Annotation of the Corpus of the Saeima with Multilingual Standards

Roberts Darģis[1], Ilze Auziņa[2], Uldis Bojārs[1,2], Pēteris Paikens[2], Artūrs Znotiņš[2]
Faculty of Computing, University of Latvia[1], Institute of Mathematics and Computer Science, University of Latvia[2]
Raina bulvaris 19, Riga, LV-1459, Latvia[1], Raina bulvaris 29, Riga, LV-1459, Latvia[2]
{ roberts.dargis, ilze.auzina, arturs.znotins, peteris.paikens }@lumii.lv, uldis.bojars@lu.lv

Abstract

This paper describes a release of corpus of the Saeima (parliament of Latvia) as open data resources for multidisciplinary research. The corpus consists of the transcription of Latvian parliamentary debates from 1993 until 2017, containing 38 million tokens from 468 speakers. Current comparative research of parliamentary debate is not sufficiently facilitated by simply providing unannotated corpora and results mostly in monolingual research by local researchers. We propose that augmenting such corpora with extra layers according to commonly used multilingual standards would make it easier to analyze and compare multiple corpora in different languages. In this regard, we believe that the key factors that need to be added are identifiers of entities mentioned in each utterance, and morphosyntactic information for linguistic analysis. For these reasons, the provided corpus is augmented with named entity linking to the Wikidata knowledge base (provided as linked data), automated translations to English, and morphological and syntactic annotations in Universal Dependency format. A part of this corpus is the LinkedSaeima dataset containing structured information about the Saeima debates published as Linked Open Data.

Keywords: syntactic annotation, entity linking, linked data, corpus, RDF, open government data

1. Introduction

The Corpus of the Saeima (Parliament of Latvia) was first published in 2016 (Dargis et al., 2016). At the time, the published corpus was in plain text format with speaker annotations and other metadata. With the increasing availability of corpora in different languages, we realized that unannotated corpora are not enough to facilitate comparative research across multiple language.

To enable researchers to conduct a comparative research across multiple languages without the need to know any of the languages, we propose augmenting corpora with extra layers according to commonly used multilingual standards.

In the paper we describe a new release of the Corpus of the Saeima. The new release contains multiple additional annotation layers:

- Morphosyntactic information for linguistic analysis (lemmas, morphological tags, syntactic).
- Automated translations to English.
- Named entity mentions with links to the Wikidata knowledge base.

The new release of the Corpus of the Saeima is published in multiple commonly used formats:

- A searchable text corpus in NoSketch query software (Rychlý, 2007).
- Syntactically parsed data according to the Universal Dependency standard (Nivre et al., 2016), containing morphological and syntactical annotations.
- LinkedSaeima – Linked Data representation of the corpus with structured information about Saeima proceedings and the entities mentioned in the corpus, represented in the dataset using Wikidata knowledge base identifiers.

To aid searchability for international researchers, the Linked Data format also contains text that was machine-translated to English. Speakers and roles are also linked to Wikidata entities where applicable.

2. The Data of the Corpus of the Saeima

The source data for this corpus was crawled from the Saeima's website[1] where verbatim reports of all the sessions of the Saeima are published in text format. The texts are processed using a semi-automatic pipeline to identify the boundaries of speeches and the speakers. The text is split into utterances, where each utterance contains a speech from only one speaker.

The Corpus of the Saeima includes transcriptions of parliamentary debate from 7 parliamentary terms (5th–12th), covering years 1993–2017. The transcriptions of the Corpus of the Saeima contain 38 million tokens, 497 thousand utterances and 468 speakers.

The available metadata for each utterance includes the date and type of the parliamentary session, speaker's name and affiliation.

3. Annotation Layers

3.1 Morphological and Syntactical Annotations

Morphological and syntactical annotation enables researchers to carry out quantitative analysis of different characteristics of the Corpus of the Saeima, for example:

- The use of gender pronouns in speech, depending on the gender of the speaker.
- The use of active and passive voice.
- The size of the vocabulary of different speakers.

[1] Saeima's website: http://saeima.lv/

The annotations contain lemma, part of speech, morphological features and syntactic dependencies according to the Universal Dependencies standard format.

To aid searching, texts are automatically tokenized, lemmatized and morphologically analyzed and tagged using CMM based tagger (Paikens et al., 2013). Syntactic dependencies are inferred by neural transition-based dependency parser (Znotins, 2016) trained on Latvian Universal Dependencies corpus version 2.1 (Pretkalniņa et al., 2016)[2].

3.2 Translation

The speeches from Latvian are translated to English using a neural machine translation system (Barone et al., 2017). The unreviewed machine-generated translation is provided for quantitative analysis and to aid searchability and understanding for international researchers. However, the text quality of automated translation is lacking, so for qualitative analysis a professional translator should be used.

3.3 Named Entities

For the purposes of this analysis, we developed a named entity linking system based on earlier research for news corpora analysis (Paikens, 2014). In this approach, we used the structured Wikidata information extracts provided by (Ismayilov et al, 2016) as the entity knowledge base. The Wikidata entity alias information is extended with Latvian morphological inflections and automatically generated variants for people and organization names to link the corpus mentions to Wikidata identifiers.

In the Corpus of the Saeima we identified 393 thousand mentions of 3 thousand unique entities. 165 thousand out of 497 thousand utterances contained entity mentions.

4. Available datasets

4.1 Universal Dependencies (CoNLL-U)

Automatic tokenization, morphological and syntactic annotations are published in CoNLL-U data format[3] with simple plain text based encoding, as shown in Figure 1. Columns contain word index, word form, lemma, part-of-speech tag, full morphological tag, morphological features, head of current word, universal dependency relation to head, and spacing information.

The CoNLL-U dataset is published as a language resource alongside this paper[4].

```
# newdoc id = 2016_03_31_355.txt_seq17
# newpar id = 2016_03_31_355.txt_seq17-p1
# sent_id = 2016_03_31_355.txt_seq17-p1s1
# text = Turpinām ar iesniegtajām izmaiņām Saeimas Prezidija apstiprinātajā
sēdes darba kārtībā.
 1  Turpinām      turpināt     _  vmnipt31pan  _  0  root   _  _
 2  ar            ar           _  sppd         _  4  case   _  _
 3  iesniegtajām  iesniegt     _  vmnpdfpdpsyp _  4  amod   _  _
 4  izmaiņām      izmaiņa      _  ncfpd4       _  1  iobj   _  _
 5  Saeimas       saeima       _  ncfsg4       _  6  nmod   _  _
 6  Prezidija     prezidijs    _  ncmsg1       _  8  nmod   _  _
 7  apstiprinātajā apstiprināt _  vmnpdfslpsyp _  8  amod   _  _
 8  sēdes         sēde         _  ncfsg5       _  10 nmod   _  _
 9  darba         darbs        _  ncmsg1       _  10 nmod   _  _
10  kārtībā       kārtība      _  ncfsl4       _  1  obl    _  SpaceAfter=No
11  .             .            _  zs           _  1  punct  _  _
```

Figure 1: A sample from CoNLL-U corpus.

4.2 Bonito corpus browser

The speeches from deputies of the Saeima are published in text corpus query software – NoSketch engine (Rychlý, 2007). The interface provides powerful corpus query system. Query can include words, lemmas, morphological tags and meta data. The result can be further filtered using positive or negative filters. The query result is displayed in concordances. From the result, frequencies and collocations can be computed in the NoSketch as well (Figure 2). The NoSketch query interface is available online with open access[5].

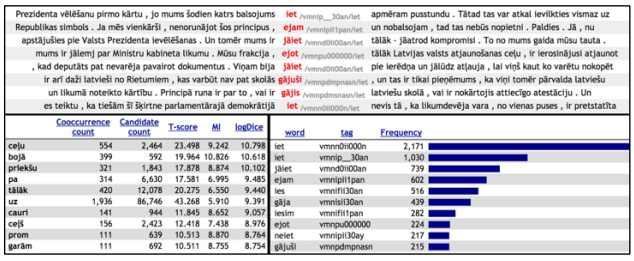

Figure 2: Screenshot of the NoSketch Engine.

4.3 Linked Data

Linked Data allows us to represent structured information about parliamentary debates by describing the properties of the objects from the domain of parliamentary meetings and relations between these objects. According to Linked Data principles, this information is represented using Resource Description Framework (RDF) (Berners-Lee, 2006).

The types of objects in the LinkedSaeima dataset[6] are:

- Meeting – a top-level concept representing one parliament meeting (a plenary) usually consisting of multiple Speeches;
- Speech – an individual speech given at a Meeting by a particular Speaker in some Role;
- Speaker – a person giving a speech;
- Role – a role (e.g. Prime Minister) which the person represented when giving a Speech.

[2] Universal Dependencies corpus version 2.1: https://github.com/UniversalDependencies/UD_Latvian
[3] CoNLL-U data format: http://universaldependencies.org/format.html
[4] The Corpus of the Saeima in CoNLL-U data format: http://saeima.korpuss.lv/datasets/ud/
[5] NoSketch server interface for the Corpus of the Saeima: dati.saeima.korpuss.lv/nosketch/
[6] LinkedSaeima dataset index page: http://dati.saeima.korpuss.lv/

For data modelling we reuse the work of the LinkedEP project (European Parliament debates as Linked Data) and their Linkedpolitics vocabulary, referenced in RDF data using prefixes *lpv* and *lpv_eu* (van Aggelen et al., 2017).

For example, a Speech is represented by *lpv_eu:Speech*, its properties include date (*dc:date*), sequence number and spoken text (*lpv:spokenText*), and it is related to the Meeting it is a part of (*dct:isPartOf*), to the Speaker (*lpv:speaker*) and its Role (*lpv:spokenAs*), and to the named entities mentioned in the text (*schema:mentions*).

The dataset is published as Linked Data and information about its objects is accessible by looking up relevant Linked Data URIs (Berners-Lee, 2006). All dataset objects have HTTP URI identifiers. The implementation uses LodLive[7] linked data browser to serve the data in HTML, RDF and multiple other formats (Figure 3).

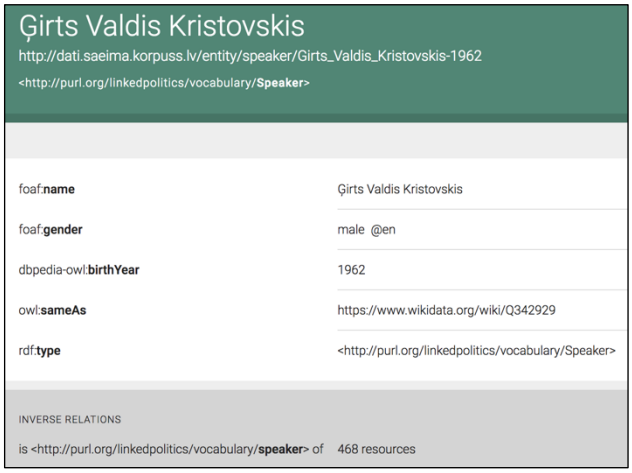

Figure 3: Screenshot of a LinkedSaeima entity in LodView.

A triple pattern fragments server and user interface[8] is published to make LinkedSaeima dataset queryable. Triple fragments server is a lightweight way for querying RDF datasets (Verborgh et al., 2016). The triple pattern fragments server can be used to query RDF dataset for RDF triplets according to any combination of subject, predicate and object (Figure 4).

The dataset is also released alongside this paper as a single RDF file that researchers can use to run more complex analysis[9].

Main innovation of this dataset, relative to the LinkedEP project, is the addition of named entity information, represented in RDF using *schema:mentions* property pointing to relevant Wikidata URI indentifiers. Another difference is that we "materialize" speaker Roles extracted from the corpus by giving them URI identifiers that can be used for querying the dataset (e.g. for speeches by presidents of the European Commission[10]).

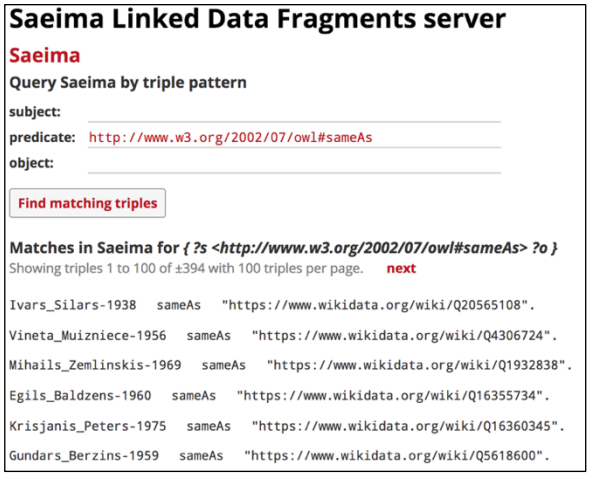

Figure 4: Screenshot of the LinkedSaeima triple pattern fragments server.

Directions for further development of the LinkedSaeima dataset include adding richer information about entity references and introducing new types of information related to the Saeima proceedings (e.g. extracting voting data). In this version, entities are represented by a property linking Speech objects to Wikidata entity identifiers. An alternative approach is to represent entity mentions as separate objects (e.g. by adapting W3C Web Annotation standard for entity references (Bojars et al., 2017). A benefit of this approach is that it can represent additional information such as the text position of the entity reference. Its downside is a larger and more complex dataset.

5. Expected use cases

Initially the corpus of the Saeima was created to facilitate the process of research for political and social scientists. The scientists have used this corpus for discourse analysis (Kruk 2007, Auzina 2007) and to oversee political and social processes in Latvia (Chojnicka 2013). It is also used by linguists as a corpus for language research (Treimane 2014).

The new annotation levels (especially named entities and translation) and its Linked Data representation will make it possible to compare Latvian parliamentary data with other national parliamentary data and to provide users with new ways for exploring this information. The described datasets have been used for different purposes:

- Annotation representation across languages for Named Entity Recognition (Ehrmann et al. 2011);
- Training and testing information extraction software;
- To produce bilingual or even multilingual cross-language resources such as dictionaries, or applications, for example, cross-lingual word sense disambiguation, cross-lingual information retrieval.

[7] LodLive linked data browser: https://github.com/dvcama/LodLive/
[8] LinkedSaeima triple pattern fragments server and user interface: http://dati.saeima.korpuss.lv/ldf/saeima
[9] LinkedSaeima RDF dump: http://saeima.korpuss.lv/datasets/rdf/
[10] URI for the President of the European Commission: http://dati.saeima.korpuss.lv/entity/role/89

6. Conclusions and further work

In conclusion, we have described a new dataset of parliamentary debate with extended annotations that should make it more useful for research and analysis.

We'd like to call upon this research community to extend their resources while keeping in mind multilingual applications. While currently parliamentary discourse analysis is fragmented, we believe that using standards that are common in NLP field we can pave the road for easy multilingual comparative analysis of many parliamentary corpora. Each country has similar data, but the language diversity and differences in technical format makes it difficult for researchers to summarize many corpora. We suggest others to investigate the possibility of providing their data in commonly used international formats, which in our opinion are Universal Dependencies for morphological and syntactic analysis, and RDF and Linked Data for entity information, with the hope of enabling new areas of research comparing parliamentary discourse of many countries.

Expected future work includes continuous processing of new debate data, improvements to entity linking and disambiguation, and extending the LinkedSaeima dataset with additional types of structured information e.g. voting data.

7. Acknowledgements

This work has been partially supported by the European Regional Development Fund (ERDF) project No. 201X/0020/2DP/2.1.1.1.0/14/ APIA/VIAA/000 at the Faculty of Computing, University of Latvia.

This work has been partially supported by the University of Latvia project AAP2016/B032 "Innovative information technologies".

The data collection process has been supported by Latvian National research program EKOSOC-LV No. 5.2.5.

The tools developed and used in this project have received financial support from the European Regional Development Fund under the grant agreement No. 1.1.1.1/16/A/219.

8. Bibliographical References

Barone, A. V. M., Helcl, J., Sennrich, R., Haddow, B., & Birch, A. (2017). Deep architectures for neural machine translation. arXiv preprint arXiv:1707.07631.

Berners-Lee, T. (2006). Linked Data – Design Issues. http://www.w3.org/DesignIssues/LinkedData.html

Bojars, U., Rasmane, A., Zogla, A. The Requirements for Semantic Annotation of Cultural Heritage Content. Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017. CEUR Workshop Proceedings, vol. 2014.

Dargis, R., Rabante-Busa, G., Auzina, I., & Kruks, S. (2016, October). ParliSearch-A System for Large Text Corpus Discourse Analysis. In Baltic HLT (pp. 115-121).

Ismayilov, A., Kontokostas, D., Auer, S., Lehmann, J., & Hellmann, S. (2016). Wikidata through the Eyes of DBpedia. Semantic Web, (Preprint), 1-11.

Nivre, J., de Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... & Tsarfaty, R. (2016, May). Universal Dependencies v1: A Multilingual Treebank Collection. In LREC.

Paikens, P. (2014, September). Latvian Newswire Information Extraction System and Entity Knowledge Base. In Baltic HLT (pp. 119-125).

Paikens, P., Rituma, L., & Pretkalnina, L. (2013, May). Morphological analysis with limited resources: Latvian example. In Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16 (No. 085, pp. 267-277). Linköping University Electronic Press.

Pretkalniņa, L., Rituma, L., & Saulīte, B. (2016, October). Universal Dependency Treebank for Latvian: a Pilot. In Human Language Technologies-The Baltic Perspective: Proceedings of the Seventh International Conference Baltic HLT 2016 (Vol. 289). IOS Press.

Rychlý, P. (2007, December). Manatee/bonito-a modular corpus manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing (pp. 65-70).

Skadina, I., & Rozis, R. (2016, October). Word Embeddings for Latvian Natural Language Processing Tools. In Human Language Technologies-The Baltic Perspective: Proceedings of the Seventh International Conference Baltic HLT 2016 (Vol. 289, p. 167). IOS Press.

van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2017). The debates of the European parliament as linked open data. Semantic Web, 8(2), 271-281.

Verborgh, R., Vander Sande, M., Hartig, O., Van Herwegen, J., De Vocht, L., De Meester, B., ... & Colpaert, P. (2016). Triple Pattern Fragments: A low-cost knowledge graph interface for the Web. Web Semantics: Science, Services and Agents on the World Wide Web, 37, 184-206.

Znotins A. (2016). Word embeddings for Latvian natural language processing tools, Proceedings of Human Language Technologies -- The Baltic Perspective, IOS Press.