Spokes – a search and exploration service for conversational corpus data

Piotr Pęzik
University of Lodz
Corpus & Computational Linguistics Laboratory
pezik@uni.lodz.pl

Keywords: conversational corpora, multimedia corpus search engine, CLARIN-PL

1. Spokes

Spokes is an online service for conversational corpus data search and exploration, currently developed as part of the Polish CLARIN infrastructure. This paper describes the data sets exposed through Spokes, the modular architecture of the service as well as a number of its planned extensions.

2. Data

The PELCRA Conversational Corpus contains over 2 million words of casual Polish spoken data collected since 1999 in a number of research projects, c.f. (Pęzik, 2012), (Waliński and Pęzik, 2007). In contrast to other speech databases and spoken corpora available for Polish, this resource includes *in vivo* recordings of casual conversations, often taken surreptitiously in everyday situations by trained acquisition agents [1]. Although the corpus has previously been released in raw source formats under open-source licenses, its full research potential has remained dormant for many potential users such as linguists and spoken discourse analysts from domains other than linguistics due to the technical difficulties related to exploring large quantities of casual conversational data. To address the need for a centralized, easy to use search and analysis service for this data, we have developed Spokes – a web-based service providing search and analysis functionality with GUI and programmatic access.

3. Architecture

A basic overview of the current Spokes architecture is presented in Fig. 1. Both the dedicated web application and other (programmatic) clients access the backend modules through a REST API service. The individual modules of this architecture are briefly described in the sections below.

3.1. Acquisition & primary data storage

The original recordings are transcribed orthographically, anonymized and aligned manually at the level of utterances with ELAN (Wittenburg et al., 2006). In addition to basic demographic metadata about the conversations and speakers, linguistic (e.g. PoS tags) and signal-processing (e.g. pitch patterns) annotation is automatically added to the original transcriptions. The data is then transferred to a relational database for management and further processing.

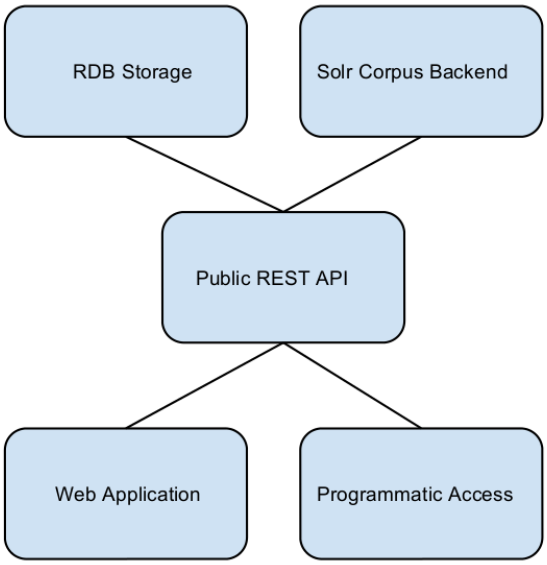

Figure 1: An overview of the Spokes architecture.

3.2. Search and exploration backend

The corpus search capabilities of the Spokes backend are based on a customized Solr index instance, which supports sequential PoS annotation queries and logical metadata filtering. The Solr backend is also used to generate on-the-fly, query-based "facets" from utterance-level metadata. For example, in addition to the basic concordance results, a corpus query such as <*pos=adj:.**> <*lemma=temat*> (which matches transcription spans corresponding to instances of adjectives preceding the lemma *temat*) returns a set of aggregated metadata counts which are instantly visualized and presented to the user as shown in Fig. 2. It should be noted that the aggregation is performed on the entire set of matching utterances rather than just the (possibly truncated) set of concordance results returned for a single query. For each utterance matching the query a link to the original audio snippet resource is provided .

In addition to these Solr-based search functionalities, a set of relational database modules provide aggregated metadata statistics for user-defined subsets of the corpus and

[1] With prior and ex post facto permissions granted by the recorded speakers to process and distribute the recordings for research purposes

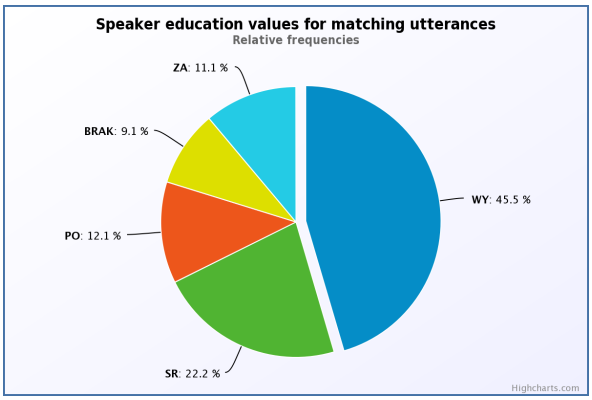

Figure 2: A facet-based demographic metadata visualization returned for a corpus query by the Solr search backend.

support annotated data retrieval. Fig. 3 shows an example of metadata aggregation results generated by the RDB backend module. The visualisations can be downloaded on generation as bitmaps and in vector graphics formats.

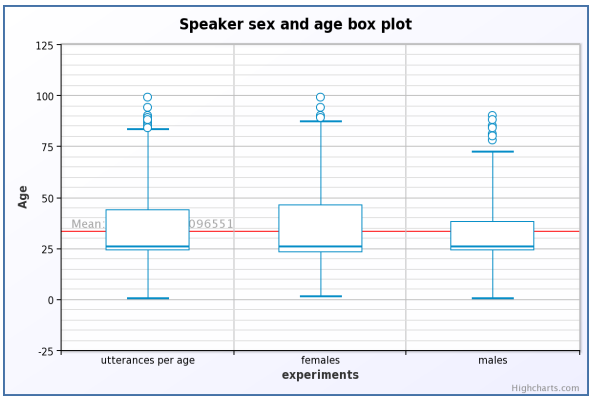

Figure 3: Box-and-whiskers plots for speaker metadata distributions generated by the relational database backend module.

3.3. Web service access

As already mentioned, the functionality of the backends is exposed exclusively through a REST web service. The Spokes web application is thus logically separated from the backend modules and it serves as a primary testbed for the publicly available REST service itself. The interactive documentation available for the service facilitates the development of client modules.

3.4. Web application

The GUI web application for Spokes provides non-technical users of the resource with search and exploration capabilities. Results obtained with the corpus search and exploration modules together with the complete source transcription can be exported on demand in popular text, image and audio formats (MS Excel, JSON, XML, SVG, JPEG, WAV). A special exploration tool for an index of formulaic conversational formulae extracted from the corpus is also available.

4. Ongoing developments

A number of new features have either been planned or are currently being implemented in the upcoming versions of Spokes.

1. To enable more advanced phonetic analyses of the data, the manually-aligned transcriptions have been automatically force-aligned at the level of word tokens. Furthermore, as illustrated in Listing 1, individual word tokens were automatically analysed into phoneme segments (represented in the example below as <p> elements) and encoded using the SAMPA alphabet (Wells and others, 1997). The phonation times shown here are relative to the offsets of the manually annotated utterance boundaries.

Listing 1: Forced alignment and segmentation in Spokes

```
<audio-segment id="Ekz6a">
   <word beg="0.090" end="0.510"
      id="nMrdkx" word="wspominac">
      <p b="0.090" e="0.120">f</p>
      <p b="0.120" e="0.150">s</p>
      <p b="0.150" e="0.210">p</p>
      <p b="0.210" e="0.240">o</p>
      <p b="0.240" e="0.290">m</p>
      <p b="0.290" e="0.330">i</p>
      <p b="0.330" e="0.360">n</p>
      <p b="0.360" e="0.430">a</p>
      <p b="0.430" e="0.510">tsi</p>
   </word>
   <word b="0.510" e="0.700"
      id="R4qxno" word="ale">
      <p b="0.510" e="0.570">a</p>
      <p b="0.570" e="0.670">l</p>
      <p b="0.670" e="0.700">e</p>
   </word>
   <word b="0.700" e="1.050"
            id="XM9dn8" word="takich">
      <p b="0.700" e="0.730">t</p>
      <p b="0.730" e="0.770">a</p>
      <p b="0.770" e="0.860">k</p>
      <p b="0.860" e="0.950">i</p>
      <p b="0.950" e="1.050">x</p>
   </word>
</audio-segment>
```

The corpus has also been analysed for selected speech signal properties, including selected prosodic features, silent pauses as well as finer-grained segmental units such as syllables. The pitch properties for the data illustrated in Listing 2 were extracted with Praat (Boersma, 2002) and they include strength, intensity and frequency values for each time point. The results of this analysis will be used to extend the search and retrieval capabilities of the service. Possible use cases include on-the-fly analysis of average phonation times for morphosyntactic patterns matching corpus queries and automatic identification and classification of pitch patterns of spans corresponding to the result sets. Due to the poor quality of some of the *in vivo*

recordings, confidence scores are used to enable filtering of "noisy" results.

Listing 2: Pitch analysis annotation.

```
<audio-segment id="Ekz6a">
   <pch s="0.778" i="0.171"
      t="0.230">164.648</pch>
   <pch s="0.899" i="0.150"
      t="0.240">164.273</pch>
   <pch s="0.915" i="0.135"
      t="0.250">164.214</pch>
   <pch s="0.936" i="0.176"
      t="0.260">163.977</pch>
   <pch s="0.960" i="0.199"
      t="0.270">163.405</pch>
   <pch s="0.934" i="0.203"
      t="0.280">161.971</pch>
   ...
</audio-segment>
```

2. Additional manual discourse annotation will be integrated in the primary database and exposed through the service. For example, subsets of the data annotated for politeness and aggression markers will be aligned with the prosodic annotation tiers and made available through the programmatic and GUI interfaces of Spokes.

3. A corpus of IDI (individual in-depth interview) transcriptions developed in a sociological research project is currently being integrated in Spokes and it will be used to showcase search and aggregation features customized for discourse analysis purposes.

4. Finally, the entire Spokes data will be available through a Federated Content Search endpoint as part of the CLARIN-PL resource center.

5. Availability

The first version of the web application is publicly available at http://clarin.pelcra.pl/Spokes. For the interactive documentation of the REST API service see http://clarin.pelcra.pl/Spokes/#help/h04.

6. Acknowledgements

The work described in this paper has been financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education. The relational database backend and the REST service was implemented mainly by Łukasz Dróżdż. Paweł Wilk and Paweł Kowalczyk are the authors of the Web application. The speech analysis modules used for the audio recordings were developed by Danijel Koržinek.

7. References

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Pęzik, P. (2012). Język mówiony w NKJP. In Przepiórkowski, A., Bańko, M., Górski, R., and Lewandowska-Tomaszczyk, B., editors, *Narodowy Korpus Języka Polskiego*, pages 37–47. Wydawnictwo Naukowe PWN, Warszawa.

Waliński, J. and Pęzik, P. (2007). Web access interface to the PELCRA referential corpus of polish. pages 65–86. Lang.

Wells, J. C. et al. (1997). Sampa computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, 4.

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.