*Human Language Technologies – The Baltic Perspective*
*I. Skadiņa and R. Rozis (Eds.)*


*doi:10.3233/978-1-61499-701-6-115*

# ParliSearch – A System for Large Text Corpus Discourse Analysis

Roberts DARĢIS [a,1], Guna RĀBANTE-BUŠA [a], Ilze AUZIŅA [a] and Sergejs KRUKS [b]
[a] *Institute of Mathematics and Computer Science, University of Latvia*
[b] *Rīga Stradiņš University*

**Abstract.** The paper illustrates the ParliSearch – the system that enables easy discourse analysis in large text corpus, providing stem search with additional search criteria. The system contains verbatim reports from debates of plenary sittings of the European Parliament and the Saeima (the Parliament of Latvia).

**Keywords.** Parliament debates, discourse analysis, corpus system, stem search

## 1. Introduction

Firstly, an easy way to facilitate information retrieval from the corpus of the transcripts of the Saeima's (Parliament of Latvia) sessions was needed.

Latvian is a highly inflected language. All the available tools did not provide full text stem search in Latvian. The functionality to specify additional search criteria (i.e. period of time, particular speakers or political parties) was also needed. Taking all the functional requirements into account, a decision to make a new platform was made. The main goal of the new platform was to make the discourse analysis in large text corpus easy, fast and available for everyone interested – researchers, students, journalists and other enthusiasts.

Later, after the search platform for corpus of the Saeima was created, there was a need for similar search functionality in other text corpora. So the adaption process of the original search platform was started to create the ParliSearch – a system for discourse analysis in large text corpus.

## 2. Functional Requirements

Analyzing user stories from previous users of the corpus of the Saeima – researchers of political and social sciences, a set of functional requirements was defined. The system must provide:

- full text stem search;
- ordering by relevance or date;
- result filtering by period of time;
- searching for specific speakers or positions they represent (i.e. political parties, officials from ministries, presidents, etc.);

[1] Corresponding Author: Roberts Darģis, e-mail: roberts.dargis@lumii.lv

- accessibility to everyone from everywhere without necessity to install any additional software – must be a website;
- search in depth – ability to limit results even further in each iteration adding more complex search criteria;
- limitation of criteria – available values for additional search parameters are limited to only those values, which are present in search results.

All the functional requirements must be implemented keeping the main goal in mind – to make discourse analysis easy, fast and available for everyone.

## 3. System Architecture

Database is the most important part in a system, that is designed for information retrieval. Many databases provide full text search, filtering and sorting, almost none provide stem search. Clusterpoint database, however, provides stem search for many languages, including Latvian.

Clusterpoint is available as a service with quite high resource limit in free tier, reducing the costs and necessary workload for server maintenance. Multiple language bindings are available for the API. PHP programming language was selected for backend development, because it is most widely supported by hosting providers.

At the moment, parameter configuration is hardcoded in the backend, since the system was originally developed to be used only for specific corpus. The main purpose for the backend is to convert input from user friendly graphical user interface to Clusterpoint search query and to prepare search results for displaying in the front-end.

Bootstrap front-end framework, along with some custom made controls, were used to make the layout responsive and provide modern and easy to use graphical user interface. At the moment, the backend and the frontend are not corpus independent. The layout is adjusted to each corpus separately in a way that would make arrangement of filtering options more compact and easier to perceive. Because of the simplicity of the backend and the frontend, it is relatively easy to adapt the system to be used with a different text corpus.

The database consists of utterances. The utterance boundaries are defined by the change of speakers, also called speaker turns. Each utterance has additional information about the speaker (name, country), language, text type (original or translated) and date.

## 4. Available Corpora

For now, two corpora are available in the ParliSearch system:

- saeima.kospuss.lv – the Corpus of the Saeima (Parliament of Latvia)
- europarl.korpuss.lv – the Corpus of the EuroParl (Parliament of Europe Union).

### *4.1. The Corpus of the Saeima*

The data for the Corpus of the Saeima was taken from Saeima's website [1] where transcriptions of all the sessions of the Saeima are published. Also, the information about members of the parliament can be found there.

The corpus contains transcriptions of plenary sittings of the Saeima from 7 parliamentary terms (from 5th to 11th). In this period, there have been 17 governments, from 1993 to 2014.

Transcriptions of the Corpus of Saeima contain 4.4 million tokens, 465,644 utterances and 647 speakers. Speakers are grouped in 7 categories and 83 subcategories.

### *4.2. The Corpus of the Europarl*

The material – transcriptions, translations and information about speakers – for the Corpus of the EuroParl was collected from the Talk of Europe [2] project.

Transcriptions of the Corpus of the EuroParl contain 2,769,433 utterances, 24 languages, 2,260 speakers from 28 countries. Speakers are grouped in 28 categories.

The original and translated versions of utterances are included from 3 parliamentary terms (1999-2014).

## 5. Search Options

### *5.1. Search Query*

By default, the search is inflection insensitive (searched by stems of words) and documents should contain at least one word from search query. Words from the search query can appear apart from each other in the document.

If a searched word should be inflection sensitive, it should be put in double quotation marks. Similarly, if a phrase must be matched exactly, the phrase should be put in double quotation marks.

To search documents that don't contain a certain word, a tilde should be put before it in the search query.

These are the options that are given in the help section of the corpus search query syntax. More advanced search query syntax expressions are not specified in the help section, because they are complex and are not necessary for an everyday user. More advanced search query syntax expressions, for example, Boolean expressions, advanced wildcard patterns and boosted search terms, are specified in the Clusterpoint documentation [3]. Before passing the query to Clusterpoint database all tokens that are separated with whitespace and all tokens not included in double quotes, are surrounded by stemming operator "$" to enable inflection insensitive search.

### *5.2. Search in Depth*

To help to narrow down matched utterances, multiple additional search parameters are available depending on the corpus of use.

- For the Corpus of the Saeima the date (from-to) can be indicated. Periods of parliamentary terms and governments are already specified and can be chosen. If more limitation is necessary, the choice of Saeima's five session's types can be used.
- For the Corpus of the EuroParl a similar limitations for the date can be set. Periods of European Parliaments are available as predefined intervals. In addition, the language and the type (original or translated) can be chosen.

Under "Speaker options" a specific speakers or the country they represent can be chosen.

To ease the process of narrowing down matched utterances, available values, for parameters that were not used in original search criteria, are limited to only those values, which are present in search results. Available values for parameters that were used in original search criteria are not disabled to allow to widen the search criteria.

### *5.3. Searching for Specific Speakers or Positions They Represent*

Search for specific speakers or positions they represent is available under Speaker Options. In the Corpus of the Saeima there are six categories representing speakers by their work place:

- 'Members of parliament', further divided into their political parties;
- 'Representatives of Institutions of Latvia', with separate subcategories for the most represented groups;
- 'Representatives of Ministries';
- 'Representatives of the Saeima', for example, 'Speakers of the Saeima', 'Secretaries of the Saeima' etc.;
- 'Government', containing ministers, presidents of ministers and presidents;
- 'Foreign Visitors' including 'Foreign Presidents', 'Representatives of Foreign Parliaments', and representatives from other multinational institutions, such as 'Union of Europe', 'NATO' and others.

In the corpus of the EuroParl, all speakers are categorized by the countries they represent (Figure 1). The data is not acquired from the EuroParl website directly, but from the Talk of Europe project [2] since the data on EuroParl website is not available in a structured form. Crawling and parsing the data from HTML documents requires a lot of work. The structure of speakers will be updated, when more detailed data becomes available in the Talk of Europe project.

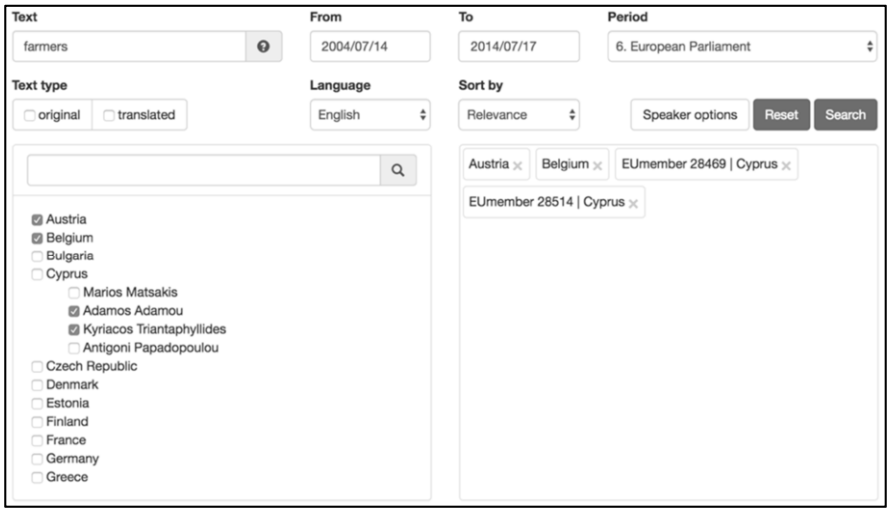


**Figure 1.** Speaker options for the corpus of the EuroParl.

To help with finding a specific speakers or positions they represent, a custom made control with search option is provided. Filtering starts when at least 4 symbols are entered to optimize systems performance.

### *5.4. Result View*

There are two possibilities to view the result – fragment view and concordance view. The fragment view allows to see the fragments that contain the words in the search query or the beginning of the speech, if no search query is provided. The concordance view is particularly useful for linguists and political context researchers, because the word or the exact phrase of interest is centered and some of the context is shown (Figure 2).

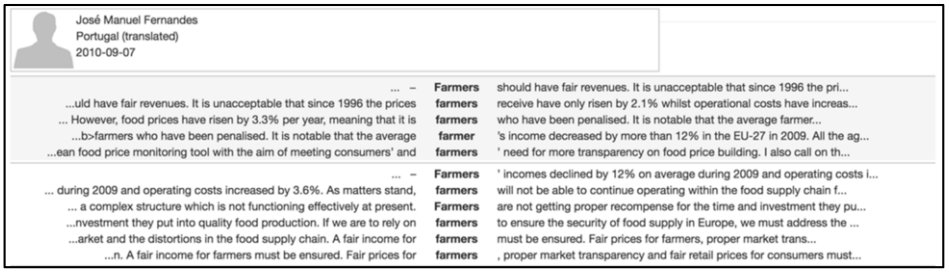


**Figure 2.** Concordances for word "farmers" in the Corpus of the EuroParl.

The results can be sorted by relevance or date (ascending or descending). When sorted by relevance, if only one word is searched, the utterances are sorted by how many times the word appears in the utterance. If more than one word is searched, the number of times the words appear in utterance is multiplied by a value inversely proportional to the distance between the words in the utterance.

Additional information about a specific utterance is shown when hovering over it. This information includes speaker, position and date of the parliament session. In the corpus of the EuroParl also the language and the text type (original or translated) are shown.

The full text of the utterance and a context is available by clicking on the specific utterance. The context includes a few utterances before and after the utterance of interest.

## 6. Statistics

To help evaluate search results, overall statistics are given after the results. Statistics consist of two parts – overview of the results (e.g. "Found 5 070 of 2 769 433 (0.18%) and fragments. Fragments found from 1999/07/20 to 2014/07/17.") and results grouped by various parameters.

Parameters, by which the results are grouped, depend on the corpus:

- For the Corpus of the Saeima, results are grouped by gender, speaker and category of the speaker.

- For the Corpus of the EuroParl, results are grouped by the country of the speaker.

In the grouped results (Figure 3) there are two different percentages:

- Percentage from the number of results, where the count is divided by the total number of results.
- Percentage from the group size, where the count is divided by the group size – the number of results in a particular group, that is acquired searching without any parameters.

## Statistics

Found 5 070 of 2 769 433 (0.18%) fragments. Fragments found from 1999/07/20 to 2014/07/17.

### Results by countries

| Country | Count | Group size | Percentage from number of results | Percentage from group size |
|---|---|---|---|---|
| Ireland | 494 | 112 128 | 9.74% | 0.44% |
| Greece | 236 | 104 316 | 4.65% | 0.23% |
| Netherlands | 218 | 98 463 | 4.30% | 0.22% |
| Austria | 232 | 109 694 | 4.58% | 0.21% |
| Luxembourg | 40 | 18 930 | 0.79% | 0.21% |
| United Kingdom | 720 | 350 352 | 14.20% | 0.21% |
| France | 549 | 268 427 | 10.83% | 0.20% |

**Figure 3.** Example of statistics for word "farmers" in the Corpus of the EuroParl.

The percentage from a group size is useful to compare different groups. For example, in Figure 3, the word "farmers" is most often (720 times or 14.20% of all the results) used by representatives from United Kingdom. Comparing percentages from group size, one can conclude, that the word "farmers", relative to group size, most often (0.44%) is used by representatives from Ireland and representatives from United Kingdom have used it only 0.21% of the time.

## 7. Use Cases

Initially the Corpus of the Saeima was created to ease the process of research for political and social scientists. They use the Corpus of Saeima to find political views of members of Saeima and to analyze how situations in the country and world are reflected in the parliament debates.

The journalists have found it useful as a resource for gathering information before interviews and political debates. For example, asking a politician about an opinion he has expressed in a parliamentary session.

Linguists can use the corpus for political discourse and language studies [4][5]. For example, to analyze language differences between genders or political parties, to research the transformation of the language usage through the time, or to investigate differences between the language in the parliament and the language in everyday use [6].

In the Corpus of the Saeima, there is a section containing a list of research that is based on this corpus.

## 8. Conclusion and Future Work

The developed system was approved by journalists, political and social scientists and students. New use cases were discovered and ideas for improvements were received.

One of the main priorities is to add the possibility to listen or watch the records of utterances. The next priority is to implement the functionality to share specific utterance on different websites using the embedded code.

Future plans also include adding new corpora. The process for the adaption of the Corpus of the Seimas (Parliament of Lithuania) has started.

By adding a new corpus to the system the process of generalizing the system has started. The aim is to create an open source system, that would be available for everyone without the necessity to modify the source code of the system.

## Acknowledgements

This work has been partially supported by Latvian State Research Programme EKOSOC-LV Nr. 9.5.

## References

[1] Saeima. http://saeima.lv/en/about-saeima/archives-all-sections
[2] A.E. van Aggelen, L. Hollink, Plenary debates of the European Parliament as Linked Open Data. http://www.talkofeurope.eu/data/. Website accessed on [October, 2015].
[3] Clusterpoint. https://www.clusterpoint.com/docs/?page=Matching_Documents
[4] I. Auziņa, LR Saeimas sēžu stenogrammu datorizēta apstrāde un analīze, *Parlamentārais diskurss Latvijā*. Latvijas Universitāte (2007), 9 – 21.
[5] I. Skadiņa, I. Auziņa, N. Grūzītis, K. Levāne-Petrova, G. Nešpore, R. Skadiņš, A. Vasiļjevs, Language Resources and Technology for the Humanities in Latvia (2004-2010), *Human Language Technologies – the Baltic Perspective. Proceedings of the Fourth International Conference Baltic HLT 2010.* IOS press (2010), 15–22.
[6] L. Treimane, *Valodas reģistrs parlamentā: sistēmiski funkcionālais skatījums* Theses PhD (2014).