi-words such as *mhm*. As short as such expressions may be, they take a variety of roles in the broad context of discourse, including task-oriented interactions. Responses may provide information about the success of communication by means of conveying agreement and acknowledgement of the partner's utterances or actions. Responses can also, sometimes simultaneously with other functions, establish and maintain contact between speakers, serve to show readiness and attention and offer turn taking cues [2].

The functional interpretation of a given response depends on many, sometimes very subtle cues, such as prosody, nonverbal signals or timing and context within the dialogue [3,4,5]. [6] analysed the prosody and discourse functions of *okay* and *uh-huh*. A pitch rise in the second syllable of these units was identified as related to their backchannel function. In a perceptual study, [4] have shown that the final intonation of *okay* in English strongly correlates with the functional interpretation of the word. The context in which the word occurs also plays a role in function disambiguation. According to [5], "both prosodic and contextual factors distinguish backchannels from other affirmative words" in American English.

As far as the differences between multimodal realisations of positive and negative feedback are concerned, in an experiment conducted by [7], subjects' sensitivity to feedback cues, as expressed by a talking head, was measured. Smile was the most salient visual cue to affirmative feedback prior to the main prosodic cue of f_0 modulations, followed by eyebrow and head movement. [2] noted that head movement often co-occurs with short responses such as *mm* and *ja* in Swedish. They also observed that nods were more subtle when responses assumed a backchanneling function rather than attitudinal functions (such as agreement, acceptance, or surprise). In the latter case, the amplitude of nods was more extensive, probably adding emphasis. On the other hand, multimodal features characteristic of negative feedback found by [7] were: "a neutral mouth configuration, interrogative intonation, a slow upwards movement of the head and eyebrow frowning". As it seems, evidence exists to support the hypothesis that prosodic, contextual and non-vocal cues form a complex relationship in order to convey the intended message and attitudes of the speakers.

The appropriate and precise identification of response functions is crucial for the efficiency of dialogue systems both on the level of conversation and task realisation. Since the amount of information that may be encoded directly, in feedback realisations is quite limited (they are short and morphologically simple) much attention has been confessed to the localisation and identification of e.g. backchannels in this context [8,9,10]. The multimodal study of feedback is also significant for the construction of naturally sounding and acting embodied conversational agents. Several applications as well as theoretical models of embodied feedback to be used in such systems have been developed recently [11,12,13]. Gesture, such as pointing, relevant in multimodal modelling of feedback, has also been generated successfully [14].

Only few attempts have been made so far towards a systematic study of the form and function of affirmative and negative responses in Polish. In [15] it was found that some prosodic properties of short responses and quasi-words may be regarded as cues to discourse functions. For example, *tak* (a close equivalent of *yes*) was significantly more often realised with a rising melody when used for the acknowledgement of instructions or orders than when it was used for the acknowledgement of statements. The quasi-word *mhm* was produced predominantly with a rising melody in positive feedback or confirmation related functions.

A comparative study by [16] showed that while Korean and Thai positive responses seemed to show relatively clear tendencies with respect to intonational contours, their Polish equivalents turned out to be much less consistent and featured a wide range of different melodies. These results either suggest some as yet unidentified influences shaping the intonational contours or the possibility that Polish allows for more idiosyncracy in the intonational realisations of feedback. We also believe that additional information disambiguating the finely-grained dialogue functions can be drawn from other sources such as, e.g. non- vocal modalities.

In the present study, we examine some multimodal aspects of Polish feedback in the task oriented-dialogue: intonation, head movement, smile and gestural context.

2.Material

The material under study comes from the DiaGest2 corpus of multimodal dialogues based on a paper folding task [17]. It comprises eight audiovisually recorded task-dialogue sessions. A spatial, origami-like structure made of paper was placed in front of subject A. Subject B was asked to produce a possibly identical structure according to the instructions

given by A. Subjects A and B could see each other, but subject B could not see the original structure. Each session was interrupted after five minutes even if the task had not been completed.

2.1.Responses

In the present study only responses of the instruction givers (4 females and 4 males) were analysed. Responses were marked on a separate tier in Praat using an existing phonetic annotation of the data. Short, one- or two-syllable responses were selected. (see: Table 1). It has to be noted that e.g.: *właśnie*, *dobrze*, *tak* and *okej* may appear in larger utterances bearing different meanings as in:

(1) "Stała [*właśnie*] tam" > "She stood [exactly] there"
(2) "Mogłaś zrobić to [*tak*]" > "You could have done it [this way]"

Such uses are not responses in Polish and were not tagged as such. *Tak* may be used as a demonstrative pronoun in a larger utterance, as can be seen in (2) but also in an independent phrase, as in:

(3) A: "Jak mam to zrobić?" B: "*Tak*!" > A: "How am I supposed to do this" B: "Thus!"

Tak and *nie* may also function as question tags and *no* as an emphatic particle. All such uses of the lexical items in question were excluded from the analysis.

Table 1. *Responses analysed in the present study.*

Polish response	IPA	English approx. equivalent
tak	/tak/	yes
mhm	/mmm/	yes, well
no	/no/	yes
właśnie	/'vwaɕɲe/	exactly, yes
o	/o/	this way, yes, exactly
dobrze	/'dobʒe/	okay, allright, yes
okej	/okej/	okay, allright, yes
nie	/ɲe/	no

2.2.Dialogue acts

The concept and taxonomy of dialogue acts used in this paper was adopted from Bunt's DIT++ system [18]. However, only the dominant function (according to the labeller's judgment) of each utterance was taken into account (cf. [19]). Additionally, the tagset distinguished between feedback given in relation to spoken utterances and to task-related actions. Most of the acts realised by the analysed lexical items fell into two major categories: (a) various kinds of feedback and (b) answers to propositional or check questions (confirmation vs. disconfirmation). The items in (b) were not labelled separately for any other functions as most of them were strictly task-related.

2.3.Intonation

Intonation was analysed instrumentally (using Praat [20]) and on the basis of perception. Nuclear melody was labelled using the following tags: High, Rise, Flat, Fall, Low. High, Low and Flat tags denoted a high, mid or low perceptually level pitch. Rise and Fall tags were used to annotate perceivable pitch movement sometimes reaching the same pitch levels as High and Low, respectively. In many cases, the entire utterance consisted of just a single syllable and, accordingly, pitch movement within a single syllable was labelled.

2.4.Visual modalities

Annotations of the visual modalities were prepared independently (with "audio mute" on) and integrated with the resulting speech annotation files in ELAN.

2.4.1.Head movement and smile

Head movement was annotated in ELAN on a separate tier. The following tags were used: Nod, Shake, Jerk and Tilt. By inspecting the video material the onset and offset of the movement was marked as well as the number of repetitions, e.g. Nod2 denoted two head nods in continuous succession. Only nods and shakes were included in the analysis since the subjects produced only single instances of jerks and tilts. Because of the quality of the recordings it was not possible to analyse eyebrow movement and other subtle facial expressions. The analysis of smiles was based on careful observation of the data using viewing tools in ELAN.

2.4.2.Manual gesture context

Gesture phrases, gesture phases and gesture types were annotated on a number of tiers. The main gesture phase tags included: Preparation, Stroke, Retraction, Post-Stroke Hold, as specified by [21]. Gesture types were labelled according to McNeill's [22] taxonomy: Deictic, Iconic, Beat, Metaphoric etc. The gestural context of positive and negative responses was inspected in ELAN.

3.Results

3.1.Responses

The number of positive and negative responses produced was 321, including repetitions within one dialogue act category. On average, 8 to 9 responses per minute were uttered by the instruction givers. The most frequently produced responses were the positive *tak* (29%) and the only negative *nie* (23%). *Mhm*, *właśnie*, *dobrze* and *no* contributed from 8 to 12 % each to the whole number of responses. We also included *o* (4%) and *okej* (2%) in the further analysis. Two cases of a positive response were realised only by a head nod.

3.2.Dialogue acts

The analysis was focused on responses produced by the instruction givers. As a result 223 acts were labelled. It was hypothesised that backchanneling functions would occur infrequently, since signalling contact, attention and perception should be more characteristic of instruction followers. Indeed, instruction givers mostly produced responses that functioned as feedback to actions executed by the instruction followers (labelled with: ActionExecutionAlloFeedback). Responses also often functioned as confirmations or disconfirmations. Table 2 summarises the proportion of the most frequent acts expressed by the units under study. Other less frequently used functions (from 2 to 7 instances) realised by the instruction givers were: own communication management, contact indicators, single instances of AcceptOffer and Agreement.

Table 2. *Distribution of dialogue acts in the corpus of positive and negative responses.*

%	Allo Feed back	Conf	Auto Feed back	Disconf	Other	Sum
pos	53	34	4	0	9	100
neg	53	0	0	42	5	100

The most frequently used response expressing both allo-feedback and confirmation was *tak*. Other two short words that were commonly shared by these two functions were *mhm* and *no*. It was not surprising to see *właśnie*, and *dobrze* being used more to convey acceptance of the partner's actions since literal meanings of *dobrze* and *właśnie* contain an added quality of positive evaluation. *O* is often used to accept an outcome of an action, while *tak* often expresses acceptance of

how the action proceeds. However, the simple agreement function of *tak* was reflected in the high number of uses in the confirmation category. A summary of the results is shown in Figure 1. The negative response *nie* split in half between dialogue act categories: 23 such realisations of negative allo-feedback were found and 22 of disconfirmation.

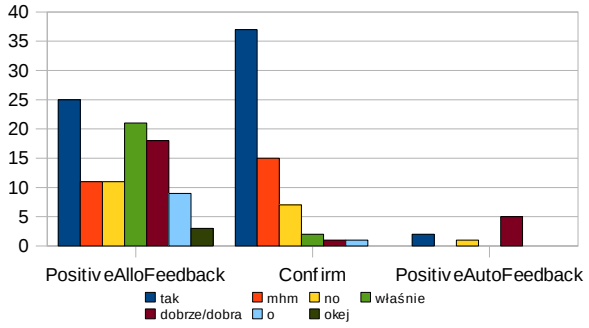

Figure 1: *The number of positive responses in the most frequent dialogue act categories used by the instruction giver.*

3.3. Intonation

As shown in Figure 2, the proportion of nuclear melody categories differs across the four most frequent dialogue act classes expressed by responses. In the realisations of negative feedback and disconfirmation (negative answers) the proportion of non-rising contours was considerably higher than in the realisations of positive feedback and confirmation, respectively. The difference in distribution was most evident between Confirm and Disconfirm acts: a very large proportion of falling (48%) and low (28%) nuclear melodies was found in the realisations of Disconfirm unlike in the expressions of Confirm (25% and 1.5%, respectively). Among the most commonly used quasi-words *mhm* was almost invariably used with a high tone on its second, accented syllable. It should be noted that the position of the accent in the realisations of this unit is atypical of Polish (as the last, rather than the penultimate syllable of a two-syllable word is accented). It was difficult to identify any strong influence of dialogue act category on the intonational response profile. A moderate tendency was observed for the realisations of *tak*: in positive feedback all employed non-falling contours, while when used as confirmation (yes-answer) falls and a large proportion of flat contours were found.

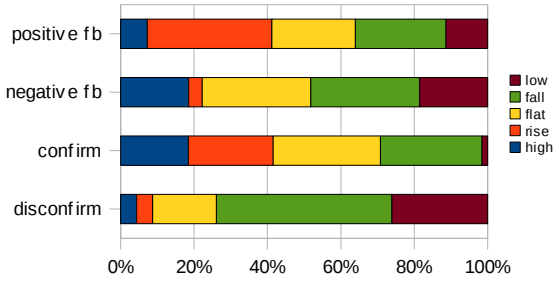

Figure 2: *Proportions of nuclear melody categories used in the realisations of four most frequent dialogue act groups.*

For the most frequent units, i.e. *mhm*, *nie* and *tak*, two basic measures of pitch frequency were calculated: the mean f_0 value and the relative range of f_0. The results are shown in Table 3. The two most numerously represented categories of dialogue acts were considered separately for each of the units. No statistically significant differences were found between the mean pitch values for different acts realised by a given word.

Table 3. *Values of the mean f_0 and relative f_0 range for the most frequent words in their most frequent functions.*

Unit	Dialogue act group	Mean f_0 value in Hz	Relative f_0 jump percentages
mhm	Confirm	191.5 (62.5)	17.7 (11.6)
	PositiveFeedback	183 (8)	25 (16)
tak	Confirm	198 (62)	11.3 (13)
	PositiveFeedback	194 (79)	14.2 (16)
nie	Disconfirm	194 (51)	17 (12)
	NegativeFeedback	208 (58)	22 (17)

For multiple repetitions of certain words (e.g., *tak*) in the realisations of the analysed dialogue acts, two basic patterns were observed. First, where each realisation forms a separate intonational unit (phrase) and second, where subsequent repetitions are joined in a single intonational phrase. Although multiple units were not directly taken into acount in the main statistics, the intonational patterns occurring there were compliant with what was observed for single realisations of the respective words.

3.4. Head movement and smile

Tak, *mhm* and *no* were accompanied by nods most frequently: 52%, 55%, 64% respectively while giving positive feedback and 62%, 40%, 57% while confirming. The number of nods produced with these three words in relation to the total number of nods was 0.9. Did the ratio arise from the high frequency of these particular words in the whole dataset? Not necessarily, since in the positive allo-feedback function the words *właśnie* and *dobrze* expressed feedback more frequently than *mhm* and *no* [cf. Figure 1] but did not tend to co-occur with nodding. However, 59% of confirmation realisations were *tak*, out of which 62% were accompanied by nodding.

Negative responses, that is, exclusively *nie*, halved its use of head shakes: 50% of negative feedback realisations and of disconfirmations overlapped with head shaking. In some of our data, smile co-occurred with multiple continuous head shakes and repeated *nie* in the negative allo-feedback function to signal a sympathetic rejection that is not judgemental. We observed only one use of a smile accompanying a positive feedback realisation.

3.5. Gesture context

Gestures tend to slightly precede or coincide with their lexical counterparts [21]. Temporal relations between gesture phases and responses were analysed by counting instances of the following relation categories: (a) a response overlapping with any gesture phase, (b) a gesture stroke target preceding the acoustic onset of a response by 0 to 300 ms. Stroke target was chosen as the point of reference in (b) because its relative synchronisation with relevant speech units was found to be the most reliable temporal correlate of semantic correspondence in this material. All search results involving strokes were additionally verified for semantic correspondence by the authors.

The vast majority of responses overlapped with gesture phases that did not carry meaning: various gesture holds (48%), retractions (19%), rests (18%) or preparations (8%). For comparison, 48% of instructing acts analysed in an equivalent corpus contained at least one stroke within their temporal limits [23]. Out of a total 292 strokes in the data, only 16 strokes (5.5%) occurred in relation (a) or (b) to the responses. In all of these cases, the gestures were iconic/deictic topic gestures and “object-oriented”, in the

sense of [24]. The few gestures that happened during a positive response often pointed towards and "fixed" the shape of the paper construction in space, as if to say "this is right, yes, leave it there". During a negative response, the pointing finger was either used to suggest "undo" and/or show the desired direction of folding. Contrary to the material in [24], all such object-oriented gestures here were definitely made for the benefit of the addressee, since the addressee could see what the instruction giver was pointing at. Emblems such as finger wagging to say *no*, were not observed in this small set of data.

4.Discussion

Our results show that short responses involve both prosodic and nonverbal components that are used to strengthen and further modify their dialogue function, especially in the negative/affirmative dimension. For *tak*, a pitch contour difference between giving feedback and confirming was observed, while *mhm* was again [cf. 15] found to employ non-falling contours consistently. Here, a comparison with the use of *mhm* as a backchannel and turn-taking cue is needed in the future. We also see the need to take inter-modal factors, such as the relationship between prosody and the non-vocal component, into account.

The most frequent, short expressions of feedback and answers in Polish i.e. *tak*, *no*, *mhm* tend to co-occur with head nodding and *nie* with shaking in approx. half of their respective instances. This result is similar to what was found for Swedish in [2]. Affirmative words that already have an added value of acceptance (*właśnie*, *dobrze*) do not tend to co-occur with nods and if they do, the head movement may form a further level of emphasis and expression of speaker attitude. Smile was not found to be produced consistently while giving positive feedback, instead, in some speakers, it served as a tool to lessen the psychological impact of rejection of the partner's actions. Such an effect does not, we believe, disprove the results by [7] since it was most likely influenced by the nature of the task. Our result might have been influenced by the fact that the task involved paying heightened attention to the paper folding by both speakers and so the use of the social and attitudinal meaning of smile typical of face-to-face interactions was limited.

A breaking of the interactive gestural aspect in the dialogue via the task could also be a possible explanation for the frequent use of topic gestures with feedback (and instructions [23]) in the present corpus. Other than what was shown in a task-oriented dialogue by e.g. [24], our subjects used topic gestures more than interactive gestures because of their spatial and attentional orientation to the folding. But since the use of interactive gestures was shown to enhance or replace backchannels and turn-taking cues, a more detailed functional analysis of gesture and feedback is needed. Especially since our data showed a large proportion of overlap with non-meaningful phases of gesture, the relationship between e.g. holds and turn-taking cues needs to be investigated.

5.Conclusions

Apart from the verification of some earlier findings, we have provided more clues to possible relations between the form and function of short responses in Polish dialogues. A substantially larger corpus is being built to gather a satisfactory number of instances for more advanced statistics. A comprehensive model may prove to be a valuable component of multimodal dialogue systems and embodied communicating agents.

6.Acknowledgements

This research is supported by the Polish Ministry of Science and Higher Education project (N N104 010337). We would like to thank Ewa Jarmołowicz-Nowikow for gesture annotations.

7.References

[1] Allen, J. and Core, M., "Draft of DAMSL: Dialogue Act Markup in Several Layers",1997.

[2] Allwood, J. and Cerrato, L., "A study of gestural feedback expressions", in P. Paggio, K. Jokinen and A. Jönsson, [Eds], First Nordic Symposium on Multimodal Communication, Copenhagen, 23-24 September, 7–22, 2003.

[3] Jurafsky, D., Shriberg, E., Fox, B., Curl, S., "Lexical, prosodic, and syntactic cues for dialog acts", ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers, 1998.

[4] Gravano, A., Benus, S., Chavez, H., Hirschberg,J., and Wilcox, L.,"On the role of context and prosody in the interpretation of okay", in Proc. of ACL, Prague, Czech Republic, June 2007.

[5] Benus, S., Gravano, A., Hirschberg, J., "The prosody of backchannels in American English", Proc. of ICPHS 2007, 1065-1068, 2007.

[6] Hockey, B. A., "Prosody and the role of okay and uh-huh in discourse", Proc. ESCOL, 128-136, 1993.

[7] Granström, B. and House, D., "Measuring and modeling audiovisual prosody for animated agents", in Proc. of Speech Prosody 2006, Dresden, 2006.

[8] Ward, N., "Using prosodic clues to decide when to produce back-channel utterances", in Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP96), 1728–1731, 1996.

[9] Noguchi, H. and Den, Y., "Prosody-based detection of the context of backchannel responses", in ICSLP- 1998, paper 0662, 1998.

[10] Cathcart, N., Carletta, J. and Klein, E., "A Shallow Model of Backchannel Continuers in Spoken Dialogue", Proc. of the Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003), Budapest, 2003.

[11] Allwood, J., Kopp, S., Grammer, K., Ahlsen, E., Oberzaucher, E., Koppensteiner, M., "The analysis of embodied communicative feedback in multimodal corpora - a prerequisite for behaviour simulation", Language Resources and Evaluation 41(3-4): 255-272, 2008.

[12] Fujie, S., Fukushima, K., Kobayashi, T., "A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information", in Proc. ICARA Int. Conference on Autonomous Robots and Agents, 379-384, 2004.

[13] Kopp, S., Allwood, J., Ahlsen, E., Grammer, K., Stocksmeier, T., "Modeling embodied feedback in a virtual human", in I. Wachsmuth and G. Knoblich [Eds], Modeling Communication With Robots And Virtual Humans, LNAI 4930, Berlin, 18-37, Springer, 2008.

[14] Kranstedt, A. and Wachsmuth, I., "Incremental generation of multimodal deixis referring to objects", in Proceedings of the 10th European Workshop on Natural Language Generation (ENLG-05), Aberdeen, Scotland, 75--82, August 2005.

[15] Karpiński, M., "Selected quasi-lexical and non-lexical units in Polish map-task dialogues", Archives of Acoustics, 32, 1, 2007.

[16] Karpiński, M., Kleśta, J., Szalkowska, E., "Non- and quasi- lexical realisations of "positive response" in Korean, Polish and Thai", in R. Hoffmann, H.-J. Mixdorf [Eds], Proc. Speech Prosody 2006, Dresden, 2006.

[17] Jarmołowicz,E., Karpiński,M., Malisz,Z., Szczyszek, M., "Gesture, prosody and lexicon in task-oriented dialogues: multimedia corpus recording and labelling", in A. Esposito, M. Faundez-Zanuy, E. Keller, M. Marinaro [Eds], Verbal and Nonverbal Communication Behaviours, LNAI 4775, 99-110, Springer, 2007.

[18] Bunt, H., "Dimensions in dialogue act annotation", Proc. of LREC 06, 919-924, 2006.

[19] Włodarczak, M. "Ranked multidimensional dialogue act annotation", unpublished MA thesis, Institute of Linguistics, Adam Mickiewicz University, Poznan, 2009.

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

[21] Kendon, A., "Gesture: Visible action as utterance", Cambridge University Press, 2004.

[22] McNeill, D., "Hand and mind", University of Chicago Press, Chicago, Illinois, 1992.

[23] Karpiński, M., "Preliminary prosodic and gestural characteristics of instructing acts in Polish task-oriented dialogues", in A. Esposito, A. Hussain, M. Marinaro, R. Martone, [Eds], Cross-modal Analysis of Speech, Gestures, Gaze and Facial Expressions, LNAI 5641, Springer, 2009.

[24] Bavelas, J. B., Gerwing, J., Sutton, C., and Prevost, D., "Gesturing on the telephone: Independent effects of dialogue and visibility", Journal of Memory and Language 58, 495-520, 2008.