CLARIN Corpora for Parliamentary Discourse Research

Darja Fišer[1,2], Jakob Lenardič[1]
[1]Faculty of Arts, University of Ljubljana, Aškerčeva 2, 1000 Ljubljana, Slovenia
[2]Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
{darja.fiser, jakob.lenardic}@ff.uni-lj.si

Parliamentary debates are an important resource for many disciplines in digital humanities and social sciences because they contain impactful information and special, formalized and often persuasive and emotional language. This paper presents the parliamentary corpora in the CLARIN infrastructure and suggests how they could be made more readily available to digital humanities and social sciences researchers in order to promote interdisciplinary, trans-national and cross-cultural studies.

1. Introduction

Parliamentary discourse displays specific institutional discursive features, complies with a set of rules and conventions, is motivated by a wide range of communicative goals such as persuasion, negotiation and agenda-setting along ideological or party lines, and is characterized by institutional role-based commitments, dialogically shaped institutional confrontation and the awareness of a multi-layered audience (Ilie, 2017). Due to their unique content, structure and language, records of parliamentary sessions have been a quintessential resource for a wide range of research questions from a number of disciplines in digital humanities and social sciences for the past 50 years (Chester and Browning, 1962; Franklin and Norton, 1993), such as political science (van Dijk, 2010), sociology (Cheng, 2015), history (Pančur and Šorn 2016), discourse analysis (Hirst et al., 2014), sociolinguistics (Rheault et al., 2015) and multilinguality (Bayley et al., 2004) but has only recently started to acquire a truly interdisciplinary scope (Bayley, 2004; Ihalainen et al., 2016). With an increasingly decisive role of parliaments and their rapidly changing relations with the public, media, government and international organizations, further empirical research and development of richly annotated and integrative analytical tools is necessary to achieve a better understanding of the specificities of parliamentary discourse and its wider societal impact, in particular with studies that take into account diverse parts of society (women, minorities, marginalized groups) and cross-cultural dimensions.

In most countries, access to parliamentary records is becoming increasingly simple due to Freedom of Information Acts, which has sparked a number of national and international initiatives that are compiling parliamentary data into valuable, often richly annotated parliamentary corpora. Several of the developed parliamentary corpora in the CLARIN infrastructure have already been successfully used in scientific research in various disciplines. In computational linguistics, the Lithuanian corpus was the basis for the development of machine learning approaches for classifying political text in accordance with its ideological position (Kapočiūtė-Dzikienė and Krupavičius, 2014), as well as for a stylometric analysis to distinguish the styles of left-wing, centre-wing and right-wing parties (Mandravickaitė and Krilavičius, 2015). Recently, Meurer (2017) has used, among other corpora, *Talk of Norway* to develop dependency relations from LFG structures. In corpus linguistics, Sverredal (2014) has used the *Korp* version of the *Riksdag's Open Data* to conduct a corpus-based analysis of the development of plural forms in Swedish finite verbs. Pančur and Šorn (2016) have argued for the necessity of using corpora in historical studies to aid with exploring large amounts of historical sources with a showcase on the Slovene parliamentary corpus *SlovParl*.

Unfortunately, corpus development efforts are seldom co-ordinated, and as a consequence the resources are not uniformly sampled, annotated, formatted or documented, and in many cases not even made easily accessible. In order to promote comparability and reproducibility of research results as well as foster interdisciplinary, trans-national and cross-cultural studies, this paper gives an overview of the parliamentary corpora available through CLARIN, the European research infrastructure for language resources and technology (Hinrichs and Krauwer, 2016). We also discuss how they could be made more readily available to the heterogeneous research community, especially colleagues without an engineering background.

2. Overview of CLARIN parliamentary corpora

Table 1: Overview of the parliamentary corpora in CLARIN, sorted by country code.

Country	Size (mil tokens)	Period	Linguistic annotation
cz	0.5	/	Speech-text alignment
de	0.4	1998-2015	/
dk	7.3	2008-2010	T, PoS, L
ee	13	1995-2001	/
el	28.7	2011-2015	/
fi	2.2	2008-2016	/
fr	0.17	2002-2012	/
lt	23.9	1990-2013	T, PoS, L
no_1	63.8	1998-2016	T, PoS, L
no_2	29	2008-2015	/
pt	1	1970-2008	T, PoS, L
se	1,250	1971-2016	T, PoS, L, Semantic
si	10.8	1990-1992	T, PoS, L
uk_1	1,600	1803-2005	T, PoS, L
uk_2	0.19	1998-2015	/
eu	588	1996-2011	Sentence alignment

In total, there are 16 parliamentary corpora accessible through the CLARIN infrastructure. Apart from the multilingual *Europarl* corpus (Koehn, 2005), which contains debates from the European parliament in 21 languages, there are 2 corpora of British parliamentary debates, 2 corpora of Norwegian debates and 1 corpus per country, for the following 11 countries: Czech Republic, Denmark, Estonia, Finland, France, Germany, Lithuania, Portugal, Slovenia, and Sweden. Table 1 gives an overview of the identified corpora in terms of size, period, and linguistic annotation.[1] The handles to the corpora are given in the Language resources section at the end of the paper.

2.1. Large monolingual corpora

Czech: ***Czech Parliament Meetings*** (Pražák and Šmidl, 2012) consists of audio recordings and related transcriptions that correspond to approximately 500,000 tokens. It is available for download on the website of the Czech repository *LINDAT* under the public CC-BY licence and for online querying through *KonText*.[2] The transcriptions of parliamentary discussions were semi-automatically aligned to the recordings and annotated with speaker-related information.

Danish: ***DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget*** (CLARIN-DK, 2011) includes Danish parliamentary proceedings from 2008-2010 and consists of 7.3 million tokens. The corpus is tokenised, PoS-tagged and lemmatised and is available for download from the Danish repository *CLARIN-DK* under a non-specific public licence.

Estonian: ***Transcripts of Riigikogu*** (University of Tartu, 2014) consists of approximately 13 million tokens and covers the time span 1995-2001. Aside from TEI-annotation, it is unclear how the corpus is annotated. The corpus can be downloaded under a non-specific academic licence on the corpus webpage or accessed online through the *Keeleveeb Query* concordancer provided by CLARIN-Estonia.

Finnish: The ***Eduskunta corpus*** **(**Bartis, 2017a; 2017b; Lennes, 2017) covers Finnish parliamentary data for the period between 2008 and 2016. The corpus consists of 2.2 million tokens. The corpus can be downloaded from the Finnish repository *Language Bank of Finland*, which also provides the associated videos of the sessions, as well as queried through the concordancer *Korp* (Finnish distribution).[3]

Greek: ***Hellenic Parliament Sittings*** (clarin:el, 2015) includes Greek parliamentary proceedings for 2011-2015 and consists of 28.7 million tokens. It is unclear how the corpus is annotated. This corpus is available for download under the academically-restricted CC BY-NC licence from the Greek repository *clarin:el*.

Lithuanian: ***Lithuanian Parliament Corpus for Authorship Attribution*** **(**Kapočiūtė-Dzikienė et al., 2017) includes Lithuanian parliamentary data for 1990-2013 and consists of 23.9 million tokens. The corpus is tokenised, PoS-tagged and lemmatised. This corpus can be downloaded from the CLARIN-LT repository under a CLARIN-LT public licence.

Norwegian: There are two Norwegian parliamentary corpora – ***Talk of Norway*** (Lapponi and Søyland 2016) and ***Proceedings of Norwegian Parliamentary debates*** (Common Language Resources and Technology Infrastructure Norway, 2015). *Talk of Norway* covers Norwegian parliamentary speech for 1998-2016, consists of 63.8 million tokens, and is available for download through the *CLARINO* repository, while *Proceedings of Norwegian Parliamentary Debates* covers a slightly shorter period, 2008-2015, consists of 29 million tokens and is only available for online querying through the concordancer *Corpuscle*.[4] Both corpora are available under the NLOD public licence.

Portuguese: ***PTPARL Corpus*** (ELRA, 2008) covers Portuguese parliamentary proceedings from 1970-2008 and consists of approximately 1 million tokens. The corpus is tokenised, PoS-tagged and lemmatised. It is listed for download in the ELRA catalogue[5] under the non-commercial ELRA END USER and commercial ELRA VAR licences.

Slovene: The ***SlovParl*** (Pančur et al., 2017) corpus covers Slovene parliamentary proceedings for 1990-1992 and in its latest version consists of 10.8 million tokens. The corpus is tokenised, PoS-tagged, and lemmatised. The corpus is available for download through the CLARIN.SI repository under CC BY and available for online querying through the CLARIN.SI concordancers.[6]

Swedish: ***Riksdag's Open Data*** consists of 1.25 billion tokens for 1971-2016 and is thus the second largest of the parliamentary corpora in the CLARIN infrastructure. The corpus is tokenised, PoS-tagged, lemmatised, and contains annotations of lemgrams, compounds and named entities. It is available through the *Språkbanken* repository and can either be downloaded through or queried online through *Korp* (Swedish distribution).[7] The corpus is available under CC BY.

UK: ***The Hansard Corpus*** **(**The SAMUELS Project, 2016) consists of 1.6 billion tokens from 1803-2005 and is the largest parliamentary corpus in the CLARIN infrastructure both in word size and temporal span. The corpus is tokenised, PoS-tagged, lemmatised and also displays seep semantic annotation. It is listed on the website of CLARIN-UK and is available for querying through the *BYU* concordancer.

[1] T = Tokenisation; PoS = Part-of-Speech tagging; L = lemmatisation

[2] http://lindat.mff.cuni.cz/services/kontext/first_form?corpname=czechparl_2012_03_28_cs_w.

[3] https://korp.csc.fi/.

[4] http://clarino.uib.no/korpuskel/page.

[5] http://catalog.elra.info/product_info.php?products_id=1179.

[6] http://www.clarin.si/info/concordances/.

[7] https://spraakbanken.gu.se/korp/.

2.2. Small monolingual corpora

In addition to the large corpora, the three smaller thematic corpora for English, French and German parliamentary speech are available for download under CC BY through the French ORTOLANG repository. These three corpora are the English ***Parliamentary Debates on Europe at the House of Commons*** (Truan, 2016a), the French ***Parliamentary Debates on Europe at the Assemblée nationale*** (Truan, 2016b), and the German ***Parliamentary Debates on Europe at the Bundestag*** (Truan, 2016c). Unlike the previously discussed large corpora, these contain only those parliamentary debates that correspond to the annual European Council meetings at the respective parliaments. The French corpus is for 2002-2012 while the English and German corpora cover a longer period, 1998-2015. In terms of token size, the English and French corpora are the smallest (approximately 190,000 and 173,000 tokens respectively), while the German corpus is slightly larger (approximately 417,000 tokens).

2.3. Multilingual corpus

The ***Europarl*** corpus (Koehn, 2005) is a multilingual parallel corpus of the sessions of the European Parliament. It covers the period 1996-2011, consists of 588 million tokens, is tokenised, sentence aligned and marked for speakers, and is freely available for download on a dedicated page[8] under no specified licence.

2.4. The state of the infrastructure

In general, the identified parliamentary corpora are well integrated within the CLARIN infrastructure. Almost all of the 16 corpora from Table 1 are listed in the Virtual Language Observatory (VLO),[9] which is the main metadata-based portal for language resources of the CLARIN infrastructure and provides the access point to finding resources across the national CLARIN centres (Uytvanck et al., 2012). The only exceptions are (i) *The Hansard Corpus*, (ii) *Hellenic Parliament Sittings* and (iii) the *Riksdag's Open Data* corpus, which are listed only in the respective national repositories (e.g. *CLARIN-UK* for *The Hansard Corpus*).

In terms of availability, 5 corpora can be both downloaded and accessed through an online concordancer (the Czech *Czech Parliament Meetings*, the Estonian *Transcripts of Riigikogu*, the Finnish *Eduskunta corpus*, the Swedish *Riksdag's Open Data*, and the Slovene *SlovParl* corpus), 3 can only be queried through an online environment (the British *Hansard Corpus*, the *Hungarian National Corpus* and *Proceedings of Norwegian Parliamentary Debates*), and the rest of the 9 corpora can only be downloaded.

Parliamentary corpora in the CLARIN infrastructure are described with high-quality metadata. Information on size and time period of the corpora is readily available (except for the temporal period included in the Czech corpus). Information on linguistic annotation is available for all the corpora except for the Finnish, Greek, Estonian and the *Proceedings of Norwegian Parliamentary debates* corpora. Although the documentation on the three thematic corpora described in section 2.2 refers to "Annotation of conversation"[10], the information on levels of linguistic annotation (e.g. PoS-tagging) is not given.

3. Findings from the CLARIN Focus Groups

In addition to evaluating the existence, findability, documentation and accessibility of parliamentary corpora in the CLARIN infrastructure presented in Section 2, we wanted to better understand how users experience the digital research infrastructure that CLARIN provides. To this aim, we conducted two half-day focus group interviews (Sanders, 2017) with 11 researchers from different disciplines from 10 European countries who are interested in CLARIN's parliamentary resources, asking them to share their experiences with the CLARIN infrastructure, obstacles they encountered, suggestions for improvement and the support and training they need.

Results indicate that both Social Sciences and Humanities researchers and speech and language technology/IT experts need more guidance about the CLARIN datasets, corpora and tools relevant for parliamentary data. First and foremost, they expressed a need for a more explicit metadata policy to ensure that high quality materials are easily available and accessible. In addition to easy access and navigation towards the relevant resources and tools, they also recommended that thorough documentation, training materials and best practice use cases for parliamentary data be provided in an enhanced online research environment. They also called for more systematic promotion campaigns, as CLARIN and its resources and tools are still unknown in many relevant research communities in their opinion. In the long run, it was recommended that CLARIN develops procedures to guarantee and monitor the quality of not only corpus metadata but also the quality of data and tools and to offer clearly visible information on recent updates of resources and tools.

4. Recommendations towards improved visibility of CLARIN parliamentary corpora

Based on the results of the resource survey and the focus group on parliamentary data we propose below recommendations to increase the visibility of these corpora to the heterogeneous and international research community, to showcase their potential for interdisciplinary, trans-national and cross-cultural studies, and to alleviate the technical obstacles that are preventing the use of the resources on a larger scale. The recommendations are comprehensive in the sense that they address all stages in the lifecycle of a resource and involve all the key players, such as resource developers, curators, infrastructure providers, knowledge sharing experts, and funders. While some of the recommendations require minimal to moderate post-production or curation efforts that can be handled centrally, others would require a substantial investment and direct involvement of the developers and curators. Despite the fact that this might not be a feasible short-term goal, the recommendations

[8] http://www.statmt.org/europarl/.
[9] https://vlo.clarin.eu.
[10] https://hdl.handle.net/11403/fr-parl/v1.

could be implemented in stages in future extensions or refinements of the existing resources, as well as by initiatives that are building new parliamentary corpora.

Intended use and users. Hughes et al. (2016) point out that "we can no longer take the impact and value of our expensive digital resources for granted, and it is not sufficient to make assumptions about use and users of digital collections". This is why we need to sample, annotate, format, document and release parliamentary corpora in such a way that they will be valuable to scholars with diverse backgrounds beyond corpus and computational linguistics which is still the prevalent situation in the CLARIN community. This issue is very important because in other disciplines different research data sampling methodologies are required (controlling for sociodemographic features, or topic-, event- or concept-based filtering etc.). An obvious development in this respect would be comprehensive data inclusion policies and regular updates of corpora with new material so that researchers could analyse the most recent but also chronologically the most diverse parliamentary activities. A more ambitious development would be semantic integration within and across parliamentary corpora. This would enable researchers to track and compare the same concepts and topics in different parliaments. A major boost would also be achieved by cross-referencing parliamentary corpora with external knowledge bases, such as place-name gazetteers and biographical lexica as well as with external documents, such as legislation and media coverage.

User interfaces and documentation. The results of the focus groups systematically show that the developers of CLARIN's tools and resources are generally overestimating users in terms of technological solutions they are offering to the researchers but underestimating them in terms of documentation about the tools and resources they believe will be relevant for the researchers. Overall, easy access to resources and straightforward user interfaces were emphasized the most and seem to carry the most impact. In addition, researchers attempting comparative studies reported interface fatigue (especially when offered in a language researchers are not proficient in, only partially localized into English or run on different platforms, resulting in different functionality as well as different results of seemingly identical functionalities). This is why researchers have expressed a need to be able to use a single tool for all parliamentary corpora that would require less time and effort to master but would also ensure that quantitative results are comparable across corpora. Good documentation was also pointed out as prerequisite for resource and tool criticism and interpreting research results (e.g. speech transcription and editing policy). On the other hand, the most frequent users expressed a desire for more complex functionalities of the interfaces and access to more advanced tools, such as distant reading, text mining and visualization applications which are currently not offered for a large majority of the available parliamentary corpora. This suggests that a balanced development of both simple and advanced solutions might be the most successful long-term solution.

Data structure and annotation. A prerequisite for a successful integration of multilingual and multinational parliamentary information into a single research environment is a systematic, incremental roadmap which requires all corpus developers to comply with a set of mutually agreed upon building blocks and text annotation, corpus encoding and metadata encoding standards. This will make the data at least formally uniform and will enable exploration and comparison across corpora.

Outreach activities and knowledge sharing. On-going promotion of parliamentary resources is of paramount importance, which was also confirmed in our focus groups. Namely, researchers will be most likely to use a resource or a tool if it is recommended to them by a colleague or in a training event they attend. While this is positive, it is not enough to result in a significant increase in users, and may be insufficient to maintain existing numbers. This is a common problem with most resources developed within projects which are funded for limited periods. However, a research infrastructure such as CLARIN has the instruments to ensure a recurrent budget for the promotion of its resources. According to the focus group results, researchers should be provided with use cases that demonstrate the importance and potential of parliamentary corpora to investigate research questions in their discipline. In addition to merely showcasing examples of research questions that can successfully be answered with parliamentary resources, the use cases should also demonstrate how advanced ICT approaches can be utilized in these kinds of studies. Apart from the use cases aimed at professional researchers, the need for educational use cases that can be integrated into university curricula have also been highlighted.

5. Conclusion

In this paper we have presented the parliamentary corpora available via the CLARIN infrastructure and analysed the level of their integration into the infrastructure, the quality of the associated metadata and ease of access. In general, the numerous parliamentary corpora are well integrated within the CLARIN infrastructure, their metadata is of high quality and most of the corpora can be downloaded. In terms of user on-line interfaces, parliamentary corpora are offered through many different concordancers which is an obstacle for users from different research backgrounds, international users and for users embarking on comparable research. In the framework of our efforts to make the corpora more visible and readily available to researchers from digital humanities and social sciences and to promote interdisciplinary, trans-national and cross-cultural studies, we have proposed some recommendations to make corpora more universally useful research datasets, to overcome technical and documentation barriers and to showcase the potential of parliamentary resources in research and education. They range from low-lying fruit to long-term policies and call for centralized interventions as well as for direct involvement of the resource developers and curators the actions of which need to be carefully motivated, planned, co-ordinated, monitored and evaluated by a designated task force.

6. Acknowledgements

The work reported in this paper has been supported by the member countries and observers in the CLARIN ERIC, and it has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 676529 for project CLARIN-PLUS. We would like to thank all the national User Involvement Coordinators and researchers who have provided invaluable feedback on our surveys. We would also like to thank the reviewers for their valuable comments.

7. Bibliographical References

Bayley, P. (Ed.). (2004). *Cross-cultural perspectives on parliamentary discourse (Vol. 10)*. John Benjamins Publishing: Amsterdam.

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., and Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *Sixth Swedish Language Technology Conference 2016*. http://people.cs.umu.se/johanna/sltc2016/abstracts/SLTC_2016_paper_31.pdf.

Cheng, J. E. (2015). Islamophobia, Muslimophobia or racism? Parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. *Discourse & Society*, 26(5): 562-586.

Chester, D. N. and Bowring, N. (1962). *Questions in parliament*. Clarendon Press: London.

Franklin, M. and Norton, P. (1993). *Parliamentary questions*. Oxford University Press: Oxford.

Hinrichs, E. and Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In N. Calzolari et. al. (Eds.), *Proceedings of LREC 2014 : 9th International Conference on Language Resources and Evaluation*, 1525–31.

Hirst, G., Feng, V. W., Cochrane, C., and Naderi, N. (2014). Argumentation, Ideology, and Issue Framing in Parliamentary Discourse. In E. Cabrio, S. Villata, A.Z. Wyner (Eds.), *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*.

Hughes, L. M., Ell, P. S., Knight, G. A., and Dobreva, M. (2013). Assessing and measuring impact of a digital collection in the humanities: An analysis of the SPHERE (Stormont Parliamentary Hansards: Embedded in Research and Education) Project. *Digital Scholarship in the Humanities*, 30(2): 183-198.

Ihalainen, P., Ilie, C., and Palonen, K. (2016). *Parliament and Parliamentarism: A Comparative History of a European Concept*. Berghahn Books: Oxford, NY.

Ilie, C. (2017). Parliamentary Debates. In R. Wodak and B. Forchtner (Eds.), *The Routledge Handbook of Language and Politics*.

Kapočiūtė-Dzikienė, J. and Krupavičius, A. (2014). Predicting Party Group from the Lithuanian Parliamentary Speeches. *Information Technology and Control*, 43(3): 321-332.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf.

Mandravickaitė, J. and Krilavičius, T. (2015). Language usage of members of the Lithuanian Parliament considering their political orientation. *Deeds and Days*, 64: 133-151.

Meurer, P. (2017). From LFG structures to dependency relations. *Bergen Language and Linguistic Studies*, 8: 183-201.

Oravecz C., Váradi, T., and Sass, B. (2014). "The Hungarian Gigaword Corpus." In N. Calzolari et al. (Eds.), *Proceedings of LREC 2014 : 9th International Conference on Language Resources and Evaluation*, 1719-23.

Pančur, A. and Šorn, M. (2016). Smart Big Data : Use of Slovenian Parliamentary Papers in Digital History. *Contributions to Contemporary History*, 56(3): 130-146.

Rayson, P., Baron, A., Piao, S., and Wattam, S. (2015). Large-scale Time-sensitive Semantic Analysis of Historical Corpora. In *Proceedings of the 36th Meeting of ICAME*.

Sanders, W. (2017). Focus Group on User Involvement conducted during the CLARIN-PLUS Workshop "Working with Parliamentary Records", Sofia, Bulgaria, 27 March 2017.

Sverredal, K. (2014). Obehöriga verb äga ej tillträde En undersökning av verbets pluralkongruens i svenska. http://uu.diva-portal.org/smash/get/diva2:850961/FULLTEXT01.pdf.

Van Dijk, T. A. (2010). Political identities in parliamentary debates. European Parliaments under Scrutiny. In C. Ilie (Ed.), *European Parliaments under Scrutiny: Discourse strategies and interaction practices*, 29-56.

Van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). Semantic metadata mapping in practice: The Virtual Language Observatory. In Calzolari et al. (Eds.), *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*, 1029-1034.

8. Language Resource References

Bartis, I. (2017a). Plenary Sessions of the Parliament of Finland, Downloadable Version 1. https://vlo.clarin.eu/record?4&docId=http_58__47__47_urn.fi_47_urn_58_nbn_58_fi_58_lb-2017030901&q=parliament+of+finland&index=1&count=1081450.

Bartis, I. (2017b). Plenary Sessions of the Parliament of Finland, Kielipankki Korp Version 1. https://vlo.clarin.eu/record?5&docId=http_58__47__47_urn.fi_47_urn_58_nbn_58_fi_58_lb-2017020202&q=Plenary+Sessions+of+the+Parliament+of+Finland&index=2&count=1096117.

Common Language Resources and Technology Infrastructure Norway. (2005). Proceedings of Norwegian parliamentary debates (2008-2015). http://clarino.uib.no/korpuskel/landing-page?resource=stortinget&view=short.

clarin:el. (2015). Hellenic Parliament Sittings (2011-2015). http://hdl.grnet.gr/11500/AEGEAN-0000-0000-2545-9

DK-CLARIN. (2011). DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget, CLARIN-DK repository. https://clarin.dk/clarindk/item.jsp?id=dkclarin:986010#body_short-0.

ELRA. (2008). PTPARL Corpus. http://catalog.elra.info/product_info.php?products_id=1179.

Kapočiūtė-Dzikienė, J., Šarkutė, L. & Utka, A. (2017). Lithuanian Parliament Corpus for Authorship Attribution, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/17.

Lapponi, E. and Søyland, M. G. (2016). Talk of Norway, Common Language Resources and Technology Infrastructure Norway (CLARINO) Bergen Repository, http://hdl.handle.net/11509/123.

Lennes, M. (2017). Plenary Sessions of the Parliament of Finland, Kielipankki LAT Version 1. https://vlo.clarin.eu/record?6&docId=http_58__47__47_urn.fi_47_urn_58_nbn_58_fi_58_lb-2017122021&q=Plenary+Sessions+of+the+Parliament+of+Finland&index=3&count=1096117.

Pančur, A., Šorn, M., and Erjavec, T. (2016). Slovenian parliamentary corpus SlovParl 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1075.

Pančur, A., Šorn, M., and Erjavec, T. (2017). Slovenian parliamentary corpus SlovParl 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1167.

Pražák, A. and Šmídl, L. (2012). Czech Parliament Meetings, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4.

The SAMUELS project. (2016). The Hansard Corpus. https://www.hansard-corpus.org/.

Truan, N. (2016a). Parliamentary Debates on Europe at the House of Commons (1998-2015) [Corpus]. ORTOLANG (Open Resources and Tools for LANGuage). https://hdl.handle.net/11403/uk-parl.

Truan, N. (2016b). Parliamentary Debates on Europe at the assemblée nationale [Corpus]. ORTOLANG (Open Resources and Tools for LANGuage). https://hdl.handle.net/11403/fr-parl/v1/.

Truan, N. (2016c). Parliamentary Debates on Europe at the Bundestag [Corpus]. ORTOLANG (Open Resources and Tools for LANGuage). https://hdl.handle.net/11403/de-parl/v1/.

University of Tartu. (2014). Transcripts of Riigikogu (Estonian Parliament). http://www.cl.ut.ee/korpused/segakorpus/riigikogu/.

Váradi, T. (2005). Hungarian National Corpus. http://hdl.handle.net/11372/LRT-345