A Preliminary Study of the Intonational Phrase, Nuclear Melody and Pauses in Polish Semi-Spontaneous Narration

Katarzyna Francuzik, Maciej Karpiński, Janusz Kleśta

Institute of Linguistics
Adam Mickiewicz University, Poznań
{kasiafr; maciejk; janklest}@amu.edu.pl

Abstract

This paper contains preliminary observations concerning the properties of the intonational phrase, the nuclear melody, and pauses in semi-spontaneous texts of short narratives in Polish.

1. Introduction

While the prosodic differences between spontaneous and read or formal texts have been intensively studied in a number of languages (e.g., [1]), very little attention has been paid so far to the prosody of Polish spontaneous speech. The majority of earlier Polish works were based on pre-prepared sets of utterances or even on quite arbitrary observations. Old corpora may therefore occur irrelevant to certain phenomena in the contemporary informal speaking styles. Owing to improved recording and analysis technologies, a more thorough investigationt has become possible now, which however may also demand a new methodological approach, too. More recent works on Polish intonation by Jassem and Demenko (e.g., [2, 3, 4]) are technologically-minded (speech synthesis), so they are based mostly on carefully pronounced or modeled speech. Nevertheless, their methodology seems to be more universal and useful also for the current study. An optional approach focused on the melody of the entire intonational phrase (e.g., [5]), is more complex and less convenient for application purposes. Moreover, for a number of languages, it has been argued that the crucial role in the domain of discourse is attributed mostly to the nuclear melody [6, 7].

The aim of this study is to review some of the basic prosodic phenomena that occur in semi-spontaneous spoken Polish. It is expected that some of the presented observations will be soon statistically confirmed with larger corpora. The PoInt database [8] will certainly offer a wider insignt into the intonation of spontaneous Polish, with its all irregularities and phenomena belonging to the LNRE class of distributions. The prosody of spontaneous speech is interesting not only from a purely cognitive viewpoint. The characteristics of spontaneous speech should also be taken into account in advanced automatic speech recognition, understanding, and synthesis systems [9, 10].

2. Speakers and recordings

The signals used in this research come from the PoInt corpus [8]. They were recorded digitally in an anechoic chamber, using Tascam CD-RW700 recorder, an AKG condenser microphone, and a Spirit Folio mixing console.

Each subject was asked to read a short, nine-picture comic strip about Goofy and Clarabella. Afterwards, the cartoon was hidden and the subjects were asked to tell the story they had read. The recordings included thirteen female and thirteen male speakers, aged between 20 and 40, with at least college education. None of them showed any serious speech deficiencies.

The speech in the recordings is referred to as "semi-spontaneous", because the spontaneity was limited by a number of factors, e.g. the “unnatural” setting of the anechoic chamber, or the fact that subjects were asked to tell a specific story, based on a given pictorial and textual material. On the other hand, a few steps were undertaken to ensure maximum possible spontaneity: (a) one of the experimenters was in the anechoic chamber so that the speaker could speak to a listener rather than to the microphone; (b) the experimenters attempted to ensure a relaxed atmosphere during the entire session; (c) the narration of the comic strip was the fourth consecutive task of the session, and the speakers were already at least partially accustomed to the recording environment.

The recordings were converted from .cda to .wav files (mono, at the sampling frequency of 44.1 Hz and 16-bit resolution). Their length varied from 21 to 78 seconds, both extreme values being observed for female speakers.

3. Annotation

3.1. Orthographic transcription

In the first step, the signals were transcribed orthographically. Standard word forms were used, regardless of actual pronunciation. Mispronounced words were marked only when it was impossible to understand them even with the aid of the context and co-text. Some words were labeled as “unintelligible”, i.e. there were not enough cues to recognize them with a reasonable probability even in the given context. Filled and unfilled pauses were also transcribed, and their length was measured. In sum, the orthographic transcription was not very problematic in this case, because the context of the story was clear, the speakers used a limited vocabulary, and simple syntactic constructions.

3.2. Pauses

Three classes of pauses were distinguished: (a) “silent” unfilled pauses (SUP), with no perceivable breath noise; (b) unfilled pauses with audible breath noise (BUP); (c) filled pauses (FP), with various types of non-lexical “fillers”. The correct classification of a given pause was not always quite obvious to the labelers not always. For example, the breath noise could be on the level not much higher than the ambient noise. When, within in a silent pause, one or a few separate, very short speech sounds occurred, it was categorized as a “mixed pause”.

The minimum length of a perceivable pause is not fixed, as it depends on many contextual factors. While the threshold

is usually set between 250 and 1000 ms [11], this value may well drop down below 200 ms in careful listening. Although even shorter pauses were labeled in the analyzed texts, only those of at least 200 ms were finally considered.

3.3. Intonational phrases (IPs)

Three labelers independently cut the signals into intonational phrases and discussed these controversial cases. The procedure was based entirely on the auditory perception. In the procedure of segmentation into IPs, a number of more or less formal cues were taken into account [2, 4, 11]. Nevertheless, many hypothetical phrase boundaries were controversial. A number of filled pauses and unfinished words were perceived as separate IPs, but they were excluded from further analysis. Certain IPs seemed to be "unnaturally long", but there was no clear indication that they could be cut into more IPs.

3.4. Nuclear melody

The intonational annotation was partially based on the strategy of the IViE system [12], but only one prominence (i.e., the nuclear accent) per IP was annotated, and this system was applied only at the phonetic level. Then, the results were reviewed and categorized into the classes which are supposed to be relevant to the phonological system of Polish nuclear intonation [13]. Since the system is still being tested for spontaneous speech (by W. Jassem), in some analyses it seemed safer to apply only broader categorizations (e.g., falling, rising, and flat pitch contours) or to operate at a lower level (e.g., phonetic).

The location of each nuclear accent was determined and discussed by two labelers. They labeled the signals independently and discussed the results in order to arrive at a common conclusion. The labeling procedure was chiefly auditory, based on careful listening. The target syllables (the pre-accentual, the accented and the post-accentual one) were extracted from the signal and listened to separately, in various sequences, to determine their relative height or pitch movement that occurred within each of them. If the last of these syllables was not the final syllable of the phrase, the remaining part of the IP was also analyzed for the direction of the pitch movement and the pitch level at the boundary. Additionally, Praat [14] was used to generate the graphic representation of the pitch movements. Another helpful function of Praat was "enhance", which played the resynthesized utterances at a slower rate, preserving their pitch height. It was used only for extremely fast speech.

Many signals could not be analyzed instrumentally, as the pitch extraction algorithm failed for creeky voice, low signal levels, and devoiced syllables. Frequency doubling/halving errors were partially handled by the error-correction procedure provided by Praat. All these phenomena occurred most frequently in the final syllables of IPs. In such a situation, the auditory analysis was considered as primary, while the instrumental one was meant just to confirm it or to provide additional cues.

Since too many signals were excluded from the instrumental analysis, most of the following findings are based on "impressionistic" annotation.

4. Findings

4.1. Intonational phrases

The labelers annotated 526 "regular" IPs. This number comprises only "lexical phrases", i.e., those based on at least one word. As has been mentioned, filled pauses (or sequences of filled and unfilled ones) could sometimes be considered as IPs, but they were excluded from further analysis in this paper. The length of the entire narration varied from nine to thirty-three "normal" IPs. Occurrence

The distribution of the phrase length is represented in Figure 1. The diagram does not show a small number of extreme length values (over 15 syllables). The average length of the IP was 7.44 (at SD=3.9), with no statistically meaningful difference between the male and female voices.

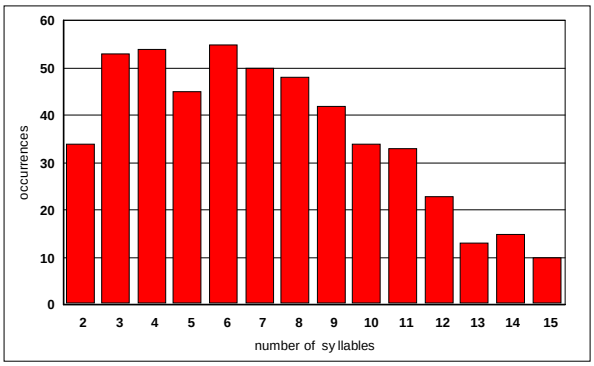

Figure 1. The distribution of the IP length.

4.2. Pauses

Altogether, 230 SUPs, 115 BUPs and 148 FPs were found in the analyzed passages. The average length was, respectively, 783 ms, 893 ms, and 591 ms (with respective standard deviations of 66.2, 43.7, and 53.3). Figure 2 illustrates the distribution of BUPs and SUPs, while Figure 3 shows the distribution of FPs. The obvious difference in the distributions of BUPs and SUPs supports the approach of treating them separately.

In a study of pauses in French by Grosjean and Deschamps [15], the mean duration of silent pauses in the description of a cartoon reached 1320 ms, while in an interview this value dropped to 520 ms. As many researchers, suggest, "the more difficult the communicative situation, the more pauses, hesitations, and stuttering events are likely to occur" [16]. The correlation of the length of the pauses only with the lexical access time and to the grammatical structure building/parsing time is an oversimplification, especially in spontaneous speech. The values obtained in this study may suggest that – if we consider the length of pauses as directly comparable across languages – the discussed task was relatively easy. This may also support the assumption that a remarkable degree of speech fluency was achieved.

It was found that the filled pauses tend to co-occur with the unfilled ones. Only 25% of the filled pauses occurred without a preceding or a consequent unfilled pause. More than 65% of pauses occurred in sequences. Some speakers produced, as a result of their hesitations, extremely long sequences of various types of pauses, like BUP-FP-FP-BUP-

FP-SUP-FP-SUP or BUP-FP-SUP-FP-BUP (both in female speakers). Certain FPs, as in the first of the above two example, were adjacent but noted separately, for their boundary was very prominent.

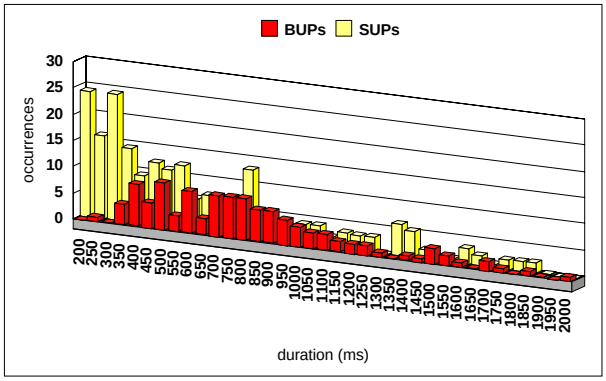

Figure 2. The distribution of BUPs and SUPs

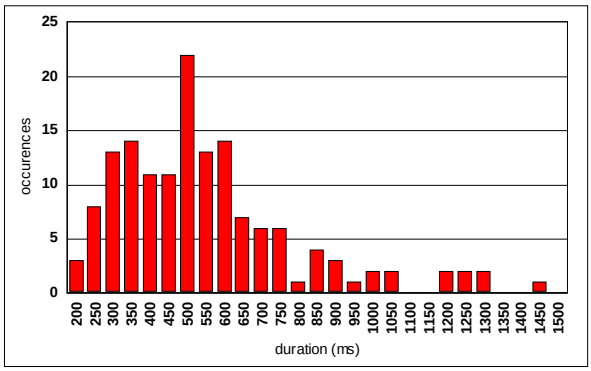

Figure 3. The distribution of FPs

Although the correlation between the length of pauses and the degree of cohesion has not been systematically studied here, the general tendency described in [17] seems to be preserved. Nevertheless, in a number of cases, syntactic and intonational phrases were broken into parts by internal pauses. In a few cases, due to strong hesitations, pauses occurred within words. In case of one of the female speakers, a within-word pause reached 600 ms (!). Only for four speakers no IP contained internal pauses of any type. In eleven speakers, compound pauses, i.e., pauses built of at least two pauses of different types, or, in the case of FPs, if their boundary was vividly perceived. The analyzed pieces of narration by different speakers contained from 6 to 29 pauses or pause sequences.

4.3. Observations concerning the nuclear melody

In the vast majority of the labeled phrases, the nuclear accent fell on the penultimate syllable of the last word in the IP. In some cases, the last word was comprised of one syllable which was accented. In 28 Ips, the nuclear accent was more distant from the end of the phrase. At the boundary, it reached "low", "mid" and "high" level, respectively, 15, 7 and 6 times.

Only 34 instances of an easily perceivable pitch movement within a syllable were labeled, and most of them occurred in the postaccentual syllable, while only 3 in the accented one. Most of them (26) were labeled on the phonetic level as an "mh" pitch movement. The pitch movement within the postaccentual syllable was a kind of approximation to the final pitch level (e.g., lMmh, lLmh), but it was prominent because of the syllable's length or other qualities. In such a situation, only seven classes of nuclear melodies were illustrated in Figure 4.

Figure 4. illustrates the distribution of the basic types of the nuclear melodies. Two most frequent patterns, HM and MH, are clearly visible. Low rising pattern was found only in 32 cases, while ML and MH occurred, respectively, 103 and 102 times.

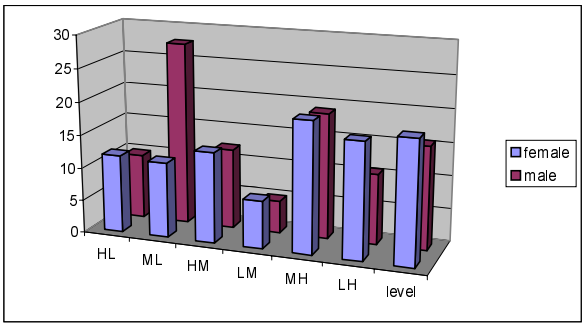

Figure 4. The proportions (%) of the nuclear melody categories in male and female speakers

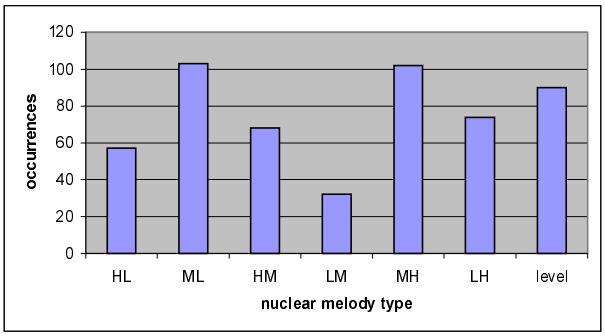

Figure 5. The distribution of the nuclear melody types

A stunning disproportion in the ML category can be observed between males and females. This can hardly be explained by the labelers' tendency to annotate male voices as "operating at lower levels", because the movement of the same range, i.e., LM, is more frequently observed in the female speakers. Only in the LH category, female speakers substantially prevailed, which may produce the impression of a "more emotional" voice.

4.4. A comment on the discourse relations and the nuclear melody patterns.

The discussed narratives, although short and simple in the expressed ideas, posed a number of descriptive problems at the level of discourse relations which, in their case, was frequently deeply distorted and probably far from the listener's expectations. Still, they remained intelligible.

Except for numerous instances of "consequence", it was hard to find clear enough examples of other discourse

relations. The texts were tested for two basic relations that seemed to be relatively easy to find and classify: "explanation" and "contrast". The relation of explanation in Polish can be marked by words like "ponieważ", "bo", "bowiem", "albowiem", "gdyż" (various equivalents of "because"), and the relation of contrast can be marked by "ale", "lecz" ("but"), but only few of these words are present in colloquial speech. Only 18 cases of "explanation" and 20 cases of "contrast" were found.

For the "explanation", no tendency was discovered. There was no regularity among the nuclear melody of the in the first or in the second phrase. It was also impossible to find any regularity in the relations between the nuclear melody of the first and the second phrases. Two times the word "ponieważ" was produced as a separate, rising IP.

In the case of "contrast", the nuclear melody of the first phrases of the analyzed pairs did not show any regularity. However, the nuclear melody of the second phrases of the pairs was usually falling (14 falling contours, 3 flat contours and 3 rising contours). Even in if this result seems quite convincing, the population was too small to carry out reliable statistic tests.

It may be hypothesized that certain discourse relations allow the speaker a wider choice of intonational realizations, while others are more "restrictive". Another hypothesis may be that the intonation of the second phrase of the related pair is more restricted. In the case of "contrast", this can be explained as follows: Not always is the entire stretch of speech (the pair of phrases) planned at once. The first phrase may be produced with any intonation; then, one decides to add another phrase, which was not consciously planned before, and which is attached to the preceding phrase with a selected discourse marker. In this context, and with a view to what is going to be said in the second phrase, the speaker may be more careful in the choice of intonational patterns. In general, it seems that appropriate intonational patterns may facilitate understanding and even convey additional information, but if they are distorted, the most crucial part of the message still can be conveyed via other means.

5. Conclusions

This study is another piece of evidence that in the research on spontaneous speech there is a need for very large corpora. We are able to find and statistically confirm certain trends only using huge collections of data. The problematic account of the discourse relations (above) can also be attributed to the limited size of the employed corpus. While the crucial role of the nuclear melody in discourse seems to be confirmed for some languages, it is still probable that, in certain circumstances, the shape of the entire intonational contour may be important. Accordingly, a part of the discussed material has been already labeled in terms of the general phrase melody and it is going to be analyzed on the discourse level, too.

The presented facts can be considered as typical of a number of types of spoken texts that convey sequences of actions. Similar phenomena may be expected in a spoken description of a roughly pre-planned travel route, etc. Some of the findings may possibly be confirmed for other classes of semi-spontaneous or even fully spontaneous texts.

Although the scope of this study and the number of findings are limited, it seems to offer a starting point for further research on the prosody of spontaneous Polish. Similar analyses will be carried out for other varieties of spoken Polish, represented in the PoInt database, e.g., the description of a painting, comments on a piece of music, a free narration on one's own life, or even conversational language.

This research was supported by KBN (1HO1D01118).

6. References

[1] Ayers M. G., 1994. Discourse functions of pitch range in spontaneous and read speech. *Ohio State University Working Papers in Linguistics*, No. 44, 1-49.

[2] Demenko, G., 2000. Automatic analysis of phrase boundary in Polish. In: W. Jassem, Cz. Basztura, G. Demenko (eds.) *Speech and Language Technology*, vol. 4. Poznan: Polish Phonetic Association.

[3] Jassem, W., 1987. Computer-based classification of basic Polish intonations. *Proceedings of the 11th International Congress of Phonetic Sciences*, Talin, 253-256.

[4] Demenko, G.; Jassem, W., 1997. Phonetic and syntactic coherence of the phrase. *In Speech and Language Technology*, vol. I, W. Jassem, Cz. Basztura (ed.). Wrocław: FORMAT, 125-139.

[5] Steffen-Batogowa, M., 1996. *Struktura przebiegu melodii języka polskiego ogólnego*. Poznań: SORUS.

[6] Mayer, J. 1997. Intonation und Bedeutung. Phonetic AIMS 2.3, IMS Universitaet Stuttgart.

[7] Pierrehumbert, J.; Hirschberg, J., 1990. The meaning of the intonational contours in the interpretation of discourse. In: Cohen, Morgan, Pollack (eds.), *Intentions in Communications*, Cambridge: MIT Press, 271-311.

[8] Karpiński, M.; Kleśta, J., 2001. Intonational Database for the Polish Language. *Proceedings of Prosody 2000 Workshop*, Kraków.

[9] Niemann, H., Noeth, E., Batliner, A., et al., 1998: Using prosodic cues in spoken dialog systems. International Workshop "Speech and Computer", St. Petersburg.

[10] Karpiński, M., 1999. Operowanie środkami fonetycznymi w wypowiedziach skierowanych do komputera. In: W. Jassem, Cz. Basztura, G. Demenko, K. Jassem (eds.) *Speech and Language Technology*, vol. 3. Poznań: Polish Phonetic Association.

[11] Cruttenden, A., 1986. *Intonation*. Cambridge: Cambridge University Press.

[12] Grabe, E., 2001. The IViE Labelling Guide, ver. 3. http://www.phon.ox.ac.uk/esther/ivyweb/guide.html

[13] Demenko, G., 1999. *Analiza cech suprasegmentalnych języka polskiego na potrzeby technologii mowy*. Poznań: Wydawnictwo Naukowe UAM.

[14] Boersma, P., 2001. Praat ver. 3.9, http://www.praat.org

[15] Grosjean, F.; Deschamps, A., 1975. Analyse contrastive des variables temporelles de l'anglais et du français. *Phonetica*, 31, 144-184.

[16] Zellner, B., 1994. Pauses and temporal structure of speech, In: E. Keller (Ed.) *Fundamentals of speech synthesis and speech recognition*, Chichester: John Wiley, 41-62.

[17] Grosjean, F.; Grosjean, L.; Lane, H., 1979. The patterns of silence: Performance structures in sentence production. *Cognitive Psychology*, 11, 58-81.