Title	Overview of Parliamentary Data and Corpora
Version	3.5
Author(s)	DF, JL
Date	20-11-2017
Status	For distribution
Distribution	NCF, UI
ID	CE-2017-1019

Contents

1 Background & Motivation

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), and the Swedish corpus has been used for the analysis of governmental policies related to Swedish film (Norén and Snickars 2016). Additionally, there have been successful CLARIN projects involving parliamentary data such as the *Talk of Europe* project; one of its results was the creation of a Linked Open Dataset on the basis of the proceedings of the European parliament.

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

In the following survey, our aim is to provide an overview of 1) parliamentary records and 2) corpora all the countries that are members or observers of CLARIN ERIC. Our motivation was to identify to what extent these resources exist and are easily available, and check which information about the resources is available, thereby highlighting the aspects in which accessibility of these corpora as well as the presentation of the relevant information can be optimised from a User Involvement perspective.

2 Approach

We took into account all 19 CLARIN member countries and observer countries: Austria, Bulgaria, the Czech Republic, Denmark, Estonia, Finland, Germany, Greece, Italy, Latvia, Lithuania, the Netherlands, Norway, Sweden, Poland, Portugal, Slovenia, the UK, France, and Hungary. We only took into account parliamentary records, not other legislative documents or public political speeches.

Section 3 gives an overview of the raw parliamentary records, the periods they cover and the text formats that they are available in. Section 4 presents a list of corpora that were identified in three steps: i) through the VLO, ii) on the repositories or websites of the national consortia, and iii) through the national UI coordinators. We also provide the key metadata (annotation, tools used for annotation, the temporal scope of the records) and key papers associated with the corpora.

3 Parliamentary records

We were able to find parliamentary records for all the countries except for Poland.[2] We primarily focussed on records that are available as transcribed text. In all cases, such parliamentary records are freely available on the relevant parliamentary websites.

Table 1: Overview of the parliamentary records available as transcribed text with respect to period and format

NC	Period	Text formats
Austria	1920–	.pdf and .html
Bulgaria	2001–	.pdf and .xls
The Czech Republic	2013–	.html
Denmark	2000–	.pdf and .html
Estonia	1998–2007	.pdf
France	2013–	.xml
Finland	X	.pdf and .html
Germany	2013–	.pdf
Greece	1990s–	.pdf and .docx
Italy	1996–	.pdf
Latvia	1920s–	.pdf and .html
Lithuania	1990s–	.pdf and .docx
The Netherlands	1814–1995 1995–	.pdf and .html .pdf, .odt and .xml
Norway	1814–	.html
Sweden	1917–	.pdf and .html

[2] The link to the Polish parliamentary data was subsequently provided by the author of the Polish parliamentary corpus.

Poland	2015–	.pdf
Portugal	1821–	.pdf
Slovenia	1990–	.html
the UK	1807–	.pdf and .html
Hungary	1990–	.html

The records differ from one another in two respects:

1 **Time period:** The periods that they cover vary widely from country to country. The oldest records are those of the UK (from 1807 on), Norway (from 1814 on) and the Netherlands (1814-1995). The contemporary records of the Netherlands are offered on a separate website, where they can be downloaded in the .pdf, .odt and .xml formats.[3]

2 **Data format:** The records are offered in various formats; in the majority of cases (15 out of 19 countries with accessible records), the records are available as .pdf files (in most of these cases other options are available as well), whereas 4 countries (The Czech Republic, Norway, Slovenia and Hungary) have the records available solely as plain .html files.

4 Corpora of parliamentary records

In total, we found 21 corpora of national parliamentary data from all CLARIN countries except Italy. For Norway and the Czech Republic we found two parliamentary corpora. We also include Europarl, containing the proceedings of the European Parliament. 17 corpora, which are described in section 4.1, are part of the CLARIN infrastructure. The data are collected in a Google Doc spreadsheet file.

4.1 Corpora within the CLARIN infrastructure

In Table 2 we describe those corpora that are part of CLARIN infrastructure on the on the basis of the following information:

- size of the corpus and the period it covers
- the type of annotation included and the tools used for the annotation
- the availability (in online environments, for download)
- related publication, if available

The hyperlinks were last accessed 15 November 2017.

Table 2: Overview of the parliamentary corpora that are part of the CLARIN infrastructure

Corpus name	Corpus description
Czech Parliamentary Meetings Czech 0.5 million tokens For download	The corpus consists of 88 hours of speech data, which corresponds roughly to 0.5 million tokens. This corpus is available for download through the CZECH repository LINDAT under the CC-BY licence.
DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget Danish 7.3 million tokens	The corpus covers Danish parliamentary data from 2008 to 2010 and consists of 7.3 million tokens. It is tokenised, PoS-tagged and lemmatised. The tool that was used for annotation is *ePOS-DSL*. It is available for download on the

[3] https://zoek.officielebekendmakingen.nl/zoeken/parlementaire_documenten.

Tokenised, PoS-tagged, lemmatised For download	webpage of DK-CLARIN under a public licence.
Hansard corpus English 1.6 billion tokens Tokenised, PoS-tagged, lemmatised, semantic tagging Concordancer	The corpus covers the British parliamentary data from 1803 to 2005 and consists of approx. 1.6 billion tokens. In addition to being tokenised, PoS-tagged, lemmatised, the corpus is also characterized by semantic tagging. The relevant tools for the semantic tagging are the *USAS semantic tagger* and the *Historical Thesaurus Semantic Tagger (HTST).* The corpus is available through an interface on the corpus webpage, the licence is unclear. For the relevant publication, see Rayson et al. (2015).
Parliamentary Debates on Europe at the House of Commons (1998-2015) English 190,000 tokens Annotation of conversation Download	The corpus covers British parliamentary data from 1998 to 2015. The corpus consists of 190,000 tokens. The documentation states that the corpus displays "annotation of conversation". The corpus is available for download in the ORTOLANG repository under CC-BY.
Parliamentary Debates on Europe at the Assemblée nationale (2002-2012) French Unclear size Unclear annotation Download	The corpus covers French parliamentary data from 2002 to 2005. The size is unclear. It is unclear how the corpus is annotated. The corpus is available for download in the ORTOLANG repository under CC-BY.
Transcripts of Riigikogu (Estonian Parliament) Estonian 13 million tokens Unclear annotation For download and concordancer	The corpus covers Estonian parliamentary data from 1995 to 2001 and consists of approx. 13 million tokens. We were unable to ascertain how the corpus is annotated (apart from the fact that the speakers are tagged), nor with which tools the annotation was carried out. It is available for download on the corpus webpage and is also accessible through a concordancer on the same webpage. The licence is CLARIN_ACA.
Plenary Sessions of the Parliament of Finland Finnish 22.4 million tokens Unclear annotation Concordancer	The corpus covers Finnish parliamentary data from 2008 to 2016 and consists of 22.4 million tokens. We were unable to discern how the corpus is annotated, nor with which tools the annotation was carried out. It is available through the concordancer *KORP* under the CC-BY licence.
Parliamentary Debates on Europe at the Bundestag (1998-2015) German 417,000 tokens Annotation of conversation For download	The corpus covers German parliamentary data from 1998 to 2015. The corpus consists of 417,095 tokens. The documentation states that the corpus displays "annotation of conversation". The corpus is available for download in the ORTOLANG repository under CC-BY.
Hellenic Parliament Sittings	The corpus covers the Greek parliamentary data from 2011 to 2015

(2011-2015) Greek 28.7 million tokens Unclear annotation For download	and consists of 28.7 million tokens. It is unclear how the corpus is annotated. It is available for download on the CLARIN:EL repository under CC-BY.
Hungarian National Corpus Hungarian 190 million tokens MSD-tagging Concordancer	The parliamentary data, which are a subset of the corpus, consist of approx. 190 million tokens. We were not able to ascertain which period the parliamentary data cover. In terms of annotation, the corpus displays automatic annotation of stems, part of speech and inflectional information. The tools used are Humor (morphologic tagger); TnT tagger (disambiguation); IMS Open Corpus Workbench (Corpus Query Tool). The corpus is accessible through an interface on the corpus webpage; however, access requires registration. For the relevant publication, see Oravecz et al. (2014)
Lithuanian Parliament Corpus for Authorship Attribution Lithuanian 23.9 million tokens Tokenised, PoS-tagged, lemmatised For download	The corpus covers Lithuanian parliamentary data from 1990 to 2013 and consists of 23.9 million tokens. It is tokenised, PoS-tagged and lemmatised. It was annotated with the following tools: Lemuoklis (morphological analyzer for lemmatization); MaltParser (generation of dependency tags). It is available for download on the corpus webpage, the licence is unknown.
Talk of Norway Norwegian 63.8 million tokens Tokenised, PoS-tagged, lemmatised For download	The corpus covers Norwegian parliamentary data from 1998 to 2016 and consists of 63.8 million tokens. It is tokenised, PoS-tagged and lemmatised. The annotation was done with the tools *angid.py* and *OBT*. It is available for download through CLARINO under the public NLOD licence.
Proceedings of Norwegian Parliamentary Debates Norwegian 29 million tokens Tokenised Concordancer	The corpus covers Norwegian parliamentary data from 2008 to 2015 and consists of 29 million tokens. The corpus is tokenised and is accessible through the concordancer *Corpuscle* through CLARINO under the public NLOD licence.
PTPARL Corpus Portuguese 1 million tokens Tokenised, PoS-tagged, lemmatised Unavailable	The corpus covers Portuguese parliamentary data from 1970 to 2008 and consists of 1 million tokens. It is tokenised, PoS-tagged and lemmatised. The relevant tools are LX-Tokenizer, LX-Tagger, MBT, MBLEM (lemmatisation). The corpus is unavailable and the licence is unknown. For the relevant publication, see Généreux et al. (2012).
SlovParl Slovenian 3.2 million tokens Tokenised, PoS-tagged, lemmatised For download and concordancer	The corpus covers Slovene parliamentary data from 1990 to 1992 and consists of 3.2 million tokens. It is tokenised, PoS-tagged and lemmatised with ToTrTaLe. It is available both for download through CLARIN.SI and through the noSketch Engine under CC-BY. For the relevant publication, see Pančur (2016).

Riksdag's Open data Swedish 1.25 billion tokens Tokenised, lemmatised For download and Concordancer	The corpus covers parliamentary data from the Swedish parliament Riksdag from 1971 to 2016 and consists of 1.25 billion tokens. It is tokenised and lemmatised with *Sparv*, which is the Språkbanken's corpus annotation pipeline infrastructure. The data can be downloaded through Språkbanken and are available through the concordancer *Korp* under CC-BY. For the relevant publication, see Borin et al. (2016).
Europarl 21 languages Tokenised For download	The corpus covers data from the European Parliament from 1996 to 2011 and consists of 588 million tokens. The corpus tokenised. The corpus is available for download on the dedicated webpage, while the licence is unclear. For the relevant publication, see Koehn (2005). Additionally, proceedings of the European parliament have been made available as a dataset enriched with Linked Open Data under the *Talk of Europe* project. This dataset covers the period from 1999 to 2014. This dataset can either be downloaded through DANS or accessed online through a SPARQL endpoint. However, it seems that this dataset cannot be found on the VLO or in a national repository.

4.1.1 Identification

In total, 17 parliamentary corpora are part of the CLARIN infrastructure. All of them were identified via the VLO except for the following 6 corpora:

(i) *Hansard corpus*
(ii) *Parliamentary Debates on Europe at the House of Commons (1998-2015)*
(iii) *Parliamentary Debates on Europe at the Bundestag (1998-2015)*
(iv) *Hellenic Parliament Sittings (2011-2015)*
(v) *Talk of Norway*
(vi) *Riksdag's Open data*

Corpus (i) was identified on the website of the British observer. Corpora (ii) and (iii) were identified through ORTOLANG. Corpus (iv) was identified through the repository of CLARIN:el, corpus (v) through the repository of CLARINO, and corpus (vi) through the repository of the Swedish consortium (i.e. Språkbanken).

4.1.2 Availability

The following 3 corpora are available both for download and through a concordancer:

(i) *Transcripts of Riigikogu (Estonian Parliament)*
(ii) *SlovParl*
(iii) *Riksdag's Open data*

The concordancer for corpus (i) is a dedicated one, the concordancer for (ii) is *noSketchEngine* and the concordancer for (iii) is *Korp*, which is provided by Språkbanken.

The following 9 corpora are available only for download:

(i) *Czech Parliamentary Meetings*, through LINDAT
(ii) *DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget*, through DK-CLARIN
(iii) *Parliamentary Debates on Europe at the Assemblée nationale (2002-2012)*, through ORTOLANG
(iv) *Parliamentary Debates on Europe at the House of Commons (1998-2015)*, through ORTOLANG
(v) *Parliamentary Debates on Europe at the Bundestag (1998-2015)*, through ORTOLANG
(vi) *Hellenic Parliament Sittings (2011-2015*), through CLARIN:el
(vii) *Talk of Norway*, through CLARINO
(viii) *Lithuanian Parliament Corpus for Authorship Attribution*, through CLARIN-LT
(ix) *Europarl*, through a dedicated webpage

The following 4 corpora are available only through a concordancer:

(i) *Hansard corpus*, where the concordancer is a dedicated one
(ii) *Plenary Sessions of the Parliament of Finland*, where the concordancer is *Korp* (Finnish distribution)
(iii) *Hungarian National Corpus*, where the concordancer is a dedicated one
(iv) *Proceedings of Norwegian Parliamentary Debates*, where the concordancer is Corpuscle (provided by CLARINO)

Finally, *PTPARL Corpus* is only listed in the VLO, but is unavailable.

4.1.3 Metadata issues

We identified three issues concerning metadata:

- It is unclear how the following 4 corpora are annotated: *Transcripts of Riigikogu (Estonian Parliament), Plenary Sessions of the Parliament of Finland*, *Parliamentary Debates on Europe at the Assemblée nationale (2002-2012)* and *Hellenic Parliament Sittings (2011-2015).* Furthermore, the size of *Parliamentary Debates on Europe at the Assemblée nationale (2002-2012)* is unknown. Otherwise, the majority of the corpora is tokenised, lemmatised and PoS-tagged, with the key exception being *Europarl* and *Proceedings of Norwegian Parliamentary Debates*, which appear to be only tokenised.
- The period that the parliamentary subset of the *Hungarian National Corpus* covers is unclear. Otherwise, the longest period is covered by *Hansard corpus* (1803-2005)
- The licence is unknown for the following 4 corpora: *Hansard corpus*, *Lithuanian Parliament Corpus for Authorship Attribution*, *Europarl,* and *PTPARL Corpus*. Otherwise, 8 corpora are available under CC-BY, 2 are available under the NLOD licence, 1 is available under an undefined public licence, 1 is available under CLARIN_ACA, and 1 corpus is available under an undefined restricted licence.

4.2 Corpora not part of the CLARIN infrastructure

In Table 3 we describe those corpora that are *not* part of CLARIN infrastructure on the on the basis of the following information:

- size of the corpus and the period it covers
- the type of annotation included and the tools used for the annotation
- the availability (in online environments, for download)
- related publication, if available

Table 3: Overview of the corpora that are not part of the CLARIN infrastructure

Country	Corpus description
Korpusbasierte Analyse österreichischer Parlamentsreden Austrian 1.2 million tokens Tokenised, PoS-tagged For download	The corpus covers Austrian parliamentary data from 2013 to 2015 and consists of 1.2 million tokens. It is PoS-tagged with the Stanford Tagger; however, it is not lemmatized. It is available for download on the corpus webpage. The link to the resource was provided by the Austrian UI coordinator. For the relevant publication, see Sippl et al. (2016).
Corpus of Bulgarian Political and Journalistic Speech Bulgarian 10 million tokens Tokenised, PoS-tagged, lemmatised Concordancer	The corpus covers Bulgarian parliamentary data from 2006 to 2012 and consists of 10 million tokens. It is tokenised, PoS-tagged, and lemmatised. We were unable to identify the tools used to annotate the corpus. It is accessible through the concordancer on the corpus webpage. The link to the resource was provided by the Bulgarian UI coordinator.
CzechParl Czech 81.9 million tokens Tokenised, MSD-tagged and lemmatised Concordancer	The corpus covers Czech parliamentary data from 1993 to 2010 and consists of 81.9 million tokens. It is tokenised, MSD-tagged and lemmatised. The tool that was used for annotation is *majka*. It is accessible through *Sketch Engine*. The link to the resource was provided by the CZ UI coordinator. For the relevant publication, see Jakubíček and Kovář (2010).
DutchParl Dutch 800 million tokens Tokenised, PoS-tagged, lemmatised For download and concordancer	The corpus covers Dutch parliamentary data from 1814 to 2014 and consists of 800 million tokens and. It is tokenised, PoS-tagged and lemmatised. The relevant tool is Frog, which is an advanced Natural Language Processing suite for Dutch. The corpus is available for download (the authors needs to be contacted) and is also accessible online through the Political Mashup environment. The link to the resource was provided by the Dutch UI coordinator. For the relevant publication, see Marx and Schuth (2010).
polmineR corpus German Unclear size Unclear annotation For download (small sample)	For Germany, we are only aware that there exists a corpus that is used for the development of the polmineR tool. A small sample can be downloaded from the GitHub webpage of the tool.
SEIMA corpus Latvian Unclear size Unclear annotation Concordancer	The corpus covers Latvian parliamentary data from 1993 to 2016. The size is unclear We were not able to find information regarding the size of the corpus and its annotation. It can be accessed through *noSketchEngine*. The link to the corpus was provided by the Latvian UI coordinator.
Polish Parliamentary Corpus Polish	The corpus covers Polish parliamentary data from 1991 to 2017 and consists of 300 million tokens.

300 million tokens Tokenised, MSD-tagged, named entities, etc. For download and concordancer	Apart from tokenisation and lemmatisation, it is characterized by utterance-level segmentation, disambiguated morphosyntactic description and tagging of syntactic words, syntactic groups and named entities. The relevant tools are *Morfeusz SGJP* (morphological analyser), *Pantera* (disambiguating tagger), *Spejd* (shallow parser), *Nerf* (named entity recognizer). The corpus is both available for download on the corpus webpage and accessible through the NKJP concordancer. The corpus could not be found through the VLO or on the webpage of the Polish consortium; the link to the resource was provided by the Polish UI coordinator. For the relevant publication, see Ogrodniczuk (2012).

4.2.1 Identification

All of the 7 corpora in Table 3 were identified with the help of national CLARIN coordinators and UI representatives except for the German corpus, which was found on Google.

4.2.2 Availability

The following 2 corpora are available both for download and through a concordancer:

(i) *DutchParl*
(ii) *Polish Parliamentary Corpus*
(iii) *Riksdag's Open data*

Corpus (i) is available through the Political Mashup environment, whereas corpus (ii) is available through the concordancer NKJP.

The following 3 corpora are available only through a concordancer:

(i) *Corpus of Bulgarian Political and Journalistic Speech*, through a dedicated concordancer
(ii) *CzechParl*, through *noSketchEngine*
(iii) *SEIMA corpus*, through *noSketchEngine*

Finally, the *Korpusbasierte Analyse österreichischer Parlamentsreden* corpus and a sample of *polmineR corpus* are available for download

4.2.3 Metadata issues

We identified two issues concerning metadata:

- The size as well as type of annotation of the following two corpora are unclear: *polmineR corpus* and *SEIMA corpus*
- The licence is unknown for all the corpora except for *Polish Parliamentary Corpus*, which is available under CC-BY.

5 Recommendations

As parliamentary corpora are of tremendous value for researchers from a wide range of research disciplines, we propose the following:

- create a virtual collection pointing to a landing page (ideally with a PID) for each parliamentary corpus;
- add the missing corpora to the VLO;
- 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* or *parliamentary* 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 multiple languages (at least English and the 'local' language) that include one of these words or an equivalent term, which will lead to higher ranking
- use have 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 (this helps for the ranking, especially if the level is/maps to PUB or ACA)
- provide relevant values that map to the *keyword* or maybe *subject* facets (we recommend using the same specific keyword *parliamentary records* so that all records can be retrieved with a single query on basis of that keyword)
- include information on corpus size, period, annotations etc.