Polish rhythmic database – new resources for speech timing and rhythm analysis

Agnieszka Wagner, Katarzyna Klessa, Jolanta Bachan
Institute of Linguistics, Adam Mickiewicz University in Poznań
Poznań, Poland
wagner@amu.edu.pl, katarzyna@klessa.pl, jolabachan@gmail.com

Abstract

This paper reports on a new database – *Polish rhythmic database* and tools developed with the aim of investigating timing phenomena and rhythmic structure in Polish including topics such as, inter alia, the effect of speaking style and tempo on timing patterns, phonotactic and phrasal properties of speech rhythm and stability of rhythm metrics. So far, 19 native and 12 non-native speakers with different first languages have been recorded. The collected speech data (5 h 14 min.) represents five different speaking styles and five different tempi. For the needs of speech corpus management, annotation and analysis, a database was developed and integrated with *Annotation Pro* (Klessa et al., 2013, Klessa, 2016). Currently, the database is the only resource for Polish which allows for a systematic study of a broad range of phenomena related to speech timing and rhythm. The paper also introduces new tools and methods developed to facilitate the database annotation and analysis with respect to various timing and rhythm measures. In the end, the results of an ongoing research and first experimental results using the new resources are reported and future work is sketched.

Keywords: speech rhythm, timing, Polish

1. Background and motivation

The current knowledge on timing patterns and speech rhythm in Polish is incomplete. Many issues have not been investigated yet, e.g. phonotactic properties of rhythm at the phrase level, cross-linguistic perception and production of rhythmic structure and grouping in Polish as first (L1) and second/foreign language (L2), or have been studied only within a limited scope, e.g. the effect of tempo on speech timing (Malisz, 2011), acoustic and perceptual correlates of prosodic prominence and phrasing (Demenko, 1999; Wagner, 2008; Malisz & Wagner, 2012) or rhythmic classification of Polish (Ramus et al., 1999; Grabe & Low, 2002; Mairano & Romano, 2011). The limitations of the current state of knowledge can be attributed to limitations of the available language resources.

In recent years, large databases of spoken Polish have been collected such as *Jurisdict* (Demenko et al., 2008), *Diagest* (Jarmołowicz et al., 2007) or *Paralingua* (Klessa et al., 2013). However, these corpora were designed either for a specific application (dictated speech for a speech recognizer – *Jurisdict*) or task (emotion portrayals – *Paralingua*), or as quasi-spontaneous conversations under variable recording conditions and mutual interlocutor visibility (*Diagest* and *Paralingua*). Such recording scenarios allow for studying only those aspects of speech rhythm that are available for the respective speaking styles.

In order to enable systematic analyses of timing- and rhythm-related phenomena in Polish, we have developed a new corpus consisting of data characterized by a highly controlled structure with respect to rhythmic organization of utterances: *Polish rhythmic database*. Additionally, we propose new tools and methods to facilitate the database annotation and analysis with respect to various timing and rhythm measures.

2. Data collection

2.1 Design criteria

With a view to application of the database to specific analyses and based on the results of previous studies (Dellwo et al., 2004; Wiget et al., 2010; Prieto et al., 2012; Arvaniti, 2012) we formulated the following criteria:

- inclusion of different elicitation methods and speaking styles,
- representation of different phonotactic, prosodic and rhythmic structures,
- speaking rate variability,
- a large sample of sentences/text materials – to provide data representative of the relevant phonological and metrical properties of Polish,
- a large number of speakers and speech samples – to ensure that the compiled corpus is representative of the Polish language,
- inclusion of non-native speakers of Polish with L1s representing different rhythm types.

2.2 Materials

The structure of the speech corpus is determined by:

- accent (Polish L1, Polish L2 and other L1s: German: a typical stress-timed language, Korean: rhythmically mixed/unclassified, Spanish: syllable-timed),
- speaking style,
- speaking rate.

The Polish L1 subcorpus includes a story ("The North Wind and the Sun"), three sets of five sentences, two mini-dialogues, excerpts from four poems and spontaneous monologue. The selection of "The North Wind and the Sun" was motivated by the fact that this text was used in many cross-linguistic studies on timing and rhythm. This makes it possible to compare the results of the analyses based on *Polish rhythmic database* with those reported elsewhere (e.g., Arvaniti, 2012; Grabe & Low, 2002; Ramus et al.,

1999; Dellwo & Wagner, 2003). Moreover, the translations of the story are available for many languages, including Spanish, German and Korean. For the sentence subcorpus we created "stress-timed" and "syllable-timed" sentences representing the maximum and the minimum phonotactic complexity, respectively, and phonotactically "uncontrolled" sentences representing "average" syllable complexity. The sentences were constructed using frequency and distribution lists of Polish syllables with word examples (Śledziński, 2013). All sentences had similar length (17-22 syllables) and syntactic and prosodic structure. The mini-dialogues were designed to incorporate different sentence types and prosodic structures. The sentence subcorpus and mini-dialogues are composed of vocabulary adjusted to the Polish L2 speakers' level. The texts in the poem subcorpus represent different poetic meters: iambic, trochaic, a hexameter and an amphibrach (Kulawik, 1997). Spontaneous monologue was elicited using two diapix tasks which consisted in finding differences between two highly similar pictures (Bradlow et al., 2007). This method has the advantage of providing some control over the lexical content of the elicited speech and similar amount of speech data from each speaker. The story, sentences and poems were recorded by Polish L1 speakers at five different tempi. For the needs of studying cross-linguistic perception and production of prominence and phrasing, stability of rhythm metrics and rhythmic classification of Polish, we also recorded non-native speakers in Polish and in their L1s (see: Sec. 2.4).

2.3 Speakers

Nineteen Polish native speakers (eight males and eleven females) and twelve Polish L2 speakers were recorded. The native speakers were students of the Institute of Linguistics at Adam Mickiewicz University in Poznan (AMU). They were all monolinguals from the region of Greater Poland and used the Cracow-Poznan pronunciation variant.

The L2 speakers were three German, five Korean and four Spanish learners of Polish as a foreign language. They all reported speaking with a standard accent in their L1s and (pre-)intermediate level in Polish which was in accordance with the level of the course they attended at AMU and was also confirmed by the language instructors. None of the speakers reported any speech or hearing disorders.

2.4 Procedures

The recordings were carried out in a professional recording studio (sampling rate: 44.1 kHz, 16-bit quantization). The Polish L1 speakers recorded consecutively the story, sentences, poems, mini-dialogues and spontaneous speech. The first three parts were first produced at the speaker's normal rate (norm), then at a fast rate (fast) and finally at the fastest tempo (vfast). After recording spontaneous speech, the texts were produced consecutively at a normal (norm2), slow (slow) and the slowest (vslow) speaking rate. As a result, each session consisted of altogether 32 recordings, each representing one speaking style at one speaking rate, e.g. recording 1: story/norm, recording 2: story/fast, recording 16: dialogue/norm.

The non-native speakers recorded first their L1 data: story "The North Wind and the Sun" and three sets of sentences differing in their phonotactic complexity, and then Polish data: story, sentences and mini-dialogues, only at the normal tempo. All the recordings were supervised by a phonetician. In case of disfluencies or too little variation in speaking rate the recording was repeated.

3. Speech database

3.1 Architecture and data management

For the purpose of annotation management, we applied the client-server methodology with Microsoft SQL Server. First, a relational database was designed, which made it possible to control the progress of annotation both locally, at the laboratory, and from remote locations. The implemented database model reflects the Polish rhythmic database contents structure and includes tables both for storing annotations themselves, and metadata (e.g., speaker's personal information, languages spoken, recording session date, access permissions, etc.). The client application (created in Visual Studio .NET WinForms C#) can be installed at any desired number of personal computers all connecting to the same central database. The application was integrated with *Annotation Pro* (Klessa et al., 2013, Klessa, 2016) and automatically opens files for annotation. The annotations are stored as database entries, but can also be easily exported as collections of self-standing annotation files (.ANT or .ANTx) or converted to any of *Annotation Pro* export formats (TextGrid, txt, csv).

3.2 Labeling procedure

The labeling was carried out in *Annotation Pro*. First, each recording session in the Polish L1 and Polish L2 subcorpora was divided into inter-pausal time groups and transcribed orthographically by four trained labelers. Then, for each time group we generated automatic transcription and alignment at the phoneme/syllable/word level using tools described in Demenko et al. (2003) and Szymański & Grocholewski (2005) which are integrated with *Annotation Pro*. The results were corrected manually by the four labelers and a final control procedure by an expert phonetician was carried out. The Polish L2 data was also annotated with mispronunciations as in Wagner & Cylwik (2010). The prosodic labeling (so far limited to Polish L1 data) consisted in marking minor and major prominences (non-prominent syllables did not receive any label) and boundaries of intonational and intermediate phrases, using perceptual and linguistic cues (cf. Breen et al., 2012; Streefkerk, 2002).

4. Ongoing work

4.1 Evaluation of the corpus

The total duration of the Polish L1 subcorpus is five hours and the mean amount of speech per speaker is 16 minutes. The Polish L2 subcorpus contains 13.3 minutes of speech. In the nearest future it is planned to record two more male speakers for the Polish L1 subcorpus in order to achieve

better balance in terms of speakers' gender, and also additional Polish L2 speakers to achieve the target of five speakers per language group. An overview of the two subcorpora is given in Table 1 and Table 2.

speaking style (L1 Polish)	min. of speech	syllables	words	inter-pausal groups
story	67	17943	9397	1583
sentences	112	33856	13455	2307
poems	70	17703	10239	1971
mini-dial.	13	5098	2571	498
spontaneous	39	12393	6341	1435
total:	301	86993	42023	7794
mean/speaker	16	4579	2212	410

Table 1: Summary statistics of the Polish L1 subcorpus.

subcorpus	min. of speech	sylla-bles	words	inter-pausal time groups
Polish L2 / German L1	3.2	2167	1013	208
Polish L2 / Spanish L1	5.5	2777	1314	400
Polish L2 / Korean L1	4.5	2817	1314	368
total:	13.3	7761	3641	976

Table 2: Summary statistics of the Polish L2 subcorpus.

A comparison between intended (ISR) and laboratory measured speaking rate (LSR) showed that speakers succeeded in differentiating the tempo of their reading: the number of syllables per second (excluding pauses) increases from normal to very fast (vfast) ISR and decreases from normal to very slow (vslow) ISR (Figure 1). Some differences in syllable rates between *norm* and *norm2* (for explanation see Sec. 2.4) can also be observed and they present an interesting topic for future study.

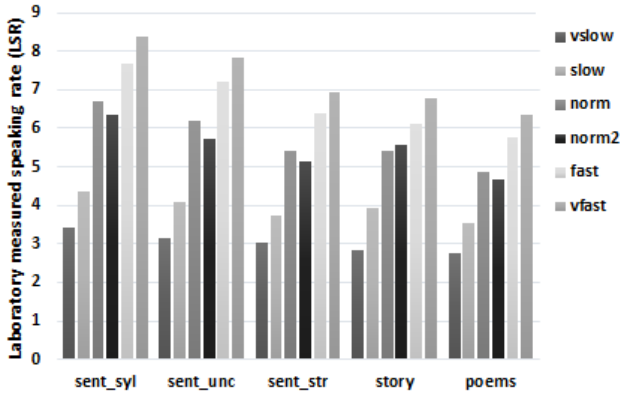

Figure 1: A comparison between intended and laboratory measured speaking rate in different subcorpora.

Generally, sentences were produced faster than story and poems. LSR was affected by syllable complexity: "syllable-timed" sentences were characterized by higher speaking rate than "uncontrolled" and "stressed-timed" sentences. LSR also differed among the speakers, with the slowest and fastest speakers achieving an overall rate of 3.29 and 5.34 syll./sec. respectively.

4.2 Automatized Procedures

In order to enhance data processing and analysis, two new *Annotation Pro* plugins (C# scripts) have been created: Category PVI plugin enabling calculation of nPVI and rPVI metrics (Grabe & Low, 2002) and Duration Intervals providing values of Varcos (Dellwo, 2006), ΔC, ΔV, and %V (Ramus et al., 1999). The plugins require input in the form of consonant-vowel label sequences, therefore we have developed another script which automatically converts phonemic transcriptions to CV sequences. The labeling on the phoneme layer is exported as one CSV file for all files using *Annotation Pro* export function. Among other things, the CSV file contains information about the source annotation filename in the .ANTx format, the phoneme label and start and duration times of the segment. This information is used to create V and C annotation files for each recording: within a given time group vowel labels and consonant labels are converted into "V" and "C", respectively. The conversion of /j/ and /w/ glides depends on the adjacent phonemes. Each V and C files have the structure: <V or C label; start time in seconds; end time in seconds>. Additionally, consecutive V or C segments are merged into one V or C interval. As a result a string of CVCVCV... intervals with time-stamps is created for each recording, and can be used as input for the Category PVI and Duration Intervals plugins.

The above solutions are largely language independent and together with *Annotation Pro + TGA* plugin developed earlier for the same software environment (Gibbon, 2013; Klessa & Gibbon, 2014), they provide a powerful tool for automatic extraction of timing and rhythm information.

Figure 2 presents *Annotation Pro* window displaying a fragment of a fully annotated utterance for which TGA parameters and rhythm metrics were calculated.

4.3 First experimental results

In a recent study (Wagner 2016) based on the whole Polish L1 subcorpus an attempt has been made to determine stability of timing patterns in Polish and to provide some insights into its rhythmic classification. The results showed that such aspects of Polish rhythm as durational variability and tendency towards isochrony expressed by PVIs are affected by speaking style, rate, poetic meter and phonotactic structure. It was observed that nPVI-V is more stable than rPVI-C and captures those timing properties which account for the perception of speech rhythm (e.g. the distinction between iambs and trochees). The results of a comparative analysis between Polish and prototypical stress- and syllable-timed languages, using data from Arvaniti (2012) and Dellwo & Wagner (2003), showed that: a) Polish has idiosyncratic patterns of timing variability that distinguish it from both stress-timed German/English

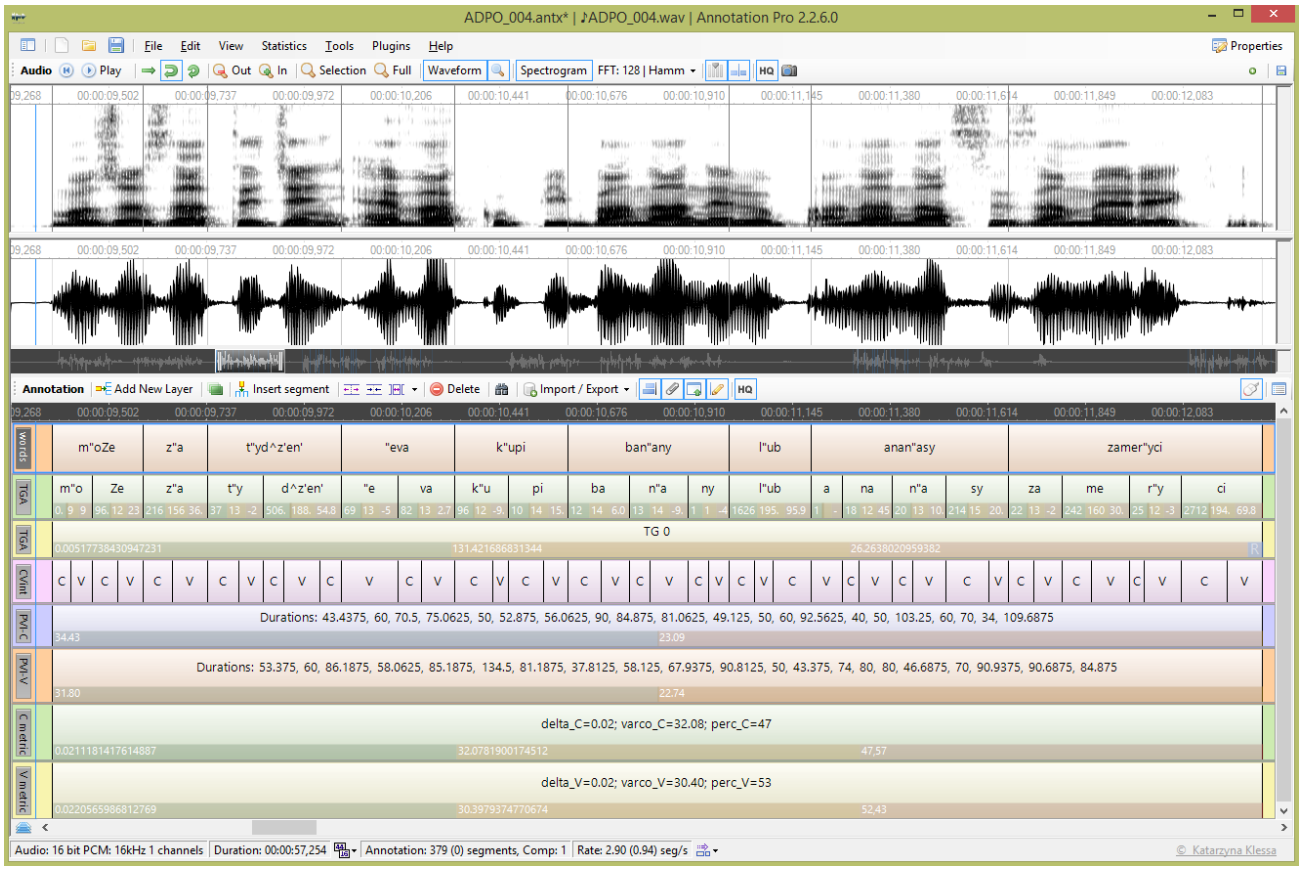

Figure 2: Annotation Pro window showing annotations and measurements for an utterance from the sent_syl subcorpus (some layers are hidden). From top to bottom: alignment and transcription at the word/syllable level, TGA results, segmentation into consonantal (C) and vocalic (V) intervals, PVIs and duration measurements for C and V intervals, deltas, Varcos and percentages of the duration of C and V intervals.

and syllable-timed Spanish/Italian, b) its position in the rhythm space is distant from the areas associated with both stress- and syllable-timing, c) it differs significantly from the other languages in terms of the patterns of tempo differentiation.

In another study (Wagner & Klessa, 2015) we investigated the relationship between tempo and speaking style on the one hand and timing properties on the other. For this purpose we performed *Time Group Analysis* (TGA; Gibbon, 2013; Klessa & Gibbon, 2014) and analyzed regression slopes and intercepts calculated for 1265 inter-pausal groups. The results showed that the intercept values were negatively correlated with tempo and the patterns were similar across speaking styles. The highest variability was found in mean slope values, indicating that the deceleration and acceleration patterns (expressed regression slope) vary more among the speaking styles than intercepts or mean values of speaking rate.

5. Conclusion and future work

The *Polish rhythmic database* is the first collection of spoken Polish designed for a systematic study of a broad range of phenomena related to speech timing, rhythm and prosody in general and it is integrated with a collection of software tools and methods designed to facilitate the database annotation and analysis with respect to various timing and rhythm measures. Currently, the collected speech data (5 h 14 min.) comprises recordings of 19 native and 12 non-native speakers of Polish and represents five different speaking styles and five different tempi. In the near future the results of prosodic labeling of the Polish L1 subcorpus will be analyzed in terms of inter-labeler agreement and intra-labeler consistency and different approaches to deriving the final markers of prominence level and boundary strength, (e.g., as in Streefkerk (2002) or Cole et al. (2010)) will be compared. The same procedures will be applied to Polish L2 subcorpus once its prosodic labeling is completed. We also intend to label the Korean/German/Spanish L1 subcorpora and to engage the speakers of these languages to provide prosodic annotations of selected data from the Polish L1 subcorpus. Apart from prominence annotation we intend to provide a higher-level description of intonation like e.g., ToBI (Beckman & Ayers, 1997; Silverman et al., 1992) for at least a part of the corpus and to automatically extract acoustic features of accents and phrase boundaries (in the domain of f0, intensity and time). First of all, timing, tempo and intonation are integrated in the speech signal and secondly, there has been a growing body of evidence that they interact in speech perception (e.g., Boswell & Arvaniti, 2010, Arvaniti & Ross, 2012). This casts doubt on the idea that rhythm is based solely on timing, as assumed by *rhythm class hypothesis*, and indicates the need of integrating all these prosodic factors in the description and analysis of rhythm.

Among research topics that we intend to investigate are: stability of rhythm metrics other than PVIs, rhythmic classification of Polish, phonotactic and phrasal properties of speech rhythm in Polish as L1 and L2, and cross-linguistic perception and production of prominence and phrasing. The final goal is to provide a comprehensive description of timing patterns and speech rhythm in Polish. Apart from that the results of the research based on *Polish rhythmic database* may also have an impact on the general understanding of rhythm and on the methodology of future research in this field.

6. Acknowledgements

This research has been carried out in the scope of the project "Rhythmic structure of utterances in the Polish language: A corpus analysis" supported by National Science Centre (NCN) grant no. 2013/11/D/HS2/04486.
The authors would like to thank the two anonymous reviewers for their suggestions and comments.

7. Bibliographical references

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics,* 40(3), pp. 351--373.

Arvaniti, A., Ross, T. (2012). Rhythm classes and speech perception. In O. Niebuhr (Ed.), *Understanding prosody: The role of context, function and communication*, pp. 75--92.

Beckman, M. E., Ayers, G. E. (1997). Guidelines for ToBI labeling, version 3: Ohio State University.

Boswell, T., Arvaniti, A. (2010). The role of tempo and pitch in rhythm class discrimination. *Journal of the Acoustical Society of America,* 128(4), pp. 2478--2478.

Bradlow, A. R., Baker, R. E., Choi, A., Kim, M., Van Engen, K. J. (2007). The Wildcat corpus of native- and foreign-accented English. *Journal of the Acoustical Society of America,* 121(5), pp. 3072--3072.

Breen, M., Dilley, L. C., Kraemer, J., Gibson, E. (2012). Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory,* 8(2), pp. 277--312.

Cole, J., Mo, Y., Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1(2), pp. 425--452.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for ΔC. In *Language and Language-Processing: Proceedings of the 38th Linguistics Colloquium,* Frankfurt am Main, Germany: Peter Lang, pp. 231--241.

Dellwo, V., Steiner, I., Aschenberner, B., Dankovicova, J., Wagner, P. (2004). BonnTempo-Corpus & BonnTempo-Tools: A database for the study of speech rhythm and rate. In *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, Sunjin Printing Company.

Demenko, G. (1999). *Analysis of Polish suprasegmentals for the needs of speech technology.* Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Dellwo, V., Wagner, P. (2003). Relations between language rhythm and speech rate. In *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, Spain, pp. 471--474.

Demenko, G., Wypych, M., Baranowska, E. (2003). Implementation of grapheme to phoneme rule and extended SAMPA alphabet in Polish text-to-speech synthesis. *Speech and Language Technology*, 7, pp. 79--95.

Demenko, G., Grocholewski, S., Klessa, K., Wagner, A., Ogórkiewicz, J., Lange, M., Śledziński, D., Cylwik, N. (2008). Jurisdic – Polish Speech Database for taking dictation of legal texts. In *Proceedings of Language Resources and Evaluation Conference (LREC),* Marrakech, Morocco, 2008.

Gibbon, D. (2013). TGA: a web tool for Time Group Analysis. In B. Brigitte, D. Hirst (Eds.), *Proceedings of TRASP (Tools and Resources for the Analysis of Speech Prosody).* Aix-en-Provence, 30 August, 2013. ISBN: 978-2-7466-6443-2.

Grabe, E., Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology,* 7, pp. 515--546.

Jarmołowicz, E., Karpinski, M., Malisz, Z., & Szczyszek, M. (2007). Gesture, prosody and lexicon in task-oriented dialogues: multimedia corpus recording and labeling. *Verbal and Nonverbal Communication Behaviours,* Berlin/Heidelberg, Germany: Springer, pp. 99--110.

Klessa, K., Gibbon, D. (2014). Annotation Pro + TGA: automation of speech timing analysis. In *Proceedings of Language Resources and Evaluation Conference (LREC),* Reykjavik, Iceland. ISBN 978-2-9517408-8-4.

Klessa, K., Wagner, A., Oleśkowicz-Popiel, M., Karpiński, M. (2013). "Paralingua" – a new speech corpus for the studies of paralinguistic features. *Procedia – Social and Behavioral Science*, 95, pp. 48--58.

Klessa, K., Karpiński, M., Wagner, A. (2013). Annotation Pro – a new software tool for annotation of linguistic and paralinguistic features. In B. Brigitte, D. Hirst (Eds.), *Proceedings of TRASP (Tools and Resources for the Analysis of Speech Prosody).* Aix-en-Provence, 30 August, 2013. ISBN: 978-2-7466-6443-2.

Kulawik., A. (1997). *Poetics: An introduction into the theory of literary work.* Krakow, Poland: Antykwa.

Mairano, P., Romano, A. (2011). Rhythm metrics for 21 languages. In *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, China, pp. 1318--1321.

Malisz, Z. (2011). Tempo differentiated analyses of timing in Polish. *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, China, pp. 1322--1225.

Malisz, Z., Wagner, P. (2012). Acoustic-phonetic realization of Polish syllable prominence: a corpus study. *Speech and Language Technology*, 14/15, pp. 105--114.

Prieto, P., Vanrell, M. D. M., Astruc, L., Payne, E., Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication*, 54(6), pp. 681--702.

Ramus, F., Nespor, M., Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3), pp. 1--28.

Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P.J., Pierrehumbert J., Hirschberg J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of the International Conference on*

Spoken Language Processing 92', pp. 867--870.
Streefkerk, B. M. (2002). *Prominence. Acoustic and lexical/syntactic correlates.* Utrecht, The Netherlands: LOT.
Szymański M., Grocholewski S. (2005). Transcription-based automatic segmentation of speech. In *Proceedings of 2nd Language & Technology Conference,* Poznan, Poland, pp. 11--15.
Śledziński, D. (2013). Syllabification of a text corpus – analysis of Polish consonant clusters. *Kwartalnik Językoznawczy*, 3, pp. 48--100.
Wagner, A. (2008). Comprehensive model of intonation for application in speech synthesis. Unpublished PhD dissertation. Adam Mickiewicz University, Poznan, Poland.
Wagner, A., Cylwik, N. (2010) Creation of the linguistic content for the pronunciation tutoring system AzAR 3.0. *Speech and Language Technology*, 12/13, pp. 145--156.
Wagner, A., Klessa, K. (2015). The influence of tempo and speaking style on timing patterns in Polish. In A. Botinis (ed.), *Proceedings of the 6th ISEL Conference on Experimental Linguistics,* Athens, Greece, June 26-27, 2015, pp. 86--89.
Wagner, A. (2016). Do languages have stable contrastive rhythmic properties? Experimental evidence from Polish. *Philological studies*, University of Warsaw (in print).
Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O. & Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *Journal of the Acoustical Society of America,* 127(3), pp. 1559--1569.

8. Language Resource References

Klessa, K. (2016). *Annotation Pro* [Software tool]. Version 2.3.1.5. Retrieved from: http://annotationpro.org/ on 2016-02-17.