Parliamentary Corpora in the CLARIN infrastructure

Darja Fišer
Department of Translation
Faculty of Arts, University of Ljubljana
darja.fiser@ff.uni-lj.si

Jakob Lenardič
Department of Translation
Faculty of Arts, University of Ljubljana
jakob.lenardic@ff.uni-lj.si

Abstract

This paper gives an overview of the parliamentary records and corpora from CLARIN countries with a focus on an analysis of their availability through the CLARIN infrastructure. Based on the results of the survey we draw a list of recommendations to optimize the depositing and cataloguing of the corpora in the CLARIN repositories in order to make them readily accessible for researchers from different disciplines.

1 Introduction

Due to its unique content, structure and language, records of parliamentary sessions have always been a quintessential resource for a wide range of research questions from a number of disciplines in Digital Humanities and Social Sciences, 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) as well as Multilinguality (Bayley et al. 2004). The good availability of parliamentary data in digitized form and granted access rights to public information in the EU countries have motivated a number of national as well as international initiatives to compile, process and analyse parliamentary corpora. The corpora were also the subject of a CLARIN-PLUS workshop[1] which aimed to bring together corpus developers and researchers using these resources. The aim of the workshop was to discuss technical issues related to proper structuring and archiving of such corpora and to address methodological questions about how to best use them in different disciplines. As examples of such use, the Finnish parliamentary corpus has already been successfully used in Discourse Analysis (Voutilainen 2017), the Swedish corpus for the analysis of governmental policies related to Swedish film (Norén and Snickars 2016), the Greek corpus for analyzing aggressive political discourse (Georgalidou 2017), and the Polish corpus for finding latent topics in parliamentary speech (Kwiatkowska 2017).

In order to gain an understanding how well the CLARIN infrastructure caters for this line of research, we conducted a survey for all member and observers CLARIN ERIC countries[2] with which we aimed to identify the existing resources and check to which extent they are integrated in the CLARIN infrastructure. In this paper we give a summary of the results, thereby highlighting aspects in which the accessibility of these corpora as well as the presentation of the relevant information can be optimised for researchers from different disciplines.

2 Parliamentary records

We focus on transcriptions of parliamentary records which are freely available on the relevant parliamentary websites for all the countries except Estonia. The records differ from one another in two respects. First, the periods that they cover vary widely from country to country. The oldest records are those of the U.K. (from 1807 on), Norway (from 1814 on) and the Netherlands (from 1814 on), while the youngest are those of Germany (from 2013 on), Poland (from 2015 on) and the Czech Republic (from 2013 on). Second, the records are offered in various formats: in the majority of cases (14 out of 18 countries), the records are available in html, docx or xls, whereas 4 countries (Germany, Italy, Poland, Portugal) have the records available only in pdf which is more difficult to process.

[1] https://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records.

[2] France was not included in the survey because the survey started before its observer status was confirmed.

3 Corpora of parliamentary records

In total, we identified 20 corpora of national parliamentary data from all CLARIN countries except Italy. We found two parliamentary corpora for Norway and the Czech Republic each. We also took into account the Europarl corpus of the proceedings of the European Parliament.

In Table 1 we summarize[3] the results of our survey by providing the following information for each corpus:

- the name of the corpus and the link where it can be accessed;
- the size of the corpus and the period it covers;
- the type of linguistic annotation included (T = tokenisation, PoS = part-of-speech tagging, L = lemmatisation); and
- how the corpus was found and how it is available (D = downloadable, C = concordancer).

Table 1: Overview of the parliamentary corpora

NC	Corpus	Size (mil tok)	Period	Anno	Found	Avail.
at	Korpusbasierte Analyse österreichischer Parlamentsreden	1.2	2013-2015	T, PoS	E-Mail	D
bg	Corpus of Bulgarian Political and Journalistic Speech	10	2006-2012	T, PoS, L	E-Mail	C
cz	CzechParl	82	1993-2010	T, PoS, L	Google	C
cz	Czech Parliament Meetings	0.5	/	/	VLO, LINDAT	D
dk	DK-CLARIN Almensprogligt korpus	7.3	2008-2010	T, PoS, L	VLO, DK-CLARIN	D
ee	Transcripts of Riigikogu	13	1995-2001	/	VLO, CLARIN-EE	D, C
fi	Eduskunta Corpus	22.5	2008-2016	/	FIN-CLARIN	C
de	PolMine Sample Corpus	/	/	/	E-Mail	D
it	/	/	/	/	/	/
el	Hellenic Parliament Sittings	28.6	2011-2015	/	CLARIN:EL	D
lv	SAEIMA	/	1993-2016	/	E-Mail	C
lt	Project Astra *STENOGRAMOS_INDV*	30	1990-2013	T, PoS, L	E-Mail	D
nl	DutchParl	800	1814-2014	T, PoS, L	E-Mail	D, C
no	Talk of Norway	64	1998-2016	T, PoS, L	Google	D
no	Proceedings of Norwegian Parliamentary Debates	29	2008-2015	T	VLO	C
se	Riksdag's Open Data	1,250	1971-2016	T, L	SWE:CLARIN	D, C
pl	The Polish Parliamentary Corpus	300	1991-2017	T, L	E-Mail	D, C
pt	PTPARL Corpus	1	1970-2008	T, PoS, L	VLO	D
si	SlovParl	3.2	1990-1992	T, PoS, L	VLO	D, C
hu	Hungarian National Corpus	22	/	T, PoS	VLO	C
uk	Hansard Corpus	1,600	1803-2005	T,PoS,L	CLARIN-UK	C
eu	Europarl Corpus	/	1996-2011	/	LINDAT	D

[3] The complete results are available here: https://office.clarin.eu/v/CE-2017-1019-Parliamentary-data-report.docx

Only half of the existing corpora in Table 1 can be found within the CLARIN infrastructure. The search in VLO yields the following 7 corpora and additional 7 corpora are available in the repositories or websites of the national consortia. Information regarding the 10 remaining identified corpora was provided either by the national User Involvement coordinators or found on Google.

Seven corpora are accessible only through online search environments – these are *KORP* for the Finnish corpus, the *Sketch Engine* for *CzechParl*, the *noSketch Engine* for the Latvian corpus, *Corpuscule* for the *Proceedings of Norwegian Parliamentary Debates* corpus, CLaRK for the Bulgarian corpus, the HNC for the Hungarian corpus and Hansard Corpus Online for the British one.

Nine corpora appear to be available for download only. These are the Austrian, Danish, German, Greek, Portuguese, Lithuanian corpora, as well as the *Czech Parliament Meetings, Talk of Norway* and *Europarl* corpora. Five corpora are available both for download and on-line searching; in these five cases, the relevant search environments are *Political Mashup* for Dutch, *Keeleveeb* for the Estonian corpus, *KORP* for the Swedish corpus, the *noSketchEngine* for the Slovenian one and the *NKJP* for the Polish one.

4 Discussion and recommendations

The results of the survey show that while parliamentary records and corpora exist for nearly all CLARIN countries, there is still a lot of room for improvement to make them readily and easily available for research through the CLARIN infrastructure. Most importantly, finding the relevant corpora within the infrastructure is not easy. At the time of our survey, only a fraction of the identified relevant corpora were found through the VLO with basic keyword searchers (*parliament* or *parliamentary*), namely the Estonian, Slovenian and *Proceedings of Norwegian Parliamentary Debates*. The Portuguese and Danish corpora were found through the VLO only after querying the full corpus name, which is, of course, a serious limitation since many users will not know the official name of the resource they are looking for. What is more, the *Czech Parliament Meetings*, Finnish, Greek, Swedish and the British corpora, on the other hand, could not be found through VLO at all, only on the repositories or websites of the national consortia while as much as half of the identified relevant corpora are yet to be incorporated to the CLARIN infrastructure.. In addition, the documentation for the surveyed corpora is incomplete. In some cases it is not obvious how the corpora are annotated (e.g. Estonia, Finland). Sometimes other important information, such as corpus size (no. of tokens) is also not readily available (e.g. *Talk of Norway*).

As parliamentary corpora are of great value for researchers from a wide range of disciplines and a lot of effort had already been invested in producing them, we propose that their developers and curators adopt the following suggestions to make them better accessible through the CLARIN infrastructure:

- create a virtual collection pointing to a landing page (ideally with a PID) for the corpora;
- add the missing corpora to the repository of a certified CLARIN centre after which they will be automatically added to the VLO via metadata harvesting;
- improve the metadata of the existing corpora in order to make them more accessible for the end user.

For improving the metadata, follow the best practices below:

- use *parliament(ary)* in the title of the metadata file, so that it gets included in target queries (e.g. https://vlo.clarin.eu/?q=name:parliament*);
- use the word *parliament(ary)* in the title (and description) and provide descriptions in English that include one of these words or an equivalent term, which will lead to higher ranking;
- use a distinctive title (not e.g. 148 times Flemish parliamentary debate https://vlo.clarin.eu/?q=Flemish+parliamentary+debate);
- when providing highly granular metadata descriptions (many + detailed), make sure to use hierarchies (cf. https://www.clarin.eu/faq/how-can-i-create-hierarchical-collection-cmdi so that the top node appears first in the VLO);

- include licencing information, which also helps with the ranking of hits in VLO, especially if the level is/maps to PUB or ACA.

5 Conclusion

In the survey we provided an overview of the parliamentary records and corpora of CLARIN member countries. We have been able to find the parliamentary records for all the countries except for Estonia and corpora for all the countries except Italy. While this is commendable, our survey highlights that not all the essential information about the corpora is easily available and, most importantly, that most of the existing corpora cannot be found through the Virtual Language Observatory. For this reason, we have drawn up a list of recommendations to improve corpus metadata in order to improve findability and ranking of the corpora by VLO. In the future, we plan to create a Virtual Collection with all the identified parliamentary corpora and develop a model to ensure interoperability of the corpora and integrate them into a single concordancer in order to make them as readily accessible for researchers from different disciplines as well as for cross-border and cross-lingual projects which is where CLARIN is in the unique position to facilitate such endeavours. In this respect, we will also draft an overview of successful applications of the corpora in Digital Humanities and Social Sciences, as such information also valuably complements the corpora. We also plan to conduct a follow-up survey in order to evaluate the effect of the proposed recommendations as well as the uptake of the improved resources at regular intervals.

6 References

[Bayley et al. 2004] Paul Bayley, Cinzia Bevitori, Elisabetta Zoni. 2004. Threat and fear in parliamentary debates in Britain, Germany and Italy, *Cross-Cultural Perspectives on Parliamentary Discourse,* 185-236.

[Cheng 2015] Jennifer E Cheng. 2015. Islamophobia, Muslimophobia or racism? Parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. http://journals.sagepub.com/doi/pdf/10.1177/0957926515581157.

[van Dijk 2010] Teun A. van Dijk. 2010. Political Identities in Parliamentary Debates. http://www.discourses.org/OldArticles/Political%20Identities%20in%20Parliamentary%20Debates.pdf.

[Georgalidou 2017] Marianthi Georgalidou. 2017. Using the Greek parliamentary speech corpus for the study of aggressive political discourse. https://www.clarin.eu/sites/default/files/4-georgalidou.pdf.

[Hirst et al. 2014] Graeme Hirst, Vanessa Wei Feng, Christopher Cochrane, Nona Naderi. 2014. Argumentation, Ideology, and Issue Framing in Parliamentary Discourse. http://ceur-ws.org/Vol-1341/paper6.pdf.

[Kwiatkowska 2017] Agnieszka Kwiatkowska. 2017. Finding latent dimensions in Polish parliamentary debates. https://www.clarin.eu/sites/default/files/6-kwiatkowska.pdf.

[Norén and Snickars 2016] Fredrik Norén, Pelle Snickars. 2016. Distant Reading the History of Swedish Film Politics—in 4,500 Governmental SOU Reports. http://pellesnickars.se/2016/12/distant-reading-the-history-of-swedish-film-politics-in-4500-governmental-sou-reports/

[Pančur and Šorn 2016] Andrej Pančur, Mojca Šorn. 2016. Smart Big Data: use of Slovenian parliamentary papers in digital history, Prispevki za novejšo zgodovino, 56:3, 130-146.

[Rheault et al. 2015] Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, Graeme Hirst. 2015. Measuring Emotion in Parliamentary Debates Using Methods of Natural Language Processing. http://www.cs.toronto.edu/pub/gh/Rheault-etal-CPSA-2015.pdf.

[Voutilainen 2017] Eero Voutilainen. 2017. Parliamentary Records as Data for Linguistic Discourse Studies. http://videolectures.net/clarinplusworkshop2017_voutilainen_studies/.