Parliamentary Corpora in the CLARIN infrastructure

Darja Fišer
Department of Translation
Faculty of Arts, University of Ljubljana
Department of Knowledge Technologies,
Jožef Stefan Institute
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 provide a comprehensive overview of the corpora as well as 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. We also analyse the recall and precision of simple and faceted search of parliamentary corpora in the Virtual Language Observatory.

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), the Norwegian corpus for developing dependency relations from LFG structures (Meurer, 2017) and the Lithuanian corpus for a stylometric analysis of parliamentary speech (Mandravickaitė and Krilavičius, 2015).

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 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 provide a comprehensive presentation of the results and highlight 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. Additionally, we evaluate the recall and precision of simple and faceted search of parliamentary corpora in the Virtual Language Observatory.

2 Corpora of parliamentary records within the CLARIN infrastructure

In total, we identified 15 corpora of national parliamentary data of CLARIN countries that are either available through a national repository or listed in the VLO. There exists one such corpus for each of the following 11 countries: The Czech Republic, Denmark, Estonia, Finland, France,

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

Germany, Greece, Lithuania, Portugal, Slovenia and Sweden. There exist two corpora for Norway and Great Britain. We also took into account the Europarl corpus of the proceedings of the European Parliament, so there are 16 corpora in total.

In what follows, we describe each corpus in turn on the basis of the results of our survey[2] by providing the following information for each corpus:

- the name of the corpus, which also provides a hyperlink to the relevant handle where it can be accessed;
- the size of the corpus and the period it covers;
- the type of linguistic annotation included;[3] and
- how the corpus is available (downloadable or through a concordancer or both).

2.1 Presentation of the CLARIN parliamentary corpora

The Hansard Corpus. This is the main corpus of the British Parliament. Covering the period between 1803 and 2005 and consisting of 1.6 billion tokens, it is the largest parliamentary corpus both in token size and temporal span. In addition to being tokenised, PoS-tagged, and lemmatised, the corpus is also characterized by deep semantic annotation pertaining to the classification of words based on historical concepts and thematic categories done with the *Historical Thesaurus Semantic Tagger* (Rayson et al., 2015). It is listed in the repository of the British observer CLARIN-UK and is presented as an online resource for querying through a dedicated concordancer. It is not listed in the VLO.

Parliamentary Debates on Europe at the House of Commons. This is the second, much smaller British parliamentary corpus. It is a thematically-focused corpus in that it contains only those parliamentary debates that correspond to the annual European Council meetings at the British parliament for the period between 1998 and 2005. It is roughly 190,000 tokens in size and its annotation consists of "mixed conversational analysis". The corpus is available for download through the French ORTOLANG repository under CC-BY and is found on the VLO.

Czech Parliament Meetings. This corpus is the only parliamentary corpus that we have identified to consist of both transcripts and associated audio recordings – there are 88 hours of speech data from the Czech parliament for an unknown period corresponding to approximately 500,000 tokens. The annotation constitutes correction of errors, adding of proper punctuation and labelling of speech sections with information about the speaker. It is available for download on the website of the Czech repository *LINDAT* under the public CC-BY-NC-ND licence and is found on the VLO.

DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget. This corpus covers Danish parliamentary proceedings for the period between 2008 and 2010 and is 7.3 million tokens in size. It is available for download from the Danish repository *CLARIN-DK* under a non-specific public licence. As for annotation, the corpus is tokenised, PoS-tagged and lemmatised. The corpus is listed in the VLO.

Transcripts of Riigikogu (Estonian Parliament). This corpus consists of Estonian parliamentary proceedings from the period between 1995 and 2001 and is approximately 13 million tokens in size. It is unclear how the corpus is linguistically annotated. It is available for download on the corpus webpage and also accessible through the *Keeleveeb Query* concordancer provided by CLARIN-Estonia. It is listed in the VLO and is available under a non-specific academic licence.

Eduskunta Corpus. The *Eduskunta Corpus* is the corpus of Finnish parliamentary debates. There are 3 versions listed in the VLO:

(1) Plenary Sessions of the Parliament of Finland, Downloadable Version 1
(2) Plenary Sessions of the Parliament of Finland, Kielipankki Korp Version 1
(3) Plenary Sessions of the Parliament of Finland, Kielipankki LAT Version 1

Each version covers Finnish parliamentary data for the period between 2008 and 2016. Versions (1) and (2) consist of the same data (2.2 million tokens), while the downloadable variant (1), which is available in the Finnish repository *Language Bank of Finland*, also provides the associated videos of the sessions. Corpus (1) is available for download under the CC-BY-NC-ND licence although the

[2] The complete results are available here: https://office.clarin.eu/v/CE-2017-1019-Parliamentary-data-report.docx
[3] We list the annotation tools only in case the information is explicitly provided by the documentation.

access is restricted in that it requires a relevant institutional account – by contrast, corpus (2), while also available under CC-BY-NC-ND, is accessible through the *Korp* concordancer without restrictions. Corpus (3) is a variant of (1) that is currently in development and is set to be made freely available. It will provide reduced variants of the videos contained in the restricted version (1) processed through the *LAT* platform.[4] The three versions of the *Eduskunta Corpus* are found on the VLO.

Parliamentary Debates on Europe at the Assemblée nationale. This is the French parliamentary corpus. Like the smaller British, it is a thematically-focused corpus in that it contains only those parliamentary debates that correspond to the annual European Council meetings at the French parliament for the period between 2002 and 2012. It is roughly 172,000 tokens in size and its annotation consists of "mixed conversational analysis". The corpus is available for download through the French ORTOLANG repository under CC-BY and is listed in the VLO.

Parliamentary Debates on Europe at the Bundestag. Like their smaller British and French counterparts, the German parliamentary corpus is thematically-focused and contains only those parliamentary debates that correspond to the annual European Council meetings at the German parliament for the period between 1998 and 2005. It is roughly 417,000 tokens in size and its annotation consists of "mixed conversational analysis". The corpus is available for download through the French ORTOLANG repository under CC-BY and is listed in the VLO.

Hellenic Parliament Sittings. This corpus consists of Greek parliamentary proceedings from the period between 2011 and 2015 and is approximately 28.7 million tokens in size. The corpus is available for download under the academically-restricted CC-BY-NC licence from the Greek repository *clarin:el*. It is unclear how the corpus is annotated and it is not listed in the VLO.

Lithuanian Parliament Corpus for Authorship Attribution. This corpus consists of Lithuanian parliamentary proceedings from the period between 1990 and 2013 and is roughly 23.9 million tokens in size. In terms of annotation, the corpus is tokenised, PoS-tagged and lemmatised with *Lemuoklis*, which is a morphological analyzer for lemmatisation, and *MaltParser*, which was used for the generation of dependency tags (Kapočiūtė-Dzikienė, et al. 2015). The corpus is available for download through the CLARIN-LT repository under a non-specific public licence. It is found on the VLO.

Talk of Norway. This corpus is one of the two Norwegian parliamentary corpora. It covers the period between 1998 and 2016 and is 63.8 million tokens in size. It was annotated with the tools *angid.py* and *OBT* and is available for download through the *CLARINO* repository under the public NLOD licence. It is found on the VLO.

Proceedings of Norwegian Parliamentary Debates. This is the other Norwegian corpus. It covers the period between 2008 and 2015, consists of 29 million tokens and displays annotation in relation to the speaker, language variety, political party to which the speaker belongs and date and time. It is only available for online querying through the concordancer *Corpuscule*, also under the NLOD licence.

PTPARL Corpus. This corpus covers Portuguese parliamentary proceedings from the period between 1970 and 2008 and consists of 1 million tokens. It is tokenised, PoS-tagged and lemmatised with *LX-Tokenizer*, *LX-Tagger* (Branco and Silva, 2006), *MBT* and *MBLEM* (Généreux et al., 2012). The VLO entry links to a listing of the corpus in the ELRA catalogue,[5] where the corpus is listed for download under a non-specific academic licence.

SlovParl. This is the Slovene corpus of parliamentary proceedings and there are two versions listed in the VLO – *Slovenian parliamentary corpus SlovParl 1.0* and *Slovenian parliamentary corpus SlovParl 2.0*. Both cover Slovene parliamentary proceedings for the period between 1990 and 1992 and differ from each other in the fact that the newer version is much larger in size – there are parliamentary proceedings from 54 sessions amounting to 2.7 million tokens in *SlovParl 1.0* in comparison with 232 sessions amounting to 10.8 million tokens in *SlovParl 2.0*. Both corpora are available for download under CC-BY in the *CLARIN.SI* repository and are extensively annotated (tokenisation, PoS-tagging and lemmatisation) with additional markup in relation to speaker and session typologies (Pančur and Šorn, 2016).

[4] https://lat.csc.fi/ds/asv/

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

Riksdag's Open Data (Swedish: *Riksdagens öppna data*). This corpus, which in total consists of 1.25 billion tokens and is thus the second largest of the parliamentary corpora, is not listed in the VLO. Rather, it is only listed in the Swedish *Språkbanken* repository and is unique among the corpora in that there is no separate entry for the entire corpus but only for its subcorpora, of which there are 21 in total.[6] Each subcorpus can be downloaded through the repository or – like the Finnish corpus – queried online through *Korp*. The corpus was tokenised, lemmatised, MSD-tagged (including additional markup in relation to semantic features, lemgrams and compounding) with *Sparv*, which is the *Språkbanken*'s corpus annotation pipeline infrastructure (Borin et al., 2016). All the subcorpora are available under CC-BY.

Europarl. This is a multilingual parallel corpus of the sessions of the European Parliament. It contains documents from the period between 1996 and 2011 amounting to 588 million tokens. The corpus is sentence aligned and is freely available for download on a dedicated page under no specific licence. The corpus is listed in the VLO.

2.2 Summary and discussion

Table 1 summarizes the salient characteristics of the parliamentary corpora discussed in the previous subsection.

Table 1: Overview of the parliamentary corpora within the CLARIN infrastructure

NC	Size (mil tok)	Period	Anno	VLO	Avail.	Location of avail.	Licence
uk_1	1,600	1803-2005	T, PoS, L, additional semantic	/	C	External	/
uk_2	0.19	1998-2015	Mixed conversational analysis	✓	D	ORTOLANG	CC-BY
cz	0.5	/	Semi-automatic alignment of transcriptions and audio records	✓	D	LINDAT	CC-BY
dk	7.3	2008-2010	T, PoS, L	✓	D	DK-CLARIN	Non-specific public
ee	13	1995-2001	TEI annotation	✓	D, C	External	Non-specific academic
fi	2.2	2008-2016	/	✓	D, C	FIN-CLARIN	CC-BY
fr	0.17	2002-2012	Mixed conversational analysis	✓	D	ORTOLANG	CC-BY
de	0.4	1998-2015	Mixed conversational analysis	✓	D	ORTOLANG	CC-BY
el	28.7	2011-2015	/	/	D	clarin:el	CC-BY
lt	23.9	1990-2013	T, PoS, L	✓	D	CLARIN-LT	CLARIN-LT public
no_1	63.8	1998-2016	T, PoS, L	✓	D	CLARINO	NLOD
no_2	29	2008-2015	Speaker, date, etc. markup	/	C	CLARINO	NLOD
pt	1	1970-2008	T, PoS, L	✓	D	External	Non-specific academic
si	10.8	1990-1992	T, PoS, L	✓	D, C	CLARIN.SI	CC-BY
se	1,250	1971-2016	T, L, PoS, semantic	/	D, C	SWE-CLARIN	CC-BY
eu	588	1996-2011	Sentence alignment, speaker markup	✓	D	External	/

[6] cf. the resources listed under "Part of the Riksdag's Open data" on https://spraakbanken.gu.se/eng/resources.

The parliamentary corpora are generally well integrated with the CLARIN infrastructure. Only 4 out of the total 16 corpora are not listed in the VLO; that is, the British *Hansard Corpus*, the Greek *Hellenic Parliament Sittings* corpus, the Norwegian *Proceedings of Norwegian Parliamentary Debates* and the Swedish *Riksdag's Open Data* corpus. *Riksdag's Open Data* is especially interesting in this respect since, as discussed in section 3.1, it consists of 21 subcorpora that are listed separately in the Swedish *Språkbanken* repository. Out of the 21 subcorpora, only one is listed in the VLO. This is the *Betänkande* subcorpus, which serves as a collection of summaries of voting results and decisions related to committee meetings.[7] However, its metadata description in the VLO is fairly poor. Consequently, the subcorpus cannot be found with the most straightforward search queries like *parliament** or *parliament* corpus*. The remaining 20 subcorpora are not listed in the VLO and it is not immediately obvious why this is so.

In terms of availability, the majority – that is, 10 corpora (the Czech, Danish, Greek, Lithuanian, Portuguese, French, German corpora, the Norwegian *Talk of Norway* Corpus, the British *Parliamentary Debates on Europe at the House of Commons* corpus and *Europarl*) – are only for download. All of these corpora can be downloaded from a relevant CLARIN repository (e.g. *Czech Parliament Meetings* through LINDAT) except for *Europarl*, which is available on a dedicated webpage, and the Portuguese *PTPARL Corpus*, which is listed in the ELRA catalogue. 4 corpora can both be downloaded and queried through a concordancer. These are the Estonian *Transcripts of Riigikogu* corpus, which is available for download on a dedicated webpage and can be accessed through the *Keeleveeb Query* concordancer provided by CLARIN Estonia; the Finnish *Eduskunta Corpus*, which can be downloaded through the Finnish repository *Language Bank of Finland* and accessed through *Korp*; the Slovene *SlovParl* corpus, which can be downloaded through the *CLARIN.SI* repository and queried online through *noSketchEngine*;[8] and *Riksdag's Open Data*, which can be downloaded from the *Språkbanken* repository and queried through *Korp*. 2 corpora can only be queried online – while the Norwegian parliamentary corpus *Proceedings of Norwegian Parliamentary Debates* is searchable through a concordancer provided by a CLARIN repository (that is, the CLARINO *Corpuscle* concordancer), the other corpus, *The Hansard Corpus*, is queried through a non-CLARIN dedicated concordancer.

The availability and thoroughness of metadata documentation is also fairly good. Information on size is available for all corpora, while information on the temporal period is missing only for *Czech Parliament Meetings*. Information on the level of linguistic annotation is likewise mostly readily available, missing only for the Greek *Hellenic Parliament Sittings* corpus and the Finnish *Eduskunta Corpus*. However, the location of the information on annotation is far from uniform – for instance, the description field of the *Lithuanian Parliament Corpus for Authorship Attribution* on the CLARIN-LT repository does not mention annotation (described only in one of the downloadable corpus files), while the *Riksdag's Open Data* subcorpora are systematically described in *Språkbanken*.

Licence information is also in most cases readily included, with most of the corpora being available under CC-BY. Here we would like to stress that there exists a slight discrepancy between the information as it is presented in the VLO on the one hand and on the relevant landing page on the other in the case of the *PTPARL Corpus*, *Lithuanian Parliament Corpus for Authorship Attribution* and *Talk of Norway*. In the relevant VLO entries for these three corpora, the licence is listed as unknown even though it is specified on the relevant landing pages – for instance, in the case of the *Talk of Norway* corpus, the licence in the CLARINO repository is specified as NLOD, which is public, so the VLO entry should follow suit and also list the corpus as publicly available.

[7] https://repo.spraakbanken.gu.se/xmlui/handle/10794/83
[8] https://www.clarin.si/noske/run.cgi/first_form

2.3 Additional national parliamentary corpora not in the VLO or CLARIN repositories

In our original survey, we identified 7 additional national parliamentary corpora, one corpus per each of the following countries: Austria, Bulgaria, the Czech Republic, the Netherlands, Germany, Latvia and Poland. However, these corpora are neither listed in the VLO nor available through a CLARIN repository, so were omitted from the current discussion. We provide a brief overview of these corpora, as they are generally well presented and would serve as welcome inclusions in relevant CLARIN repositories and the VLO. Note that the bolded names of the corpora below provide hyperlinks to relevant landing pages.

- **Korpusbasierte Analyse österreichischer Parlamentsreden**. Austrian parliamentary corpus for 2013-2015; 1.2 million tokens; tokenised and PoS-tagged (Sippl et al., 2016);
- **Corpus of Bulgarian Political and Journalistic Speech**. Bulgarian parliamentary corpus for 2006-2012; 10 million tokens; tokenised, PoS-tagged, and lemmatised;
- **CzechParl**. Czech parliamentary corpus for 1993-2010; 81.9 million tokens; tokenised, MSD-tagged, and lemmatised (Jakubíček and Kovár, 2010);
- **DutchParl**.[9] Dutch parliamentary corpus for 1814-2014; 800 million tokens; tokenised, PoS-tagged and lemmatised (Marx and Schuth, 2010);
- **polmineR corpus**. German parliamentary corpus; size, period and annotation unknown;
- **SEIMA corpus**. Latvian parliamentary corpus for 1993-2016; unclear size; unclear annotation; and
- **Polish Parliamentary Corpus**.[10] Polish parliamentary corpus for 1991-2017; 300 million tokens; tokenisation, MSD-tagging, lemmatisation, utterance-level segmentation, named entities (Ogrodniczuk, 2012).

3 Identifying parliamentary corpora through the VLO[11]

In this section, we discuss the identification of the 12 VLO corpora both in terms of simple queries and the faceted search option and highlight problematic aspects.

3.1 Simple search

We first focus our discussion on simple search (in other words, using only the search field) on the basis of two salient search strings – *parliament** and *parliament* corpus*.

In the case of the simple search string *parliament**,[12] Table 2 lists the top 20 search results:

[9] The Dutch CLARIN consortium has been involved in important projects on parliamentary data. One such project is *War in Parliament* (http://www.clarin.nl/node/410); another is the *DiLiPaD* project (http://dilipad.history.ac.uk/), which applies Linked Data to British, Dutch and Canadian parliamentary proceedings. Resulting datasets are available through the online PoliticalMashup environment (http://politicalmashup.nl/).

[10] This corpus is set to be included in the CLARIN-PL D-Space repository in June 2018 (Ogrodniczuk, personal correspondence).

[11] The results presented in this section reflect VLO version 4.3.2. from January 2018.

[12] https://vlo.clarin.eu/?q=parliament*

Table 2: Results for the search string *parliament*. The bolded results correspond to a subset of the national parliamentary corpora described in section 2.1.* [13]

#	Name of VLO entry
1	**Plenary Sessions of the Parliament of Finland, Kielipankki Korp Version 1**
2	**Plenary Sessions of the Parliament of Finland, Downloadable Version 1**
3	**Slovenian parliamentary corpus SlovParl 2.0**
4	**Slovenian parliamentary corpus SlovParl 1.0**
5	Plenary Sessions of the Parliament of Finland
6	**Czech Parliament Meetings**
7	**Corpus of the Proceedings of Estonian Parliament**
8	TC-STAR Transcriptions of Spanish Parliamentary Speech
9	European Parliament Interpretation Corpus (EPIC)
10	Information in Sign Language on the Tasks of the Parliamentary Ombudsman of Finland
11	**Lithuanian Parliament Corpus for Authorship Attribution**
12	**Europarl: European Parliament Proceedings Parallel Corpus 1996-2003**
13	**Corpus of the Proceedings of Estonian Parliament**
14	**Parliamentary Debates on Europe at the House of Commons (1998-2015)**
15	Dataset of European Parliament roll-call votes and Twitter activities MEP 1.0
16	**Parliamentary Debates on Europe at the Bundestag (1998-2015)**
17	**Parliamentary Debates on Europe at the Assemblée nationale (2002-2012)**
18	**Plenary Sessions of the Parliament of Finland, Kielipankki LAT Version 1**
19	On the legislative authority of the British Parliament [Electronic resource] / James Wilson
20	The Present State Of Westminster Bridge: Containing A Description [...]

With this search string we are able to find the Finnish corpus (results 1, 2, 18), the Slovene corpus (results 3, 4), the Czech corpus (result 6), the Estonian corpus (results 7, 13), the European parliament corpus (result 12), the Lithuanian corpus (result 11) and the three thematic parliamentary corpora (results 14, 16, 17), so 13 hits for 9 out of the total 12 VLO corpora. The discrepancy between the higher number of VLO entries and smaller number of actual corpora is due to the fact that 3 corpora – that is, the Finnish, Slovenian, Estonian corpora – are associated with more than one VLO entry each. While this is to be expected in the case of the Finnish and Slovene corpus since each result corresponds to a different version, in the case of the Estonian corpus (results 7, 13), the two VLO entries point to exactly the same resource, so the lower-ranked version, which differs from the higher-ranked one only in that it contains a less detailed metadata description, is likely redundant. All in all, 75% of the corpora described in section 2.1 can be found by using the simple search string, but not all.[14]

We now turn to the phrasal search string *parliament* corpus*.[15] As shown below, the relevant results are more scattered in comparison with the previous query and are in several cases ranked below the 20th hit.

[13] After (18), the results of the simple *parliament** query begin corresponding to resources which turn out to be entries for singleton documents like specific letters, or transcriptions of a single speech, so not collections of a series of proceedings and are thus not relevant for our survey.

[14] We believe that the the remaining three corpora (the Danish *DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget*, the Norwegian *Talk of Norway*, and the Portuguese *PTPARL Corpus*) are not yielded by this search string because the name or their metadata description does not contain the term *parliament(ary*).

[15] https://vlo.clarin.eu/?q=parliament*+corpus

Table 3: Results for the search string *parliament* corpus*[16]

#	Name of VLO entry
1	**Lithuanian Parliament Corpus for Authorship Attribution**
2	**PTPARL Corpus**
3	**Slovenian parliamentary corpus SlovParl 2.0**
4	**Slovenian parliamentary corpus SlovParl 1.0**
5	European Parliament Interpretation Corpus (EPIC)
6	Amaryllis Corpus - Evaluation Package
7	KOTUS Finnish-Swedish Parallel Corpus
8	**Europarl: European Parliament Proceedings Parallel Corpus 1996-2003**
9	GeFRePaC - German French Reciprocal Parallel Corpus
10	**Corpus of the Proceedings of Estonian Parliament**
11	**Corpus of the Proceedings of Estonian Parliament**
12	Grenelle II - Subpart 1: audio/video
13	Grenelle II - Subpart 2: audio/video
14	Grenelle II on environnement: multimodal annotation
15	Grenelle II - Subpart 2: audio/video
16	Grenelle II - Subpart 1: audio/video
17	Grenelle II on environnement: multimodal annotation
18	Corpus of Early Modern English Statutes 1491-1707
19	Helsinki Corpus of Swahili 2.0 (HCS 2.0) Annotated Version
20	**Parliamentary Debates on Europe at the House of Commons (1998-2015)**
21	**Parliamentary Debates on Europe at the Bundestag (1998-2015)**
22	**Parliamentary Debates on Europe at the Assemblée nationale (2002-2012)**
23	**Plenary Sessions of the Parliament of Finland, Downloadable Version 1**
24	[...]
25	**Plenary Sessions of the Parliament of Finland, Kielipankki LAT Version 1**
26	[...]
27	**Plenary Sessions of the Parliament of Finland, Kielipankki Korp Version 1**
28	**Czech Parliament Meetings**
29-45	[...]
46	**Talk of Norway**

On a positive note, this complex search string lists two more corpora than the simpler search *parliament**; that is, the Portuguese *PTPARL Corpus* (result 2) and the Norwegian *Talk of Norway* corpus (result 46). However, notice the relative low ranking of the three versions of the Finnish *Eduskunta Corpus* (results 23, 25, 27) and *Czech Parliament Meetings* (result 28) in comparison with the simpler search string in Table 2, in which case they were listed as top results. Why this is so is unclear, as all of these low-ranked entries have rich and granular metadata in which the term *corpus* is both used in the general resource description tab in the VLO and is selected as the value for *resourceType* under the extensive *All metadata* tab.[17] By contrast, resources like the subparts of the *Grenelle* collection (results 12-17) display a poorer metadata description in which the term *corpus* is not used in the general description tab nor is it specified as the value for *resourceType*. Such results would be expected to be ranked lower than the three versions of the Finnish corpus and *Czech Parliament Meetings*. It is also unclear why the *LAT* version of the *Eduskunta Corpus* (result 25) is ranked higher than the *Korp* version (result 27), since it is, as described in section 2.1, still under development and therefore its VLO entry lacks a detailed metadata description in comparison with that of the *Korp* version. Additionally, *Talk of Norway* comes up as a very low-ranked result (46), so

[16] The numbering in the table below again corresponds to the ranking in the VLO and, for reasons of space, we omit the irrelevant results beyond (20).

[17] As of version 4.3.2. of the VLO, terms in phrasal search strings are conjoined by the AND operator, so a search string like *parliament* corpus* should provide an intersection of resources that pertain to the *parliament** search string and those resources that pertain to the *corpus* search string.

it is unlikely that a potential researcher would find easily. The fairly scattered presentation of the results in Table 3 is counterintuitive for anyone interested in finding parliamentary corpora with such a straightforward string as *parliament* corpus*.

3.2 Faceted search

In the VLO, resources can also be found by means of faceted search, which allows the user to narrow down a general query by applying filters under various facets such as Language, Resource Type, Genre, Modality, Subject and so forth. Following Odijk (2014), we focus on two facets that seem to be the most relevant for narrowing our search down to relevant parliamentary corpora. These are Resource Type, which should in the case of the simple search string *parliament** give us the option to select *corpus* as a value, and Subject, which should presumably narrow the search down to subjects related to parliamentary data.

In the case of the simple search string *parliament**, the Resource Type facet returns the following list of values, with the number of resources for each value listed in the parentheses: *Text* (51), *Sound* (37), Info*:eu-repo/semantics/dataset* (3), *Bioscoop* (2), *Corpus* (2), *Politics* (2), *Boek* (1), *Dataset* (1), and so forth. Surprisingly, the value *Corpus* lists only two resources: *Czech Parliament Meetings* and *Europarl: European Parliament Proceedings Parallel Corpus 1996-2003*. In other words, the vast majority of the corpora we can identify with the simple query *parliament** (cf. Table 2) are not captured in this facet and it is unclear why this is so, especially since a resource like *Plenary Sessions of the Parliament of Finland, Downloadable Version 1* has the value *corpus* under *resourceType* in the metadata description. Similarly, recall from section 2.1 that *Czech Parliament Meetings* is a corpus of audio records, yet selecting the value *Sound* fails to list it. In short, narrowing the search down through Resource Type does not yield the desired results.

Sticking to the same simple search string, the Subject facet presents the following list of values: *text_and_corpus linguistics* (37), *corpus* (6), *audio* (4), *débat politique* (4), *political debate* (4), *video* (4), *vidéo* (4), *discours politique* (3), *débats parlementaires* (3), *europe* (3) and so forth. On the one hand, values for the same type of subject are given twice (e.g. *discours politique* and *débat politique*) and selecting one value filters the results of the other. On the other hand, several values are clearly more suitable for the Resource Type facet, yet selecting for instance *corpus* (6) does not yield any of the parliamentary corpora in Tables 2 and 3, nor does selecting *audio* (4) yield *Czech Parliament Meetings*, contrary to expectations. Our findings correspond with Odijk's experience (2014).

4 Discussion and proposals

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 listed in section 2.3 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.

However, a bigger limitation that needs immediate attention seems to be the VLO. We have shown that identifying parliamentary corpora in the current version of VLO is counterintuitive, since the best results (Table 2) are yielded by the simplest search string, whereas further specifications either by a narrower search string (Table 3) or the use of faceted search yields in substantially lower precision as well as recall. Additionally, while some corpora like *Talk of Norway* or the Danish *DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget* are listed in the VLO, they can only be identified after querying the full corpus name, which we believe is a limitation since many users will not know the official name of the resource they are looking for. On the other hand, a handful of highly relevant corpora, like the Swedish *Riksdag's Open Data* and the British *Hansard Corpus*, are not listed in the VLO at all, only in the national repositories, and we believe their inclusion would be beneficial for the comprehensive representation of CLARIN parliamentary corpora in the VLO.

5 Conclusion

In this paper we presented a survey of the parliamentary corpora in the CLARIN infrastructure. We have been able to find 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 easily be found through the Virtual Language Observatory. For this reason, we have drawn up a list of recommendations for corpus metadata in order to improve findability and ranking of the corpora by VLO as well as documented issues with the VLO that should be taken into account in future development of the service. This is of paramount importance as the VLO is the main gateway to the invaluable CLARIN resources.

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 common 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. With this in mind, we will also collect showcases of successful applications of parliamentary corpora in Digital Humanities and Social Sciences, as such information 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.

[Borin et al. 2016] Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. http://www8.cs.umu.se/~johanna/sltc2016/abstracts/SLTC_2016_paper_31.pdf. Last accessed on 11 January 2018.

[Branco and Silva 2006] António Branco and João Silva. 2006. A Suite of Shallow Processing Tools for Portuguese: LX-Suite, In *Proceedings of EACL2006 – 11th Conference of the European Chapter of the Association for Computational Linguistics,* 179–182.

[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.

[Généreux et al. 2012] Michel Généreux, Iris Hendrickx, Amália Mendes. 2012. "A Large Portuguese Corpus On-Line: Cleaning and Preprocessing." *Conference: Computational Processing of the Portuguese Language (PROPOR)*.

[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.

[Jakubíček and Kovář 2010] Miloš Jakubíček, Vojtěch Kovář. 2010. "CzechParl: Corpus of Stenographic Protocols from Czech Parliament". In P. Sojka, A. Horák (eds.) *RASLAN 2010 Recent Advances in Slavonic Natural Language Processing*.

[Kapočiūtė-Dzikienė et al. 2015] Jurgita Kapočiūtė-Dzikienė, Andrius Utka, Ligita Šarkutė. 2015. "Authorship attribution of internet comments with thousand candidate authors." *ICIST 2015 : 21st International Conference on Information and Software Technologies*, 433-448. Springer International Publishing.

[Mandravickaitė and Krilavičius 2015] Justina Mandravickaitė, Tomas Krilavičius. 2015. Language usage of members of the Lithuanian Parliament considering their political orientation. *Deeds and Days* 64: 133-151.

[Meurer 2017] Paul Meurer. 2017. From LFG structures to dependency *relations. Bergen Language and Linguistic Studies* 8: 183-201.

[Marx and Schuth 2010] Maarten Marx and Anne Schuth. "DutchParl: The Parliamentary Documents in Dutch." http://politicalmashup.nl/new/uploads/2010/03/lrecfinalversionlong.pdf. Last accessed on 7 January 2018.

[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/

[Odijk 2014] Jan Odijk. 2014. "Discovering Resources in CLARIN: Problems and Suggestions for Solutions." http://www.clarin.nl/sites/default/files/Searching%20with%20the%20VLO.pdf. Last accessed on 11 January 2017.

[Ogrodniczuk 2012] Maciej Ogrodniczuk. 2012. "The Polish Sejm Corpus." http://www.lrec-conf.org/proceedings/lrec2012/pdf/653_Paper.pdf. Last accessed on 8 January 2018.

[Oravect et al. 2014] Csaba Oravecz, Tamás Váradi, Bálint Sass. 2014. "The Hungarian Gigaword Corpus." http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf. Last accessed on 10 January 2018.

[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/.

[Rayson et al. 2015] Paul Rayson, Alistair Baron, Scott Piao, Steve Wattam. 2015. "Large-scale Time-sensitive Semantic Analysis of Historical Corpora." http://ucrel.lancs.ac.uk/samuels/papers/SAMUELS_ICAME36_Software_Demo_Handout.pdf. Last accessed on 7 January 2018.

[Sippl et al. 2016] Colin Sippl, Manuel Burghardt, Christian Wolff, Bettina Mielke. 2016. "Korpusbasierte Analyse österreichischer Parlamentsreden."