ARCHIVES OF ACOUSTICS
32, 1, 51–65 (2007)

SELECTED QUASI-LEXICAL AND NON-LEXICAL UNITS IN POLISH MAP TASK DIALOGUES

Maciej KARPIŃSKI

Adam Mickiewicz University
Institute of Linguistics
Międzychodzka 5, 61-376 Poznań, Poland
e-mail: maciej.karpinski@amu.edu.pl

(*received October 16, 2006; accepted December 16, 2006*)

The present study is focused on selected types of fillers, quasi-words and non-lexical words that are generally categorized as expressing positive or negative response in Polish task-oriented dialogues. Basic phonetic properties of such units are analyzed with a special focus on intonation. Some of their possible realizations are shown and some relations between their intonational form and meaning are hypothesized. A brief note on comparative background from our recent work is also provided and some implications for speech technology are mentioned.

Keywords: dialogue, intonation.

1. Introduction

The units analyzed in the present paper do not form a homogeneous group. Some of them may be referred to as non-lexical words, while others are merely non-linguistic vocalizations that may be attributed certain meanings in specific contexts, or are just quasi-volatile expressions of hesitation (ROSE, [23]; WARD, [29]). However, they share one important property: Their function and meaning (if any) usually strongly depends on the prosodic realization. While some of them frequently act as discourse particles, organizing conversational turns or exchanges, many may be used to perform specific dialogue moves. For example, *tak* (a close equivalent of English "yes"), may be used to express agreement, confirmation, doubt, as a sort of exophoric reference ("in this way?", "like that?"), as a filler (usually [ta:k]) that signals prolonged thinking or, sometimes in conjunctions with some other words or quasi-words (*no tak*, *teraz tak*, *no to tak*), as a discourse marker, internally organizing conversational turns.

The role of fillers, hesitations and other phenomena "from the verge of language" has been studied from a few perspectives. Hesitation may results in filled or unfilled pauses that provide information about the stages of utterance production processes, GOLDMAN–EISLER, [8]. On the other hand, filled pauses as well as quasi-words may

act as discourse markers (SWERTS, WICHMAN and BEUN, [26]; SCHIFFRIN, [25]). They also add a touch of naturalness to spoken language. Hesitation-free utterances, deprived of fillers and quasi-words, may sound artificial, similarly to speech signal deprived of any prosodic variance (like pitch or speech rate changes). Nevertheless, one must also note that, in a number of situations, certain kinds of disfluencies are regularly overlooked or automatically ignored as irrelevant (LICKLEY, [16]; LICKLEY and BARD, [15]; BARD and LICKLEY, [1]).

For the present study, the words, quasi-words and vocalizations were selected which might be used as replacements for *tak* and *nie* (close equivalents of English "yes" and "no", respectively) and categorized as "positive" and "negative" responses.

An additional remark on the transcription of fillers and quasi-words seems to be necessary. While the units in question are immanent components of utterances, their linguistic status may be doubtful. They may not follow the phonotactics of a given language or may be built of segments that can be hardly classified as realizations of its phonemes. The use of phonological transcription for nonlinguistic vocalization is obviously risky. Although one may achieve a better representation of what was heard or said with the use of the phonetic transcription, it may prove hardly useful due to, e.g., infinite realization variance and simultaneous lack of categorization possibilities. The third way, chosen for this paper, is the conventional orthographic transcription enriched with appropriate visualizations, e.g., spectrograms and intonograms.

2. The data

A set of ten map task dialogues (CARLETTA *et al.,* [4]) was selected for this study from the *PoInt Corpus*(*) . The sessions selected for analyses were 6 min 30 sec to 19 min 49 sec long (the total duration of the analyzed sessions exceeded two hours). The number of female and male participants was balanced (ten speakers of each gender) and there were five "mixed" and five "same gender" pairs.

The recordings were segmented into "working units" (WU). For well-formed utterances, they corresponded to intonational phrases (IP). In other cases, more technical criteria were applied for segmentation. The units were transcribed using an extended orthographic transcription system and labeled as realizations of dialogue moves. Those tasks were carried out with Praat (BOERSMA and WENINK, [2]) and spreadsheet software.

From over two hours of recordings, 1588 targets were extracted. When a filler, a word or a quasi-word formed a separate IP or simply did not belong to any larger intonational unit, it was extracted without any context. When it formed a part of an WU, it was extracted with the entire WU (not necessarily being an IP). In 46 cases, a number of different targets occurred within one WU. They were treated as "sequences" and extracted as whole units. Each target (or a sequence of targets) was described in terms of its form and the move it realized or the move realized by the entire WU it occurred in.

(*) Polish Intonational Database project by M. Karpiński, W. Jassem and J. Kleśta (e.g., KARPIŃSKI, KLEŚTA, [12]).

Table 1. The categories of extracted units, selected possible English equivalents and the number of occurrences in the analyzed corpus.

No	Target category	Some possible English equivalents	Number of occurrences
1	monosegmental and other fillers	–	365
2	tak /tak/	yes	327
3	no /no/	well, er, argh, yes	236
4	mhm /mm̥m/	yes, sure, okay	167
5	dobra /dobra/	okay, all right, yes	114
6	aha /axa/	I see, yes	100
7	dobrze /dobʒe/	okay, all right, yes	58
8	others	–	56
9	sequences of units	–	46
10	okej /okej/	okay, all right, yes	40
11	nie /ɲe/	no	26
12	no dobra /no dobra/	okay, yes	22
13	audibly prolonged segments (vowels or cons.)	–	19
14	no dobrze /no dobʒe/	okay, yes	12

3. Data analysis

3.1. Dialogue moves labeling

Discourse-level labeling was based on the well-known, existing concept of dialogue moves. A new inventory of dialogue moves was built according to the needs of the Pol'n'Asia project. It differed from the older ones in a few ways. The division into initiating moves and responses was abandoned. The number of categories was larger than usually (about thirty categories and subcategories). While a touch of subjectivity is nearly always added to any interpretation and categorization of naturally occurring utterances, much can be achieved by the design of appropriate definitions for the units in question. The definitions of dialogue moves for the present study were based on: (a) the intention of the speaker and the obligation put on the hearer; (b) the result achieved in the hearer; (c) the semantic content of the utterance. This approach stems from the early works of Edinburgh Map Task team (CARLETTA *et al.*, [4]) as well as from other studies pertaining to the analysis of intentions, expectations, obligations in discourse (KREUTEL and MATHESON, [13]; CRISTEA and WEBBER, [6]; TRAUM and ALLEN, [27]; CARLSON, [5]; MANN, [18–20]).

Below, a working inventory of broad dialogue move categories used in this work is presented (detailed subcategories are not listed).

- Statement, Instruction;
- Y/N question (polar question), WH-question ("detailed" question), Question for Confirmation (often relatively similar to English question tags);

- YES-answer, NO-answer, Detailed Answer (to a WH-question), Extended Answer (adding something that exceeds the basic expectations of the asker);
- Turn Organization, Exchange Organization, Checking, Confirmation;
- Acknowledgment (confirming that some information has been received), Acceptance or Rejection (adds the element of evaluation).

3.2. Findings concerning selected classes of units

A subset of move categories relevant to this study is divided into two groups that are referred to as "positive responses" and "negative responses". Obviously, many of these units may play some other roles in various dialogue contexts and in different prosodic realizations as it is explained in the following paragraphs.

Table 2. "Positive" and "negative" responses under study.

General category	Some possible realizations	Move categories
positive response	*tak*, *no*, *aha*, *mhm*, *dobra*, *okej*	YES-answer, Confirmation, Acknowledgment
negative response	*nie*, *e-e*, *m-m*	NO-answer, Negation, Rejection

3.2.1. no

No often plays an emphatic role in colloquial Polish (e.g., as an emphatic particle). However, it can also be employed to perform a number of other discourse functions (see: Table 3). It may be incorporated into a bigger intonational phrase or function as a "stand-alone" unit. The position of *no* within a phrase varies. Among sixty-four cases of within-phrase occurrences, fifty-nine initial and four final realizations were found. Additionally, one phrase was both opened and closed with *no*. 164 cases of single *no* realized as a separate IP were found. Twenty cases of sequences comprising of two to four realizations of *no* (forming one IP) were found in the corpus. Besides its occurrences as a separate unit, *no* may join some other words, forming sequences like *no dobra*, *no tak*, *no to dobra*, realizing mostly Turn Organization or Exchange Organization moves.

Table 3. Dialogue moves realized by *no* (occurring as a separate WU).

Dialogue move category	Number of cases
Acknowledgment	109
YES-answer	14
Turn or Exchange Organization	15
Others (confirmation, readiness, emphatic, etc.)	26

For *no* sequences, it was found that in ten (among twenty) cases, the entire sequence had a rising melody, in five cases the melody was falling and in five it was flat. In the sequences comprising 2 to 4 realizations of *no* and showing an overall rising tendency, it was always the last *no* in the sequence that were perceivably high pitched (cf. Fig. 1).

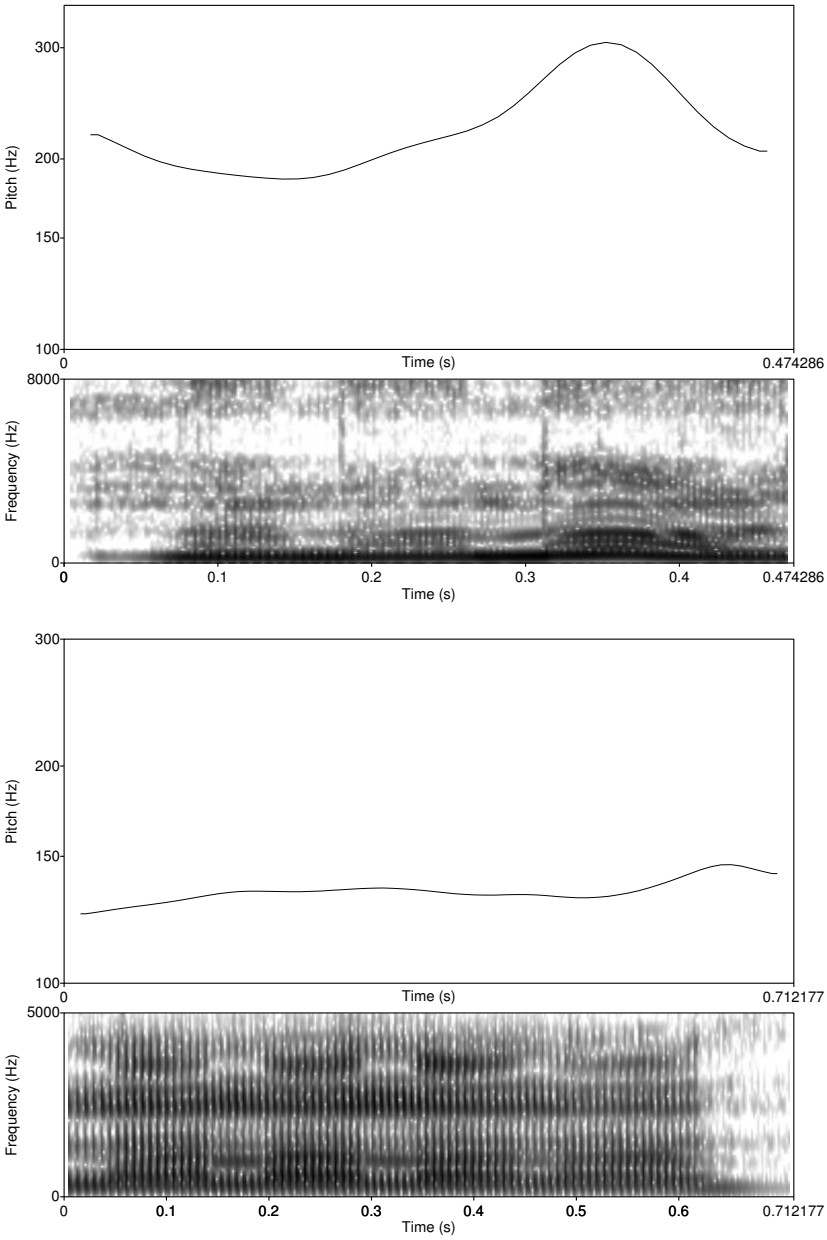

Fig. 1. Two examples of triple *no* realizations with perceivable rises on final syllables (i.e., on the last realization of *no*).

3.2.2. tak

This word is a close equivalent of English "yes". Similarly to "yes", it may play a number of discourse roles. Below, the moves realized by *tak* in the analyzed dialogues are listed.

Table 4. Dialogue moves realized by *tak* (occurring as a separate intonational phrase).

	Dialogue move category	Number of cases
1.	Acknowledgment	160 (including 9 multiple)
2.	YES-answer	76 (including 6 multiple)
3.	Checking for Acknowledgment/Understanding	46
4.	Confirmation (for checking moves)	24
5.	Other move categories (e.g., Turn Organization, Exchange Organization)	22

For Checking for Acknowledgment/Understanding, 44 out of 46 pitch contours were clearly rising, while only two were categorized as flat. For *tak* employed as YES-answer, among 64 contours (the remaining contours were rejected as impossible to analyze due to creaky voice or a low S/N ratio), 46 were falling, 15 flat, and 3 – rising. This difference is quite obvious when one takes into account that Checking for Acknowledgment/Understanding usually takes the grammatical form of a question.

While the mean duration of *tak* when used for Acknowledgment of Statement was shorter than when it was used for Acknowledgment of Instruction (385 ms vs. 432 ms), the difference turned out to be statistically meaningful only at 0.10. In the acknowledgments of statements, 50% pitch contours were rising, while in the acknowledgments of instructions/orders, that proportion exceeded 73%. The proportions of the falling contours reached 28% and 11%, respectively.

In Fig. 2, two realizations of *tak* are represented, a plain falling one and a compound, falling-rising one. The former seems to signal finality and to close an exchange. The latter simultaneously seems to function as a call for continuation and, on the affective level, as the expression of close listening to the conversational partner.

3.2.3. nie

This word may can be used similarly to *no* in English, but its other discourse roles may be quite different. For example, it is frequently placed at the end of a sentence in order to form a question for confirmation. Oddly, in colloquial Polish it can be replaced by "tak" in this position and function, with no substantial change to the meaning. However, for the purpose of this study, only those realizations of *nie* were extracted that functioned "independently", not as parts of such phrases. Surprisingly, only 23 single and four multiple realizations of *nie* meeting this condition were found.

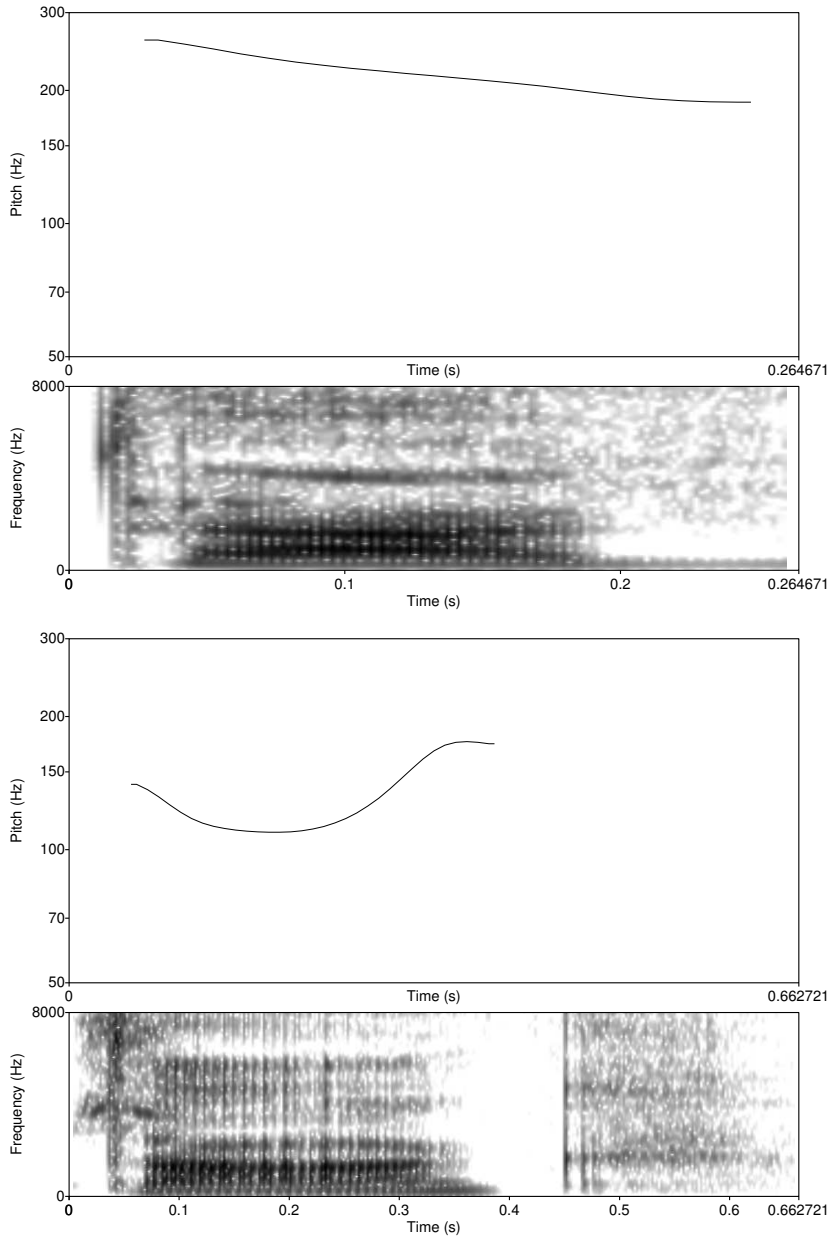

Fig. 2. A plain falling (female; upper panel) and a compound, falling-rising (male; lower panel) intonational realization of *tak*.

A small group of five naïve listeners, informally tested with 20 selected signals, confirmed the author's impression that, depending on the pitch contour shape and alignment, *nie* may be perceived as having one or two tones. The subjects were more prone to notice any rise when the pitch contour was from the class represented in the lower

panel of Fig. 3, not just a plain, gradual rise (as represented in the upper panel). Similar phenomenon may occur for other one-syllable quasi-words or fillers and it is worth further investigations.

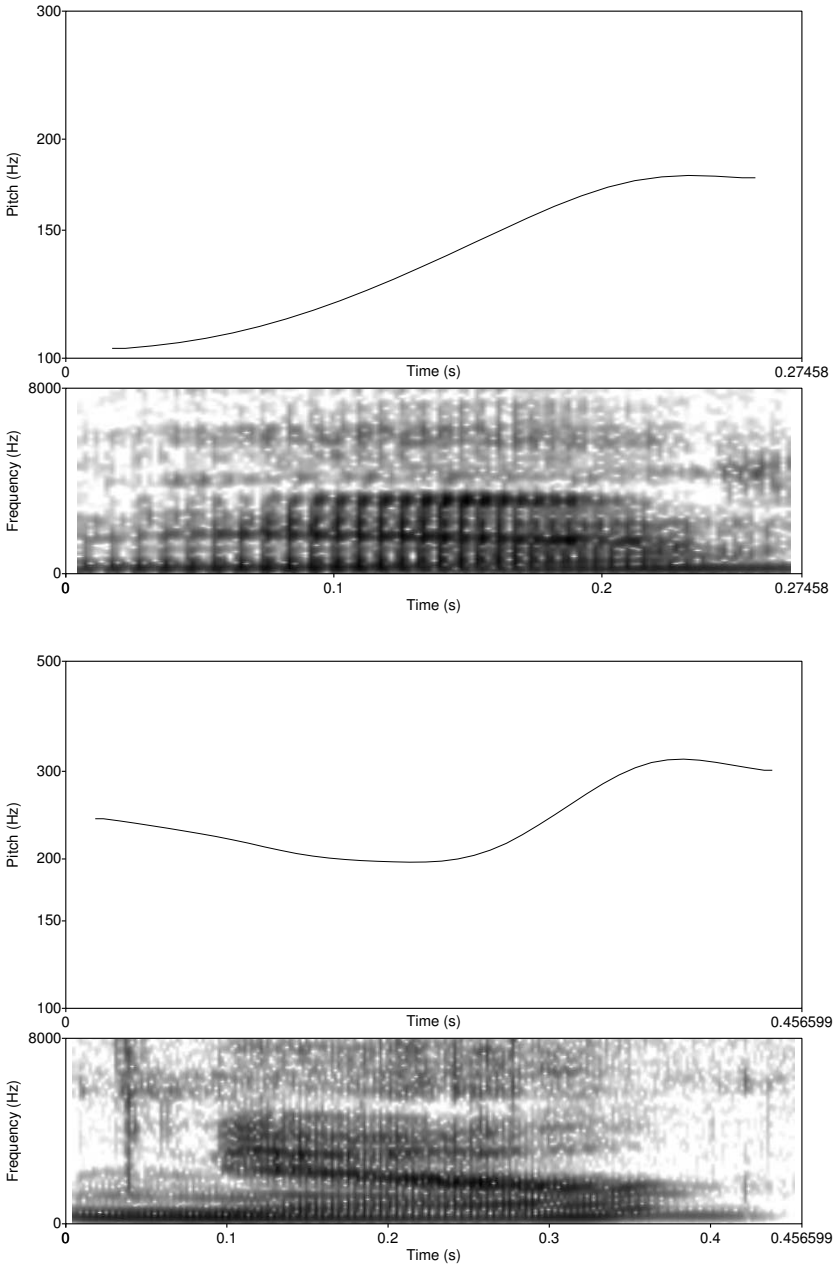

Fig. 3. Two varieties of rising contours in *nie*: The pitch rises gradually in time (upper panel); the rise starts later and it is steeper (lower panel).

3.2.4. mhm

The quasi-word *mhm* was used to perform mostly three dialogue moves: Acceptance/Acknowledgment, Confirmation, and YES-answer. In Table 5, the number of occurrences for each type of dialogue move as well as the proportion of rising pitch con-

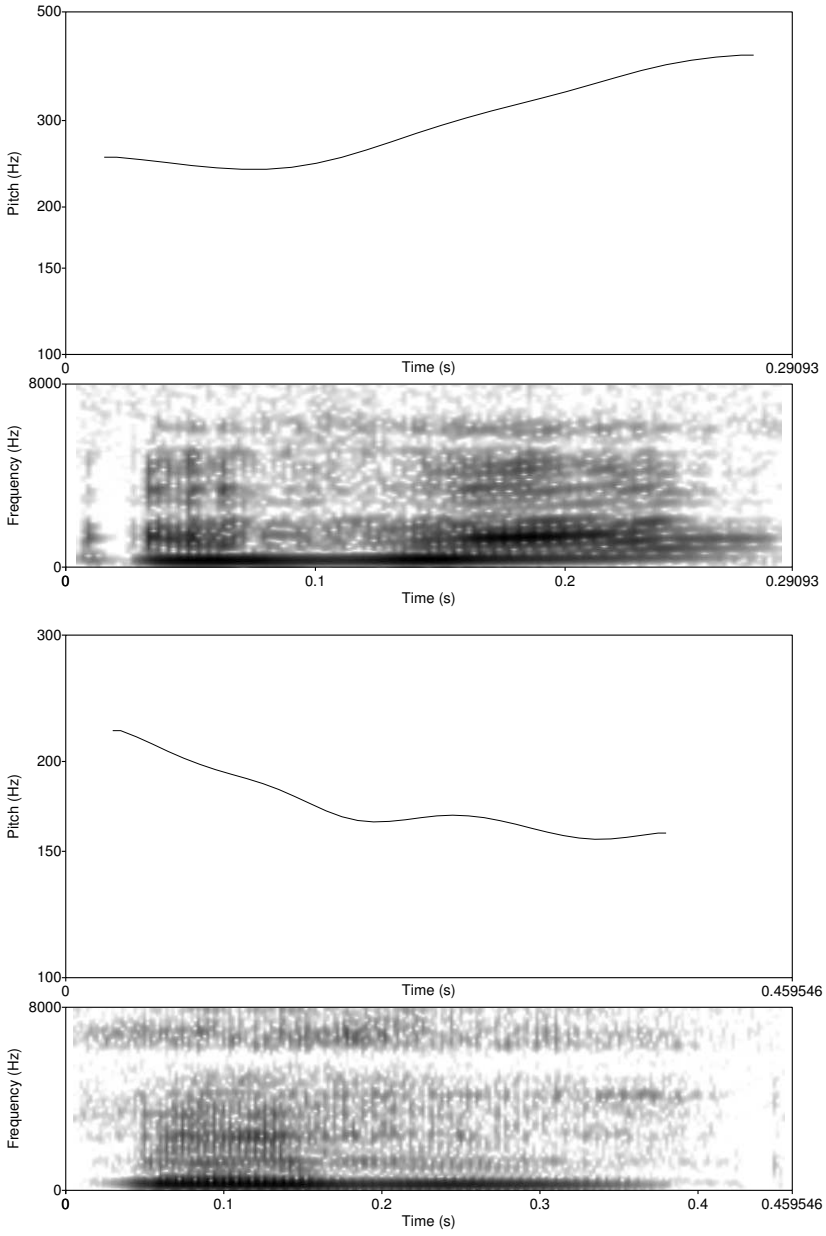

Fig. 4. A typical, rising realization of *mhm* by a female speaker (upper panel) and a rare example of a falling realization by a male speaker (lower panel).

tours in their realizations are listed. In Fig. 4, examples of a typical and an untypical intonational realizations of *mhm* are presented.

Table 5. Dialogue moves realized by *mhm.*

Dialogue move category	Number of occurrences	Proportion of rising pitch contours
Acknowledgment/Acceptance	120	80%
Confirmation	19	95%
YES-answer	16	100%

3.2.5. aha

This quasi-word was employed mostly for various subclasses of Acknowledgment, Acceptance, YES-answer and Confirmation moves. Among one hundred occurrences of *aha*, seventy-three realized various types of Acknowledgment/Acceptance moves. It was found that 52% of male *aha* realizations in this function had a falling pitch contour, while for female speakers this value reached only 33%. Two examples of *aha* realizations are shown in Fig. 5, the one shown in the upper panel being more emotional.

Two triple and one quadruple realization of *aha* were found. All of them functioned as Statement Acknowledgment/Acceptance moves.

While the data in hand are too limited to prove it statistically, one may put forward a hypothesis that Acknowledgment/Acceptance moves realized by *aha* more frequently have rising pitch contours when they simultaneously express a certain degree of surprise by a new or less expected piece of information. *Aha* with a rising pitch contour is mostly a response to an expected piece of information. Another situation in which rising realizations of *aha* occur are probably those that integrate Acknowledgment/Acceptance with a sort of prompt for further action (e.g., next instructions).

3.2.6. Monosegmental fillers

Monosegmental fillers are usually realized as centralized or nasalized vowel-like sounds, frequently reminding [ə]. In total, 365 stretches of speech including various realizations of MFs were extracted from the recordings. Forty-six of them contained sequences of two or more MFs produced within one IP. Only seven examples of MF sequences with two acoustically different, subsequent components (not separated by a perceivable pause) were found. Of course, they can be regarded as "bisegmental fillers", but being represented so sparsely, they are not analyzed further anyway. Their compound structure is most probably not intended to convey any additional meaning.

Fifty-two MFs that did not belong to larger WU were analyzed in more detail. Most of them were categorized as signs of "hesitation" or "prolonged thinking". Typically, they were realized as centralized or nasalized vowels, frequently with a glottal stop at the beginning. Those expressing hesitation were produced, as a rule, with a relatively flat pitch contour.

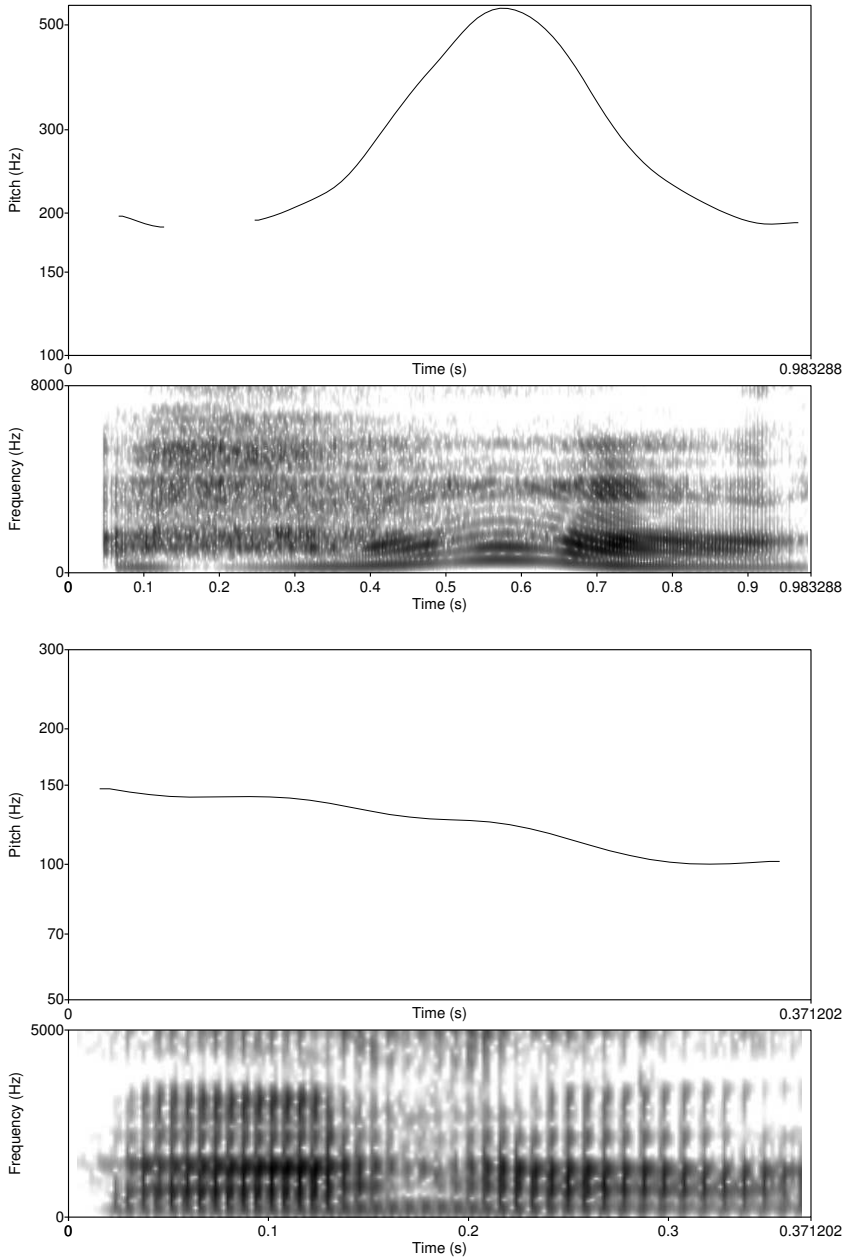

Fig. 5. An expressive realization of *aha* with a rising-falling pitch contour (upper panel) and a typical falling realization (lower panel).

Among 52 MFs that were not incorporated into any larger intonational unit, most expressed hesitation or signalized prolonged thinking/considering. There were only few examples of Statement Acknowledgment/Acceptation, Instruction Acknowledg-

Table 6. The positions of MFs realized as parts of IPs.

Position of MF	Number of cases
at the beginning of an intonational phrase	138
inside an intonational phrase	111
at the end of an intonational phrase	23

ment/Acceptance, and TurnOrganization. Their length varied from 102 to 1228 msec (mean = 605 msec with the standard deviation for the sample = 223 msec).

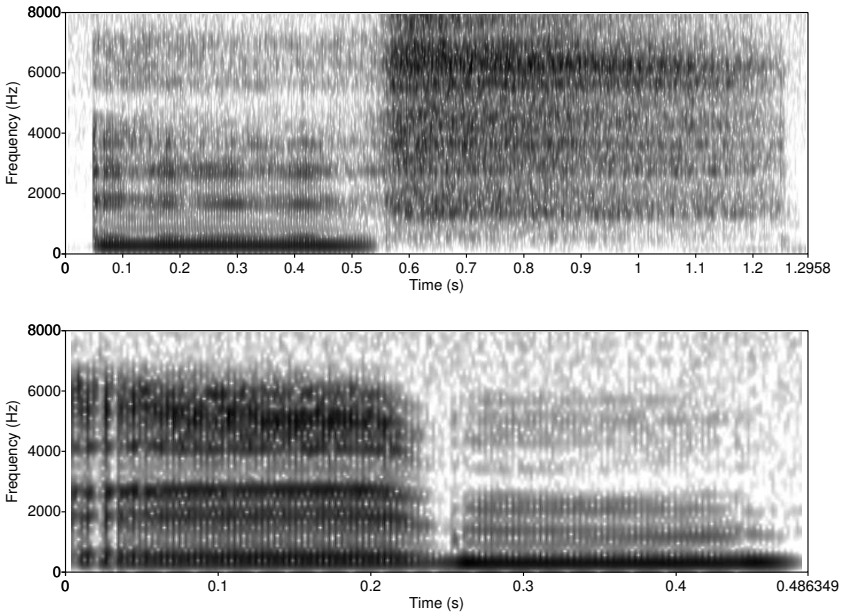

Fig. 6. Two compound fillers, each composed of two directly adhering monosegmental fillers (upper panel – approximate transcription [ɨ:f:]; lower panel – approximate transcription [ɨ:m:]).

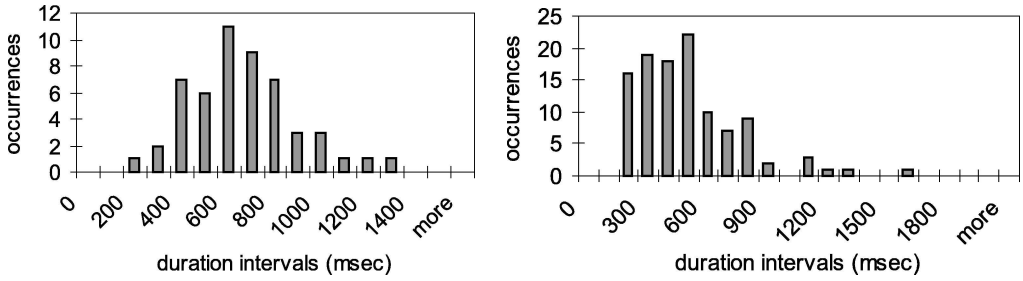

Fig. 7. The distribution of durations for MFs realized as separate units follows the normal pattern (left panel), but for the MFs occurring within IPs, it is skewed to the left (right panel).

3.2.7. Other units

Due to space limitations, some other units like *dobra*, *dobrze* or *okej* are not analyzed in this paper. However, one can mention that they usually realized Acknowledgment/Acceptance move. They were also used in Turn or Exchange Organization moves and to confirm the realization of an instruction or to signal speakers' readiness (similarly to "Done", "I've got it", "Ready"). While both rising and falling pitch contours were found in their realizations, any hypotheses that would explain their occurrences were not formed.

4. A comparative view

The research described here was recently extended to include two more languages, namely Korean and Thai. However, the comparative study was focused only on the moves collectively termed as "positive responses". We found that the average frequency of such units is similar in all the languages under study (from 6.5 per minute for Polish, through 7.2 for Korean to 7.5 for Thai). Also, the overall number of expression categories constituting the inventories of individual languages was almost identical (10 for Polish, 11 for Korean and Thai). Four or five most popular of them accounted for about 90% of all the analyzed tokens. We discovered Thai and Korean positive responses to be intonationally more consistent than those in Polish (within individual expression categories). It was also found that the nasal vocalization *mhm* /m̥m̥m/ was not only popular in the three languages but also quite consistently produced with a rising melody. Another surprising founding was that the normalized pitch change, calculated as $(F_{0\,\max} - F_{0\,\min})/F_{0\,\max}$, reached no more than 40% for the Asian languages, while in the case of Polish we noted examples of 60% change and higher average values of this coefficient. More detailed results are presented in KARPIŃSKI, KLEŚTA, SZALKOWSKA [11].

5. Conclusions

The findings described in the present paper support the thesis that quasi-words and fillers are important components of task-oriented dialogues. Although the repertoire of discourse function they realize is surprisingly wide, the number of their most popular applications is quite limited. Consequently, even with a relatively large corpus, one may face the problem of categories represented by a very small number of realizations available for analysis (MÖBIUS, [21]). The intonational realizations of the units in question are extremely varied and it is obvious that in some cases pitch contour changes may lead to dialogue move category shift. However, the shape of intonational contours of the units in question is linked not only to dialogue moves they realize, but also to other factors like gender or emotional expression. The boundary between the linguistic and non-linguistic intonation, discussed in depth by GUSSENHOVEN [9], is quite hazy here.

While it is possible to discuss it in terms of the three biological codes (GUSSENHOVEN [9], few practical cues for the differentiation methods have resulted from this approach so far.

Further research in this field may encompass a more detailed analysis of intonational contours (not only flat, rising and falling categories). Another important step would be to undertake other studies of the phonetic form of the units in question so that they can be effectively identified by speech recognition systems or more precisely synthesized in Text-to-Speech applications. KIM [14] studies some approaches to the automatic detection of disfluencies and fillers in spontaneous speech. The general aim is to improve the performance of a speech-to-text system. FISCHER and WREDE, [7] proved that phonetic and linguistic modelling of such units might be very important in human-computer communication systems. On the other hand, this knowledge becomes crucial in any attempts to analyze semi-spontaneous interpersonal communication. Since the non-linguistic component of intonation seems to be especially prominent in some of these units, they may be used to determine the psychological state of the speaker. In general, studies of this kind may result in rethinking the concept of the language system and to enriching it with components that have been regarded as non- or paralinguistic so far.

Acknowledgment

This study is part of the Pol'n'Asia project, funded by the Polish Ministry of Science and Higher Education (project code H01D 006 27).

References

[1] BARD E. G., LICKLEY R. J., *On not Remembering Disfluencies*, Proc. Eurospeech 97, Rhodes, pp. 2855–2858, 1997.

[2] BOERSMA P., WENINK D., *Praat: Doing phonetics by computer (Version 4.3.16)*, [Computer program], Retrieved June 22, 2005, from http://www.praat.org/.

[3] BYRON D. K., HEEMAN P. A., *Discourse marker use in task-oriented spoken dialog*, Proc. Eurospeech 97, pp. 2223–2226, 1997.

[4] CARLETTA J., ISARD A., ISARD S., KOWTKO J., DOHERTY–SNEDDON G., ANDERSON A., *HCRC dialogue structure coding manual*, Human Communications Research Centre, University of Edinburgh, Edinburgh, HCRC TR – 82, 1996.

[5] CARLSON L., *Dialogue Games: An approach to discourse analysis*, D. Reidel, Boston 1983.

[6] CRISTEA D., WEBBER B., *Expectations in Incremental Discourse Processing*, Proc. the 35th Annual Meeting of the Association for Computational Linguistics, Madrid 1997.

[7] FISCHER K., WREDE B., *Discourse particles in female and male human-computer-interaction*, Proceedings of Women into Computing, Intellect Press, Exeter 1997.

[8] GOLDMAN–EISLER F., *Hesitation and information in speech*, [in:] Information Theory, C. Cherry [Ed.], pp. 162–174, Butterworth, London 1961.

[9] GUSSENHOVEN C., *Intonation and Interpretation: Phonetics and Phonology*, [in:] Speech Prosody 2002, B. Bel, I. Marlien [Eds.], pp. 47–57, 2002.

[10] GUSSENHOVEN C., *The Phonology of Tone and Intonation*, CUP, Cambridge 2004.

[11] KARPIŃSKI M., KLEŚTA J., SZALKOWSKA E., *Non- and Quasi-lexical Realizations of "Positive Response" in Korean, Polish and Thai*, [in:] Pro. Speech Prosody Conference, R. Hoffmann, H.-J. Mixdorff [Eds.], Dresden, PS7-16-133, 2006.

[12] KARPIŃSKI M., KLEŚTA J., *The project of an intonational database for the Polish language*, [in:] Prosody 2000, St. Puppel, G. Demenko [Eds.], Wydział Neofilologii UAM, pp. 113–118, Poznań, 2001.

[13] KREUTEL J., MATHESON C., *Obligations, intentions, and the notion of conversational games*, [in:] Information States, Obligations and Intentional Structure in Dialogue Modelling, Proc. the 3rd International Workshop on Human-Computer Conversation, Bellagio 2000.

[14] KIM J., *Automatic Detection of Sentence Boundaries, Disuencies, and Conversational Fillers in Spontaneous Speech*, MSc Thesis, University of Washington, 2004.

[15] LICKLEY R. J., BARD E. G., *On not Recognizing Disfluencies in Dialogue*, Proceedings of the ICSLP, Philadelphia 1996.

[16] LICKLEY R. J., *Missing Disfluencies*, *Proc. the ICPhS*, Stockholm, **4,** 192–195, 1995.

[17] LICKLEY R. J., *Juncture Cues to Disfluency*, Proc. the ICSLP, Philadelphia 1996.

[18] MANN W. C., *Dialogue Games*, USC Information Sciences Institute, Marina del Rey, CA, ISI/RR-79-77, 1979.

[19] MANN W. C., *Models of Intentions in Language*, 5th Workshop on Formal Semantics and Pragmatics of Dialogue, Bielefeld 2001.

[20] MANN W. C., *Dialogue Macrogame Theory*, SIGdial Workshop 2002, Philadelphia 2002.

[21] MÖBIUS B., *Rare Events and Closed Domains: Two Questionable Concepts in Speech Synthesis*, [in:] Research Papers from the Phonetics Lab, **6,** 4, 119–134 (2000).

[22] OWEN M., *Conversational units and the use of 'well...'*, [in:] Conversation and Discourse: Structure and Interpretation, Paul Werth [Ed.], St. Martin's Press, pp. 99–116, 1981.

[23] ROSE R. L., *The communicative value of filled pauses in spontaneous speech*, MA Thesis, University of Birmingham, 1998.

[24] SCHIFFRIN D., *Conversational coherence: The role of 'well'*, Language: Journal of the Linguistic Society of America, **61**, 3, 640–665 (1985).

[25] SCHIFFRIN D., *Discourse markers*, CUP, Cambridge 1987.

[26] SWERTS M., WICHMANN A., BEUN R., *Filled pauses as markers of discourse structure*, Proc. ICSLP, 1033–1036, 1996.

[27] TRAUM D. R., ALLEN F. J., *Discourse obligations in dialogue processing*, Proc. ACL '94, P94-1001, 1994.

[28] TRAUM D. R., HEEMAN P. A., *Utterance units in spoken dialogue*, [in:] Dialogue Processing in Spoken Language Systems, E. Maier, M. Mast and S. Luperfoy [Eds.], Springer-Verlag, pp. 125–140, Heidelberg 1997.

[29] WARD N., *Non-Lexical Conversational Sounds in American English*, Pragmatics and Cognition, **14**, 1, 129–182 (2006).