Increasing the Accessibility of Time-Aligned Speech Corpora with *Spokes Mix*

Piotr Pęzik
University of Lodz
Pomorska 171/173, 90-236 Lodz, Poland
piotr.pezik@uni.lodz.pl

Abstract

Spokes Mix is an online service providing access to a number of spoken corpora of Polish, including three newly released time-aligned collections of manually transcribed spoken-conversational data. The purpose of this service is two-fold. Firstly, it functions as a programmatic interface to a number of unique collections of conversational Polish and potentially also spoken corpora of other languages, exposing their full content with complete metadata and annotations. Equally important, however, is its second function of increasing the general accessibility of these resources for research on spoken and conversational language by providing a centralized, easy-to-use corpus query engine with a responsive web-based user interface.

Keywords: Speech corpora, corpus search engine, online language services

1. Introduction

High-quality corpora of spoken conversational language are expensive to acquire. Worse still, even when they are collected, annotated and made publicly available, their potential as sources of primary data may remain largely unrealized, due to the lack of search and exploration tools which could deal with the identification of speech-specific linguistic and multimodal phenomena. Apart from their research value, spoken language resources are also indispensable in the development of speech recognition systems and related language technologies. However, in order to sufficiently serve this purpose, they have to be standardized and interoperable enough to be used for both online and offline processing. Spokes Mix is an online service developed as part of the CLARIN-PL infrastructure to address both of these challenges for Polish. Firstly, the search engine of Spokes Mix makes it possible to explore not only the linguistically annotated transcriptions of spoken language, but also some characteristics of the underlying speech signal, such as duration and prosodic contours. These functionalities can be accessed through a web application, which also makes it relatively easy to customize, export and save search results by non-technical users. Secondly, the entire contents of the indexed corpora is available for programmatic search clients and it can exported on demand together with the time-aligned sound files. The present paper briefly describes both the newly and previously available corpora available through Spokes Mix. Next, some of the features of its search engine and API are described to illustrate its accessibility as a language resource service for both technical and casual users. Finally, some of the key planned improvements of Spokes Mix are also briefly outlined.

2. The Data

Spokes Mix provides access to a number of spoken corpora, including three entirely new collections of conversational Polish. To demonstrate and test the ability of Spokes Mix to simultaneously serve speech corpora in multiple languages, it also includes a collection of well-known speech corpora of English. All of these datasets are briefly specified below.

2.1. Annotation and Interoperability

The corpora currently indexed in Spokes Mix have been unified with respect to the varying degrees of bibliographic metadata and sociolinguistic annotation they originally contained. Every transcription has been automatically tokenized into word segments, part-of-speech tagged and time-aligned with the original recording. Spans of transcriptions are explicitly linked to recordings at the level of utterances, words and sound segments. The latter two levels of annotation are provided automatically using the forced-alignment tools developed by (Korżinek et al., 2017). In addition to the JSON (JavaScript Object Notation)-based proprietary data format used to serve search results and export the textual contents of the indexed corpora, popular formats of annotating speech samples are available for selected features of the search engine, such as time-aligned utterances matching corpus queries, which can be downloaded as TextGrid (Boersma, 2006) or .eaf files. Additionally, the entire corpora indexed in Spokes Mix are stored internally as ELAN-encoded files (Wittenburg et al., 2006) and EMU-databases (Cassidy and Harrington, 2001). As indicated below, we are planning to release these versions as self-contained, offline resources.

2.2. Casual-Spoken Data

The PELCRA Conversational corpus is the largest of the Polish collections available through Spokes Mix. It contains just over 218 hours (ca. 2.2 million word tokens) of unplanned, casual conversations recorded in vivo contexts. Released previously through the first version of Spokes (Pęzik, 2015), this corpus has found numerous applications in basic research on spoken Polish, e.g. (Guz, 2015), (Guz, 2017). Among the new features in this edition of the corpus is the word-token level time-alignment and annotation of intonation contours. The original CC-BY-NC license for this corpus extends to this new edition as well.

2.3. Parliamentary Proceedings

As mentioned above, Spokes Mix provides access to three new collections of spoken conversational Polish released under a CC-BY license. The first one of these is a corpus of 50 sampled recordings of parliamentary sessions and com-

mittee meetings, totalling over 12 hours (ca. 101 000 word tokens) of formal spoken Polish. The recordings were manually transcribed with a separate tier for each speaker and force-aligned word segment boundaries.

2.4. Focused Interviews

The second newly released collection of Polish spoken data contains 30 structured interviews (17 hours, ca. 200 000 word tokens) focused around the topic of selected emotions in personal experiences of the interviewed speakers. The interviews are dialogues recorded in stereo with a separate channel for each speaker. The recordings are aligned with the transcriptions both manually at the level of turns and automatically at the level of word tokens.

2.5. Semi-scripted Interviews

The last of the three new collections is still under preparation. Its currently available section contains recordings of 25 interviews with partly scripted questions (15 hours, ca. 165 000 word tokens) covering a variety of everyday topics. Similarly to the previous two collections, the corpus of semi-scripted interviews is released under a CC-BY license. The Polish corpora currently available in Spokes Mix are summarized in Table 1.

Corpus	Register	Hours	Words	License
Conversational	Casual	218	2.2 M	CC-BY--NC
Parliamentary	Formal	12	0.1 M	CC-BY
Focused Intvs.	Mixed	17	0.2 M	CC-BY
Open Intvs.	Mixed	15	0.16 M	CC-BY

Table 1: Summary of Polish speech corpora in Spokes Mix.

2.6. Spoken Corpora of English

The ability of Spokes Mix to serve multilingual speech corpora has been tested on a collection of spoken corpora of English. Currently, users can search, browse and export samples of the spoken subcorpus of the British National Corpus, which was relatively recently time-aligned and released by (Coleman et al., 2012). The service also features a large collection of English read speech in the form of the Librispeech corpus (Panayotov et al., 2015) and an experimental version of the TEDLIUM Corpus (Rousseau et al., 2014).

3. Corpus Search Engine

One of the main objectives of developing Spokes Mix was to provide a corpus search and exploration service which would make its underlying resources accessible for non-technical users. The current implementation of the search engine supports positional and simple lexico-grammatic queries which build upon the syntax described in (Pęzik, 2015). In addition to returning a specified number of contexts matching a query, the engine also calculates aggregated statistics called 'facets' computer on the entire result set and presents them as charts and other visualizations. Full results can be downloaded as Excel spreadsheets, while individual utterances matching the query can be downloaded or played in the web browser interactively using time-aligned word span highlights.

Selected tiers of automatic prosodic annotation can be integrated on demand with spans matching a corpus query. This feature is illustrated in Fig. 1, where the query for the phrase *you know* matches a total of 2720 turns in the indexed corpora. The duration of every matching span is calculated dynamically and presented in a separate column, which can be used as a sorting key. Additionally, pitch codes are presented as a separate piece of prosodic annotation for each matching span. The encoding of pitch and intensity contours is performed using the Momel technique for simplifying contours and the INTSINT scheme (Hirst et al., 2000), (Hirst, 2007) for encoding intonation patterns, with two global codes (T – top, M – middle and B – bottom) representing normalized pitch points and five relative ones (H – higher, U – up-stepped, S – same, D – down-stepped and L – lower).

In addition to presenting such dynamically mapped prosodic annotations for each concordance, it is possible to obtain a number of aggregated views of these data by simply clicking on the `Prosody` tab in the main results window. For example, a requested subset of spans matching the query *you know* can be analyzed to obtain a summary of the distribution of INTSINT codes for this phrase in the underlying corpora of English speech. An example of such a summary is shown in Table 2, which lists the ten most frequent INTSINT code combinations found in a sample of 1000 instances of *you know*. There are relatively few occurrences of this phrase which coincide with local pitch maxima (marked as T), which may be explained by the fact that the semantic bleaching of *you know* in its discourse-marking function results in prosodic and phonetic reduction. Table 3 shows some descriptive statistics calculated for a sample of 1000 duration values of *you know*, which are also displayed in the `Prosody` tab of the search results screen.

Needless to say, such hypotheses about possible correlation of prosodic features and discourse function should be verified by a closer inspection of the underlying data. The link between aggregated statistics and individual data samples is preserved in Table 2 and users can access all instances of a specific code with a single click on a given row. It is even possible to use different features or their combinations to create cluster-like visualisations of prosodically 'similar' instances of spans matching a corpus query. Fig. 2 shows a graph generated by Spokes Mix for combinations of INTSINT codes aligned with matching concordance spans. The red nodes in the graph represent INTSINT codes and they are linked with yellow nodes representing matching utterances. User may click on any of the red nodes to get an instant listing of the utterances containing the corresponding matching span.

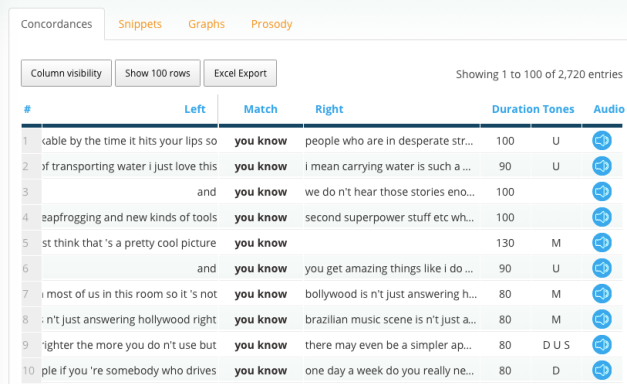

Figure 1: Dynamic mapping of prosodic annotation in corpus search results.

IntSint Code	Frequency
M	130
U	120
H	78
D	78
L	69
S	69
B	67
T	38
M_S	21
U_D	21

Table 2: Top 10 InstSint Codes in a sample of 1000 instances of the phrase *you know* in Spokes Mix.

Statistic	Value
Sample size	1000
Missing values	4
Minimum	59
1st Quartile	80
Median	100
3rd Quartile	114.5
Max	510
Mean	103.04
Variance	1840.70
St. dev	42.90

Table 3: A summary of duration values measured for 1000 occurrences of the phrase *you know* in Spokes Mix.

4. Programmatic Access

The REST (Representational state transfer)-based public API of Spokes Mix can be used not only to selectively search the underlying resources, but also to export their full transcriptions with metadata and sound files. An example response to a concordance query is illustrated in the listings shown below. The same API is used by the Spokes Mix web application, which means that, with a few exceptions, any functionality available through the web-application can

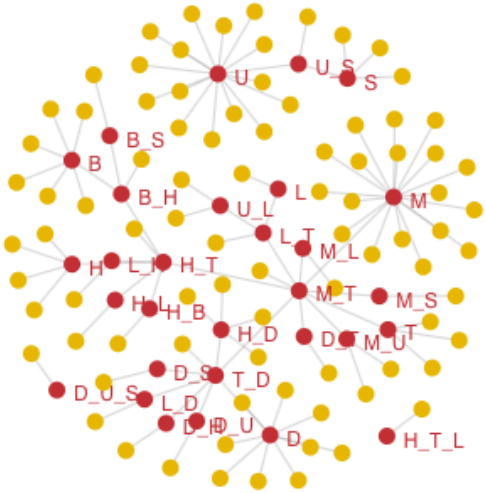

Figure 2: Interactive clusters of IntSint code combinations for 100 occurrences of *you know*.

be requested automatically by a client program using our REST service.


```
{
  "totalHits": 754,
  "docCount": 20,
  "spanCount": 21,
  "spans": [
    {
      "lp": 1,
      "sid": "1_0LDeL",
      "text_id": "sp-pl_Roee",
      "metadata": {
        "id": "1_0LDeL",
        "start": 1862390,
        "stop": 1865010,
        "seq": 647,
        "title_a": "Pogadanki rodzinne",
        "origin": "PELCRA",
        "lang": "pol",
        "source_id": "0LDeL",
        "time_aligned": true,
        "source_corpus_id": "1",
        "speaker_age": 24,
        "speaker_education": "WY",
        "speaker_role": "INTVEE",
        "speaker_sex": "FEMALE",
        "speaker_age_prec": "EXACT"}
      }]
       "matchSpan": [
      {
        "id": "154973893",
        "seq": 84,
        "tag": "adv",
        "lemma": "razem",
        "wordOrig": "razem"
      }
    ],
    "leftTxt": "kumpel z grupy z którym",
    "matchTxt": "razem",
    "rightTxt": "byliśmy w zespole ",
```

```
        "duplicate": false
    },
}
```

The media files can currently be retrieved as 16-bit WAV streams matching an arbitrary section of a recording. They may also correspond to a specific structural unit of transcription such as an utterance, a word token or a span of words matching a corpus query. For example, the timestamps specified in the `start` and `stop` fields of the concordance response shown in the listing above can be used to retrieve the corresponding audio stream for offline use. It is also possible to use an arbitrary offset and retrieve a larger context, which can then be used for manual or automatic re-alignment, further processing and annotation.

5. Planned Developments

We are planning to deposit the three newly released corpora described above in the CLARIN-PL repository as downloadable EMU databases (Cassidy and Harrington, 2001) with metadata descriptions in the CMDI format (Broeder et al., 2012), in order to increase their general visibility and reusability as offline speech databases. More collections of spoken Polish corpora are planned for inclusion as well.

6. Availability

The current version of *Spokes Mix* is available at `http://pelcra.clarin-pl.eu/spokes2-web`. It should be noted that the first stable version of this service is officially planned for June 2018, which is also when its source code will be released. Nevertheless, the current version can be used to test most of the features described in this paper.

7. Acknowledgments

Research and development described in this paper was financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

8. References

Boersma, P. (2006). Praat: doing phonetics by computer. *http://www. praat. org/*.

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., and Trippel, T. (2012). CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, volume 1.

Cassidy, S. and Harrington, J. (2001). Multi-level annotation in the Emu speech database management system. *Speech Communication*, 33(1):61–77.

Coleman, J., Baghai-Ravary, L., Pybus, J., and Grau, S. (2012). Audio BNC: the audio edition of the Spoken British National Corpus. *Phonetics Laboratory, University of Oxford*.

Guz, W. (2015). The structural non-integration of wh-clefts. *English Language & Linguistics*, 19(3):477–503.

Guz, W. (2017). Resumptive pronouns in Polish co-relative clauses. *Journal of Slavic Linguistics*, 25(1):95–130.

Hirst, D., Di Cristo, A., and Espesser, R., (2000). *Levels of Representation and Levels of Analysis for the Description of Intonation Systems*, pages 51–87. Springer Netherlands, Dordrecht.

Hirst, D. J. (2007). A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. In *Proceedings of the XVIth International Conference of Phonetic Sciences*, volume 12331236, pages 1223–1236.

Koržinek, D., Marasek, K., Brocki, Ł., and Wołk, K. (2017). Polish read speech corpus for speech tools and services. *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016*, 136:54–62.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

Pęzik, P. (2015). Spokes – a search and exploration service for conversational corpus data. In *Selected Papers from CLARIN 2014*, Linköping Electronic Conference Proceedings, pages 99–109. Linköping University Electronic Press, Linköpings universitet.

Rousseau, A., Deléglise, P., and Estève, Y. (2014). Enhancing the TEDLIUM corpus with selected data for language modeling and more TED talks. In *Proceedings of LREC*, pages 3935–3939.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006, page 5th.