Automatic Syllabification of Polish

Brigitte Bigi*, Katarzyna Klessa†

*Laboratoire Parole et Langage, CNRS, Aix-Marseille Université
5 avenue Pasteur, 13100 Aix-en-Provence, France
brigitte.bigi@lpl-aix.fr

†Institute of Linguistics, Adam Mickiewicz University in Poznań
Instytut Językoznawstwa UAM, al. Niepodległości 4, pok. 218 B, 61-874 Poznań
klessa@amu.edu.pl

Abstract

This paper presents an approach to automatic detection of syllable boundaries for Polish speech based on a phonetized text input. First, we discuss selected issues of syllable structure in Polish with a special focus on the needs of rule-based automatic insertion of syllable boundaries. We describe and verify an existing rule-set for Polish, which is subsequently used as an input information for automatic syllabification with SPPAS, a freely available multiplatform software tool. Then, the applied syllabification methodology is described and illustrated with examples obtained with a Polish speech corpus. Finally, the paper provides information about the syllabification module for Polish that has been implemented as one of the latest extensions of SPPAS.

Keywords: automatic syllabification, syllabification rules, speech annotation, free software tools

1. Introduction

Despite controversies regarding the definition of syllable, e.g., (Laver, 1994), (Roach, 1990), and the discussion over its role in speech technology (Kishore et al., 2003), (Hu et al., 1996), the syllable is credited as a linguistic unit conditioning both segmental (e.g., consonant or vowel lengthening) and prosodic phonology (e.g., tune-text association, rhythmical alternations) (Campbell, 1992), (Roach, 1982). Automatic annotation on the syllable level represents a valuable tool for quantitative analyses of large speech data sets.

While the phonological structure of the syllable is similar across different languages, phonological and phonotactic rules of syllabification are language-specific. Automatic approaches to syllable detection have thus to incorporate such constraints to precisely locate syllable boundaries. The question then arises of how to obtain an acceptable syllabification for a particular language and for a specific corpus (a list of words, a written text or transcripts of an oral corpus of more or less casual speech).

Although a number of automatic syllabifiers are currently available online as freeware for some languages (e.g., for French, English or German), the availability of such resources for Polish is still limited. The present work aims at filling this gap by proposing a freely available automatic segmentation tool implemented as a new module in SPPAS (Bigi, 2012).

The main aspect of the SPPAS automatic syllabification reported in this paper is as follows:

- to propose *a generic and easy-to-use tool* to identify syllabic segments from phonemes;
- to propose *a generic algorithm*, then a set of rules for the particular context of Polish spontaneous speech.

In this context, "generic" means that the phone set, the classes and the rules are easily changeable; and "easy-to-use" means that the system can be used by any user.

Section 2 of the paper introduces selected Polish-specific issues related to the syllable that are potentially relevant for the task of constructing rules for automatic syllabification. Section 3 describes the methodology and implementation of the new, freely available syllabification module in SPPAS, while Section 4 summarizes the study and outlines the necessary future work.

2. Establishing a rule-set for automatic syllabification of Polish

Polish is known to have a rich consonant inventory and to allow complex consonant clusters to exist within utterance structures both word-initially, word-internally, and word-finally. Many 4-5-element clusters can be easily found within words occurring in typical Polish texts (Śledziński, 2013), and when cross-word sequences are considered, the clusters can be even longer, reaching e.g., eight consonantal elements such as in the phrase: *sierść z pstrym wzorem* (Eng. 'fur of motley pattern') where the SAMPA (Wells et al., 1997) transcription for the underlined fragment would be /r s' ts' s p s t r/. Obviously, in spontaneous speech various types of reductions or elisions might be expected, resulting in simplification of such clusters, but anyway, their proper pronunciation would not pose serious problems for Polish native speakers.

Finding the optimal location for syllable boundaries appears particularly challenging in case of languages characterized by complex consonant clusters. The departing points for syllabification are usually associated with two principles, widely discussed in the subject literature, i.e. the Sonority Sequencing Principle (sonority rise within the syllable onset and its fall within the coda, the nucleus being the most sonorous element) (Selkirk, 1984) and the Maximal Onset Principle (assigning intervocalic consonants to the syllable onset unless it contradicts the sonority principle) (Kahn, 1976). Unconditioned application of both of

the principles might become problematic in practice, e.g., because of the same sonority level in neighbouring consonants or simply due to the number of consonants in a sequence. Various approaches have been reported so far for Polish corpus-based research. A detailed report of a corpus-based analysis of Polish syllables and consonant clusters can be found in (Śledziński, 2013) who postulates, among others, postulating among others, purposeful violation of one or both of the principles due to morphological constraints. Other researchers prefer to primarily use the sonority principle, together with a number of additional language-specific constraints (Malisz and Wagner, 2012).

In the present study, we make use of a list including above 1800 phoneme sequences with boundary placement information (henceforth referred to as the syllable pattern list). The list was constructed manually based on the analysis of subject literature (e.g., (Rubach, 1990), (Szpyra-Kozłowska, 1998), (Śledziński, 2007)) and subsequently refined and extended based on experiences gained during annotation of several speech databases for Polish such as (Klessa et al., 2009) or (Klessa et al., 2013b). For most of the patterns, the boundary positions are defined according to the Maximal Onset Principle on condition that they do not violate the sonority hierarchy (following the approach used e.g., by Klessa and Śledziński, 2006).

The current version of the list has been found to be sufficient to successfully syllabify a number of Polish transcripts derived from annotations of spoken language corpora composed both of read narratives and quasi-spontaneous dialogues or monologues. A single entry in the list consists of:

- a sequence of transcription labels for consonants located between two vowels or between a vowel and a pause; we use SAMPA (Wells et al., 1997) as the phonetic alphabet in an extended version (Demenko et al., 2010);
- two V labels representing vowels (without distinguishing between particular vowel labels);
- a boundary separator.

The pattern list is used as one of the components in Annotation Pro software tool (Klessa et al., 2013a). Syllable boundaries are inserted into a previously phonetized input string (the phonetization needs to be done either manually or with an external tool). In case of a missing syllable pattern, the software displays a report dialogue window and it is possible for the user to add the missing pattern to the list. As it was mentioned above, the current version of the list suffices for syllabification of standard texts but the possibility to add non-default patterns might be still helpful for users automatically processing non-standard texts, e.g., when constructing test signals for the needs of applications for measuring and fitting hearing aids where many atypical sequences may occur (a recent work for Polish speech test data has been reported by Habasińska, 2015).

The syllabification method used in Annotation Pro is based on a very straightforward boundary insertion algorithm. The algorithm performs a pattern matching procedure using the whole list of syllable patterns as a reference to match with the sequences found in the transcribed text. No generalisation or grouping of the rules has been implemented. The pattern-matching process is designed so that it always begins with the longest pattern in the list and if no match is found, an attempt is made to match a shorter one. The constraints for the syllable boundary insertion are as follows:

1. a syllable may include only one vowel label;
2. an acoustic pause is a syllable boundary;
3. a word boundary reflected by a space in orthography is a syllable boundary with the exception of word sequences with the Polish prepositions *w* and *z* (Eng. 'in' and 'from/with') which are orthographically spelt as separate units, e.g., 'w domu' (Eng. 'at home') but do not constitute separate syllables. The resulting syllabification in such case would be /vdo/ /mu/.

A more detailed discussion of the above syllabification algorithm is beyond the scope of the present paper. Here, we use the syllable pattern list as an initial input for developing a new, class-based set of rules for the Polish SPPAS module.

3. SPPAS syllabification: Method description

In this section, we report on the adaptation of a rule-based system for automatic syllabification of phoneme strings of the size greater than a graphic word. The system was initially developed for French (Bigi et al., 2010) and then adapted for Italian (Bigi and Petrone, 2014). It is here adapted to Polish.

The system proposed in this paper is included in SPPAS (Bigi, 2012), a tool distributed under the terms of the GNU Public License[1]. It is implemented using the programming language Python 2.7. Among other functions, SPPAS offers an automatic speech segmentation at the phone and token levels for French, English, Spanish, Italian, Catalan, Polish, Mandarin Chinese, Cantonese, Taiwanese and Japanese.

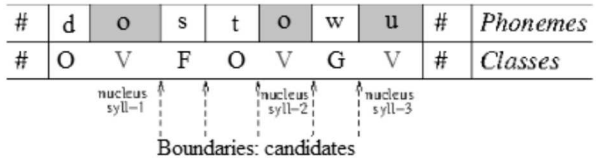

Figure 1: Main principles ("#" denotes a pause)

The problem we deal with is the automatic syllabification of a phoneme sequence. The proposed phoneme-to-syllable segmentation system is based on 2 main principles (Figure 1):

1. a syllable contains a vowel, and only one;
2. a pause is a syllable boundary.

[1]See: `http://www.gnu.org/licenses/gpl-3.0.en.html` for details

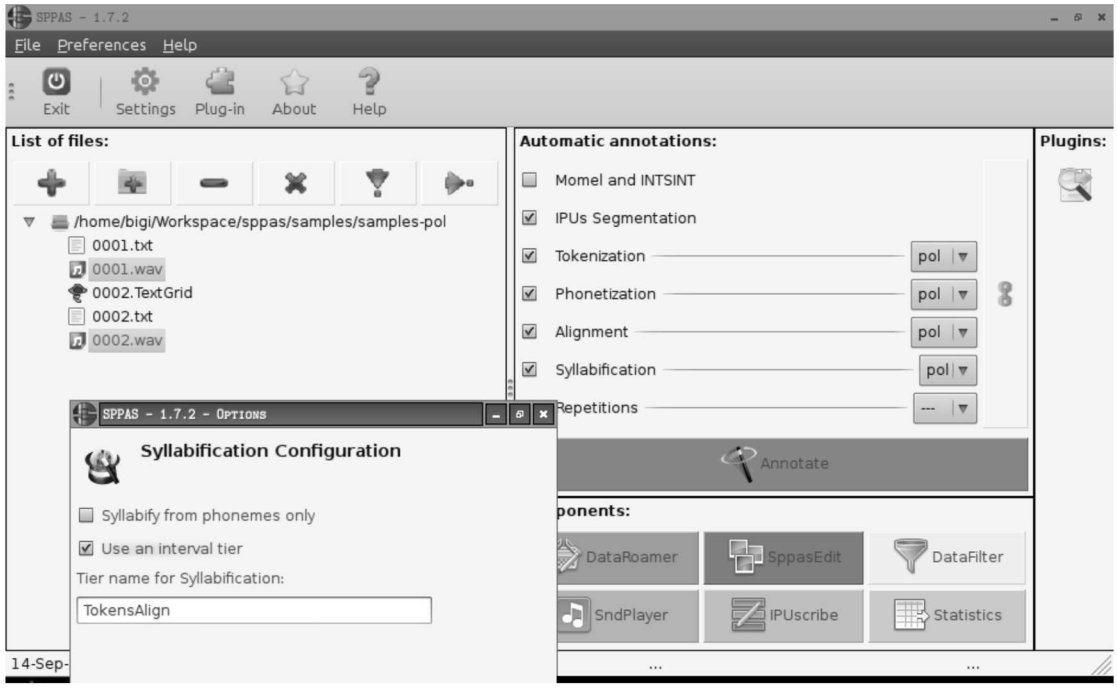

Figure 2: SPPAS: Graphical User Interface with syllabification options

These two principles focus the problem on the task of finding a syllabic boundary between two vowels, in each Inter-Pausal Unit (IPU).

As in the initial system for French, we group phonemes into classes and establish language-specific rules dealing with these classes. The identification of relevant classes is then very important.

The following classes were defined for Polish based on the contents of the syllable patterns list described in the previous section of this paper:

V - Vowels: a e e˜ i o o˜ u y
G - Glides: j w
L - Liquids: l r
O - Occlusives: p t k b d g Q c
F - Fricatives: dz dZ f v x tS ts ts‘ dz‘ z z‘ Z
S - Fricatives: s‘ s S
N - Nasals: n n‘ m N J

Uppercase bold-letters indicate the abbreviations used for classes throughout this paper. The letter **C** is also used to mention one of G, L, O, N, F, S.

The system firstly checks if the observed sequence of classes corresponds to an exception. If not, the general rules are applied (see Table 1).

For VCCV, the exception rules are:

- VOCV is segmented as V.OCV
- VFCV is segmented as V.FCV
- VSCV is segmented asV.SCV, except for C=S
- VCGV is segmented as V.CGV

For VCCCV, the exception rules are:

- VOOCV is segmented asV.OOCV
- VFFGV is segmented as V.FFGV

	Observed sequence	Segmentation rule
1	VV	V.V
2	VCV	V.CV
3	VCCV	VC.CV
4	VCCCV	VC.CCV
5	VCCCCV	VC.CCCV
6	VCCCCCV	VCC.CCCV
7	VCCCCCCV	VCC.CCCCV

Table 1: General Rules (V and C are phonological vowels/consonant respectively)

SPPAS also offers the possibility to fix segmentation in a set of specific rules that deal with phoneme sequences, to which our general or exception rules do not apply. that in some cases, rules are not relevant. These specific rules are not used for Italian and just a few are fixed for French. However, due to the large number of consonant clusters in Polish, these specific rules on phonemes sequences are widely used:

- 89 specific rules of phoneme sequences in VCCV;
- 135 specific rules of phoneme sequences in VCCCV;
- 86 specific rules of phoneme sequences in VCCCCV;
- 20 specific rules of phoneme sequences in VCCCCCV;

Then, it represents a total of 330 specific rules that correspond to 18% of the pattern list defined in the previous section. The general rules and the exception rules previously defined cover all the other cases.

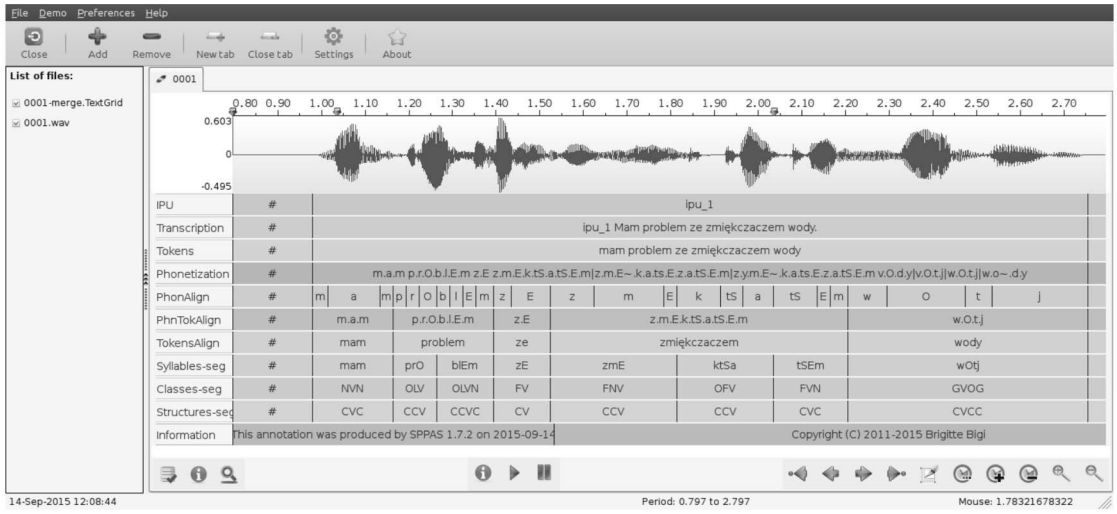

Figure 3: SPPAS output example on a read sentence in Polish.

Finally, in the system described in (Bigi et al., 2010), the syllabification is performed between 2 pauses (as defined in the main principles). From this system, we added the possibility to perform the syllabification between any kind of boundaries. In such case, a "reference tier" is given by the user to the system. Table 2 shows an example when the time-aligned tokens are used as a reference tier.

segment type	sentence	phonemes	syllables
sentence	do domu	/dodomu/	do.do.mu
	do stołu	/dostowu/	dos.to.lu
token	do.domu	/do/ /domu/	do.do.mu
	do.stołu	/do/ /stowu/	do.sto.wu

Table 2: Syllabification into segments, without changing the rules.

Of course, the reference tier can contain any type of annotation (we used word tokens in the example, but prosodic contours, syntactic segments, etc. can be used if their annotation is available). The use of tokens is particularly relevant for Polish. Figure 3 illustrates the result SPPAS can produce.

In SPPAS, a simple text file that the user can change as needed contains the phoneset and the rules for the syllabification process.

4. Conclusion & future work

This paper describes an implementation of an automatic syllabification system for Polish. The implemented functionality is freely available as part of SPPAS software (Bigi, 2012). The first tests of its performance have been conducted using the conversational speech recordings from the Paralingua corpus (Klessa et al., 2013b), however a more systematic evaluation is a necessary further step. An interesting future task would also be a comparison of the output of automatic syllabification using the two different methodological approaches: the generalised class approach implemented in SPPAS vs. the straightforward pattern-matching method used in Annotation Pro (Klessa et al., 2013a).

An additional outcome of the reported study was an improvement of the Polish syllable pattern list mostly related to the removal of unnecessary patterns (duplicates or spelling errors).

In the course of the present work we have also developed import/export modules to enable data transfer between the native formats of Annotation Pro and SPPAS. Consequently, it is now possible to use SPPAS syllabification output also within Annotation Pro, not only for Polish but also for all other languages available in SPPAS. Moreover, thanks to the interoperability between the two tools, another types of analyses become supported, e.g., combined studies based on linguistic information automatically generated with SPPAS (e.g., part-of-speech tagging or prosodic labelling) with perception test results or perception-based annotations obtained with Annotation Pro (see Figure 4 for an example view including syllabified data and the graphical representation of the feature space, useful for perception tests).

5. References

Bigi, B., 2012. SPPAS: a tool for the phonetic segmentations of Speech. In *The eighth international conference on Language Resources and Evaluation, ISBN 978-2-9517408-7-7*. Istanbul, Turkey.

Bigi, B., C. Meunier, I. Nesterenko, and R. Bertrand, 2010. Automatic detection of syllable boundaries in spontaneous speech. In *Language Resource and Evaluation Conference*. La Valetta (Malta).

Bigi, B. and C. Petrone, 2014. A generic tool for the automatic syllabification of italian. In *Proceedings of the First Italian Conference on Computational Linguistics and the Fourth International Workshop EVALITA 2014*.

Campbell, W Nick, 1992. Syllable-based segmental duration. *Talking machines: Theories, models, and designs*:211–224.

Demenko, G., K. Klessa, M. Szymański, S. Breuer, and

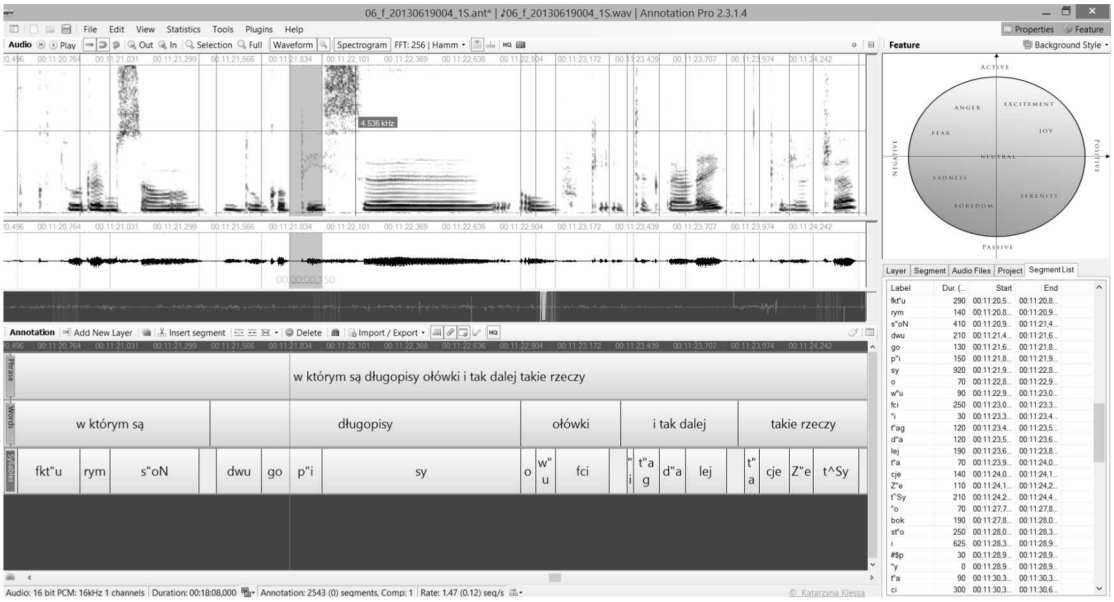

Figure 4: Annotation Pro interface: example syllabification and graphical representation of the feature space.

W. & Hess, 2010. Polish unit selection speech synthesis with BOSS: extensions and speech corpora. *International Journal of Speech Technology*, 13(2):85–99.

Habasińska, D., 2015. Stworzenie polskiego testowego sygnału mowopodobnego do wykorzystania w miernictwie i dopasowaniu aparatów słuchowych (En. The development of the Polish speech test signal for measuring and fitting hearing aids). MA Thesis, The Institute of Acoustics, Adam Mickiewicz University in Poznań.

Hu, Z., J. Schalkwyk, E. Barnard, and R. Cole, 1996. Speech recognition using syllable-like units. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 2. IEEE.

Kahn, D., 1976. *Syllable-based generalizations in English phonology*, volume 156. Indiana University Linguistics Club Bloomington.

Kishore, S., Prahallad, and Alan W. Black, 2003. Unit size in unit selection speech synthesis. In *INTERSPEECH*.

Klessa, K, M Karpiński, O Bałdys, and G Demenko, 2009. Speechlabs ASR. Ppolish lexical database for speech technology: Design and architecture. *Speech and Language Technology*, 12(13):191–207.

Klessa, K., M. Karpiński, and A. Wagner, 2013a. Annotation pro-a new software tool for annotation of linguistic and paralinguistic features. In *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence*.

Klessa, K. and D. Śledziński, 2006. A study of chosen temporal relations within syllable structure in Polish. *Speech and Language Technology*, 9.

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

Laver, J., 1994. *Principles of phonetics*. Cambridge University Press.

Malisz, Z. and P. Wagner, 2012. Acoustic-phonetic realisation of Polish syllable prominence: a corpus study. *Speech and Language Technology. Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem.*, 14.

Roach, P., 1982. On the distinction between 'stress-timed'and 'syllable-timed'languages. *Linguistic controversies*:73–79.

Roach, P., 1990. *English Phonetics and Phonology: a practical course*. Cambridge: Cambridge University Press.

Rubach, & Booij G., J., 1990. Syllable structure assignment in Polish. *Phonology*, 7(01):121–158.

Selkirk, E. O., 1984. On the major class features and syllable theory:107–136.

Śledziński, D., 2007. *Fonetyczno-akustyczna analiza struktury sylaby w języku polskim na potrzeby technologii mowy*. Ph.D. thesis, Adam Mickiewicz University, Poznań, Poland.

Śledziński, D., 2013. Podział korpusu tekstów na sylaby – analiza polskich grup spółgłoskowych. *Kwartalnik Językoznawczy*, 2013(3):48–99.

Szpyra-Kozłowska, J., 1998. The sonority scale and phonetic syllabification in Polish. *Biuletyn Polskiego Towarzystwa Językoznawczego, Bulletin de la Société Polonaise de Linguistique*, 54:63–82.

Wells, John C et al., 1997. SAMPA computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, 4.