ParlaMint Ngram Viewer: Multilingual Comparative Diachronic Search Across 26 Parliaments

Asher de Jong, Taja Kuzman, Maik Larooij, Maarten Marx
Information Retrieval Lab, Informatics Institute, University of Amsterdam,
Department of Knowledge Technologies, Jožef Stefan Institute, Slovenia
asher2912@gmail.com, taja.kuzman@ijs.si, {larooij|maartenmarx}@uva.nl

Abstract

We demonstrate the multilingual search engine and Ngram viewer that was built on top of the Parlamint dataset (Erjavec et al., 2023), using the recently available translations (Kuzman et al., 2023). The user interface and SERP are carefully designed for querying parliamentary proceedings and for the intended use by citizens, journalists and political scholars. **Demo:** https://debateabase.wooverheid.nl/

Keywords: Multilingual Search, Parliamentary Proceedings, Ngram Viewer, Machine Translation

1. Introduction

The ParlaMint collection contains the complete parliamentary proceedings of 26 European national and regional parliaments, all in the same XML format, from the period 2015–2022 (Erjavec et al., 2023; Kuzman et al., 2023). Strong analysis tools like the Sketch Engine concordancer are available for (corpus) linguists, but access to this valuable dataset for social scientists and the general public has been lacking. So we decided to build a dedicated parliamentary search engine for ParlaMint. The availability of good quality automatic translations of all corpora to English (Kuzman et al., 2023) made it possible to develop a multilingual search and analysis tool, allowing both scholars and ordinary citizens to compare stances, opinions, and policies about a topic across different nations. We developed two integrated information systems for this data. The first entry after a query is a diachronic comparative saliency analysis tool, reminiscent of Google's Ngram viewer (Mann et al., 2014), that provides a fast and clear overview of the development of topics through time and across nations. From this in essence unordered faceted presentation of search results, the user can enter the vertical search engine yielding relevance ranked speeches given in various parliaments.

This paper describes the broad technical details, zooms in on the design choices made for the user interaction, and provides details of the automatic translation process. Our demo is available at https://debateabase.wooverheid.nl/, the raw data at http://hdl.handle.net/11356/1810, and the code for creating the demo at https://github.com/AsherIDE/Debate-a-Base.

Related Work With more and more easy to process parliamentary corpora becoming available, we saw several non-governmental initiatives to open up the proceedings to the general public with specialized vertical search engines e.g., Marx (2009); Beelen et al. (2017); Kaptein and Marx (2010), a process that started in 2003 with TheyWorkForYou.com in the UK. The proceedings of the European Parliament were multilingual from the early beginning, and the EuroParl corpus (Koehn, 2005) kickstarted the field of statistical machine translation. Cross-language information retrieval is an active research field since the late 1990's (Oard and Diekema, 1998) and is still very relevant today (Nie, 2022). Ngram viewers have been used to visualize and analyse temporal and comparative trends in multilingual corpus linguistics (Lin et al., 2012), psychology (Pettit, 2016), geosciences (Brandt, 2018), and political speech (de Goede et al., 2013).

2. The ParlaMint Dataset

The search engine uses the Parlamint.ana 3.0 dataset[1] (Erjavec et al., 2023) and its machine-translated English version, ParlaMint-en.ana 3.0[2] (Kuzman et al., 2023). The corpora were collected in the ParlaMint II project[3], which focused on creation and curation of parliamentary corpora from different countries in a harmonised and uniform format (Erjavec et al., 2023).

The ParlaMint 3.0 corpora include parliamentary sessions from 26 national and regional parliaments with a total of over 1.2 billion words (Erjavec et al., 2023). All corpora encompass the sessions held in the 8 years between 2015 and 2022, with many also including earlier sessions. The corpus collection consists of 27 languages; 24 in the Latin alphabet, 2 in Cyrillic (Bulgarian and Ukrainian corpus) and 1 in

[1] http://hdl.handle.net/11356/1488
[2] http://hdl.handle.net/11356/1810
[3] https://www.clarin.eu/parlamint

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 110–115

Country	Years	Speeches	EN tokens	Tokens	Speakers	Parties	Languages
Austria	27	228K	67M	66M	853	9	German
Bosnia-Hz.	25	126K	22M	18M	603	40	Bosnian
Belgium	9	199K	43M	43M	787	66	Dutch, French
Bulgaria	9	210K	30M	27M	1,033	19	Bulgarian
Czech Republic	10	181K	34M	28M	592	19	Czech
Denmark	9	399K	43M	41M	383	19	Danish
Estonia	12	228K	32M	23M	488	6	Estonian
Spain: Catalonia	8	50K	16M	16M	364	21	Catalan, Spanish
Spain: Galicia	8	83K	19M	18M	227	7	Galician
France	6	715K	47M	49M	908	26	French
Great Britain	8	671K	-	126M	1,951	2	English
Greece	8	342K	53M	50M	635	13	Greek
Croatia	20	504K	103M	88M	1,036	45	Croatian
Hungary	9	105K	35M	28M	426	9	Hungarian
Iceland	8	95K	33M	31M	261	9	Icelandic
Italy	10	173K	34M	31M	771	45	Italian
Latvia	9	163K	13M	9M	234	11	Latvian
Netherlands	9	609K	68M	68M	586	35	Dutch
Norway	25	399K	99M	89M	1,106	13	Norwegian
Poland	8	228K	44M	36M	1,223	9	Polish
Portugal	8	171K	18M	18M	723	10	Portuguese
Serbia	26	316K	99M	85M	1,724	71	Serbian
Sweden	8	85K	33M	29M	650	13	Swedish
Slovenia	23	311K	83M	70M	973	27	Slovenian
Turkiye	12	681K	63M	45M	1,346	5	Turkish
Ukraine	12	196K	23M	19M	2,192	48	Ukrainian, Russian
Total	-	7.5M	1.2B	1.2B	22K	597	27 langs

Table 1: For each corpus in the Parlamint collection: number of years, speeches, tokens in English and in the original language, number of different speakers, parties, and the languages of the proceedings. *Note*: Total English tokens represent the number of tokens in the corpora that were machine-translated into English. As the British parliamentary corpus is originally in English, it was not included in the machine-translated ParlaMint-en.ana corpus.

the Greek alphabet. Certain corpora are bilingual, such as the Belgian and Catalan corpus. The sizes of the corpora are presented in Table 1.

While the ParlaMint corpora in original languages are a very rich source of information, most users would be able to search only a small part of the corpus that is in the languages which they understand. That is why we included in the search engine the translated version as well – the ParlaMint-en.ana 3.0[4] corpus (Kuzman et al., 2023), which allows the users to browse through the entire dataset at once in one language.

The ParlaMint-en.ana 3.0 corpora (Kuzman et al., 2023) provide the English translations obtained with machine translation using the pre-trained OPUS-MT models (Tiedemann and Thottingal, 2020). These freely-available[5] Transformer-based models are based on the MarianNMT neural machine translation toolbox (Junczys-Dowmunt et al., 2018) and were trained on parallel corpora from the OPUS repository (Tiedemann, 2012). For each language, a manual evaluation of a translated sample was conducted to determine the most suitable model. The evaluations confirmed that the translations exhibited satisfactory quality. However, it is important for users of the search engine to be aware that the translations contain errors. The manual evaluation revealed incorrect translations of proper names, terms, and multi-word expressions, as well as repetitions, insertions, and incorrect translations that are unrelated to the source sentence (commonly referred to as "hallucinations" of MT systems). The search engine's interface allows users to verify the accuracy of the translations by toggling between the translated and the source text.

[4] http://hdl.handle.net/11356/1810

[5] https://github.com/Helsinki-NLP/Opus-MT

3. The Demo

The aim of the Ngram viewer (Figure 1) is to provide insight into the relative use of a phrase (ngram) through time, and to compare these temporal developments across countries: it is a diachronic comparison tool. Users can temporally zoom in on the visualization and view the relative counts also in months and even days. If the user is interested in the debates that were held at a certain day, she simply clicks in the ngram graph and is redirected to the search engine result page listing all speeches of that day relevant for the given phrase.

Users can search, read and compare debates on the debates page (Figure 2). Through this page a user can also gain insights into the actual statements that politicians made, by only having to provide the topic they are interested in. It is possible to filter on country, person, political party and date.

3.1. Interface Design

The interface design of the search engine was based on the SERP (Search Engine Result Page) design principles laid out by Hearst (Hearst, 2009). It features two main screens, the Ngram viewer (Figure 1) and the SERP combined with a document inspector (Figure 2). An important design choice was to use all the non-linguistic metadata in the collection (like name, gender, party affiliation of speakers) in the SERP. The added numbers 1–7 in Figure 2 highlight some of the design choices specially made for multilingual parliamentary search: 1: the ranked list of speeches that match the query; 2: inspection of a user-opened speech shown in the context of the surrounding debate; 3: filters for querying with the values of the used filters highlighted; 4: Debate file identifier (same #tag identifier means same debate); 5: move to next and previous speeches in the debate which are hits for the query; 6: highlighting of used search terms; 7: button to switch between the original language and English translation of the debate.

The design of the Ngram viewer is standard. To normalize counts across parliaments, it shows the fraction of speeches containing the N-gram. To reduce clutter, parliaments with few hits for a Ngram are ignored in the viewer, and the user can remove more. Users can temporally zoom in, and clicking on a line brings the user to the SERP for the Ngram as query restricted to the parliament connected to the clicked line.

3.2. Back-End Framework

The website is built with the Python based[6] Flask web framework. The search engine is built in Elasticsearch (ES) and uses the default BM25 ranking[7], but with slight tweaks to return exact string matches for Ngram queries. Normal debate speech queries only have to contain the queried word. Both the website and ES are placed in a Docker container. Following https://www.theyworkforyou.com/, we took the individual speeches as the objects, which are indexed and returned after a query.

All XML files were extracted with a variety of Python scripts that can be found on our Github repo. The complete corpus contains 7.5 million speeches. For the debates overview, all data was uploaded to ES, where one row contained one speech. In ES, one can respond to Ngram queries using phrase-queries but this turned out to be much too slow. So, in line with other Ngram viewer architectures, we simply precomputed the number of hits for each Ngram (N between 1 and 5) for each parliament, and for each day, month and year and stored these as documents in ES. With 5.8 billion different Ngrams, this did not fit into a regular ES index that has a limit of about 2.1 billion[8], so we created an index for each N.

3.3. Search Engine Evaluation

After the creation of the website, it was tested by 10 participants, with a mean age of 30 and varying educational backgrounds. Participants had to answer 7 questions using our Ngram viewer and search engine. An intervention only occurred if the participant got stuck on a question. During the experiment, participants were encouraged to think aloud constantly. From the results it became apparent that the debates page was not clear enough about which search results (speeches) belonged to the same debate. We improved the design by adding a document number next to each search result. Some participants did not realize they could jump to the speeches from the Ngrams page. To solve this, we added instructions below the visualization.

4. Conclusion

Our goal was to make the ParlaMint corpus easily available to a much wider audience, in particular to people with very little technical background. As the main aim of ParlaMint is the ability to *compare* speeches through time and across nations, we designed our interface based on that principle. The

[6] https://www.fullstackpython.com/flask.html

[7] https://www.elastic.co/guide/en/elasticsearch/reference/7.17/similarity.html

[8] https://issues.apache.org/jira/browse/LUCENE-5843

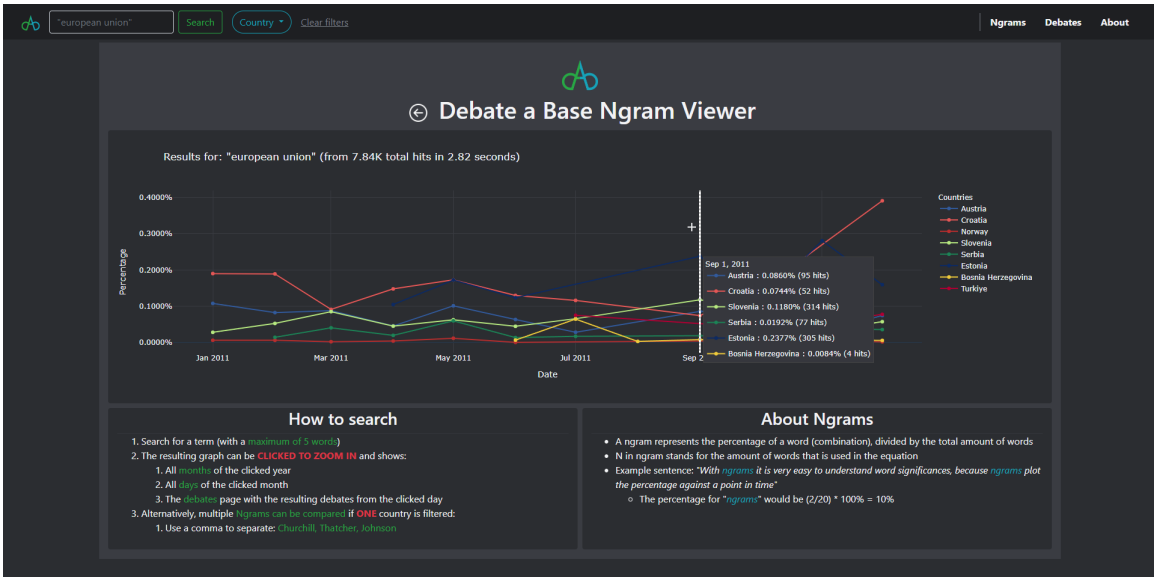

Figure 1: Debateabase Ngram Viewer

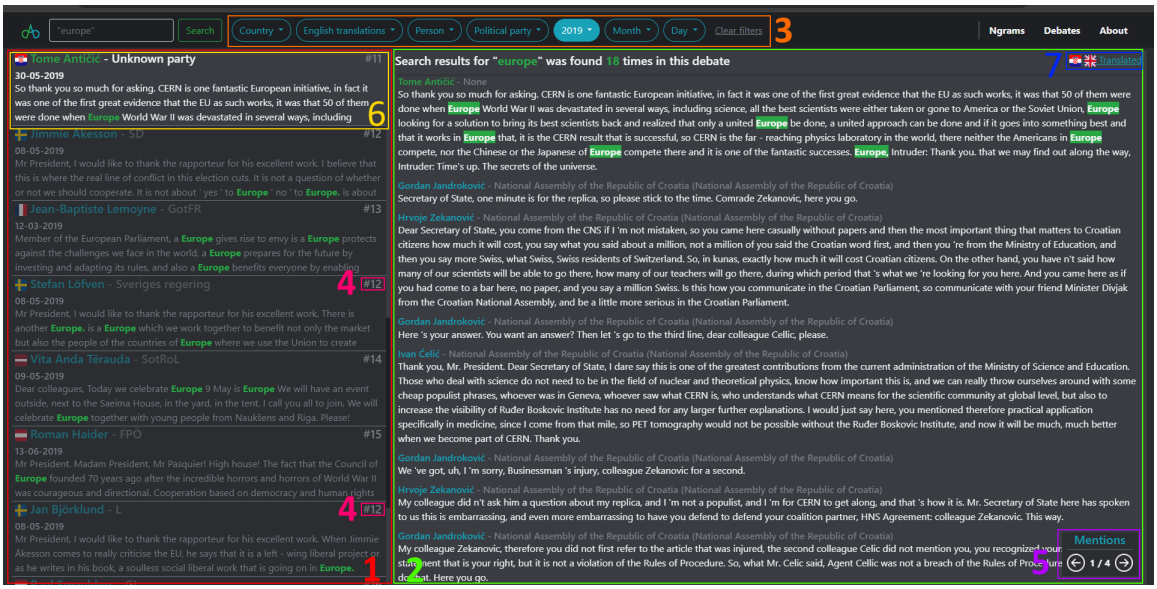

Figure 2: Debateabase SERP; design choices highlighted by numerals 1-7.

ParlaMint corpus made corpus linguistic comparisons possible by standardizing the technical format of all debates, but it is the availability of the translations into one language that makes *comparisons on content* possible. This also opens up the corpus to a far wider group of users.

We designed our system using time-tested examples: a Google style search engine, speeches as the unit of retrieval and counting, as initiated by TheyWorkForYou, the Ngram viewer in which we can compare normalized saliency timelines across parliaments, and proven-to-work interface choices for dealing with facets and multilinguality.

The goal of the demo is really to show the richness of the ParlaMint corpus (that is why we also included e.g., the political party of a speaker and more information), and to provide a somewhat familiar manner to explore its vast possibilities. Our hope is that (the idea of) this demo is taken up by a party which can sustain it and hopefully also keep the whole corpus up to date. The demo shows that with limited computing resources and freely available software a strong prototype covering all of ParlaMint can indeed be created. The value of a corpus lies in its use. Our aim with the demo is to widen that use both to a new audience and to new types of questions asked to the ParlaMint corpus.

5. Acknowledgements

This research was supported in part by the Netherlands Organization for Scientific Research (NWO) through the ACCESS project grant CISC.CC.016 and an NWO Open Science Fund grant nr 01607400. The creation of the ParlaMint corpora was funded by CLARIN ERIC and by in-kind contributions of partners from ParlaMint I, (https://www.clarin.eu/parlamint#Partners), including POIR.04.02.00-00C002/19 and the Polish Ministry of Education and Science 2022/WK/09, programs Language resources and technologies for Slovene P6-0411 and Digital Humanities P6-0436, projects LiLaH N6-0099 and MEZZANINE J7-4642, all funded by the Slovenian Research Agency; CLaDA-BG DO01-301/17.12.21, funded by the Bulgarian Ministry of Education and Science; the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

6. Bibliographical References

Kaspar Beelen, Timothy Alberdingk Thijm, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, Ludovic Rheault, et al. 2017. Digitization of the Canadian parliamentary debates. *Canadian Journal of Political Science/Revue Canadienne de Science Politique*, 50(3):849–864.

Danita S Brandt. 2018. Charting the geosciences with Google Ngram Viewer. *GSA Today*, 5:66–67.

Bart de Goede, Justin van Wees, Maarten Marx, and Ridho Reinanda. 2013. PoliticalMashup Ngramviewer. In *Research and Advanced Technology for Digital Libraries*, pages 446–449. Springer.

Tomaz Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubesic, Kiril Simov, Andrej Pancur, Michal Rudolf, Matyás Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çagri Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevicius, Tomas Krilavicius, Roberts Dargis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fiser. 2023. The ParlaMint corpora of parliamentary proceedings. volume 57, pages 415–448.

Marti Hearst. 2009. *Search User Interfaces*. Cambridge University Press.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Rianne Kaptein and Maarten Marx. 2010. Focused retrieval and result aggregation with political data. *Information retrieval*, 13:412–433.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit X: Papers*, pages 79–86.

Yuri Lin, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google books Ngram corpus. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 169–174.

Jason Mann, David Zhang, Lu Yang, Dipanjan Das, and Slav Petrov. 2014. Enhanced search with wildcards and morphological inflections in the google books ngram viewer. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 115–120.

Maarten Marx. 2009. Advanced information access to parliamentary debates. *Journal of Digital Information*, 10(6).

Maarten Marx, Nelleke Aders, and Anne Schuth. 2010. Digital sustainable publication of legacy parliamentary proceedings. In *Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities*, dg.o '10, page 99–104. Digital Government Society of North America.

Jian-Yun Nie. 2022. *Cross-language information retrieval*. Springer Nature.

Douglas W Oard and Anne R Diekema. 1998. Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33:223–56.

Michael Pettit. 2016. Historical time in the age of big data: Cultural psychology, historical change, and the Google Books Ngram Viewer. *History of psychology*, 19(2):141.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *LREC*, volume 2012, pages 2214–2218. Citeseer.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT–Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

7. Language Resource References

Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej and Osenova, Petya and Fišer, Darja et al. 2023. *Multilingual comparable corpora of parliamentary debates ParlaMint 3.0*. Slovenian language resource repository CLARIN.SI.

Kuzman, Taja and Ljubešić, Nikola and Erjavec, Tomaž and Kopp, Matyáš and Ogrodniczuk, Maciej et al. 2023. *Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 3.0*. Slovenian language resource repository CLARIN.SI.