Title	Overview of Parliamentary Data and Corpora
Version	2
Author(s)	DF, JL
Date	09-04-2017
Status	Draft
Distribution	NCF, BoD, SAB
ID	CE-2017-1019

1 Background & Motivation

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 and Hungary.[1] 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 and Estonia.[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	X	X

[1] France was not included in this task because France did not yet have an appointed UI representative in February 2017, when the research was carried out.

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

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
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 (14 out of 18 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 20 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. The data are collected in a Google Doc spreadsheet file. In Table 2 we describe each corpus 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 2: Overview of the parliamentary corpora

Country	Corpus description
Austria	The Austrian parliamentary corpus is *Korpusbasierte Analyse österreichischer Parlamentsreden*. It consists of 1,225,109 tokens and compiles Austrian parliamentary data from 2013 to 2015. It is PoS-tagged with the Stanford Tagger; however, it is not lemmatized. It is available for download on the corpus webpage. The corpus could not be found through the VLO or on the webpage of the Austrian

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

	consortium; the link to the resource was provided by the Austrian UI coordinator. For the relevant publication, see Sippl et al. (2016).
Bulgaria	The Bulgarian parliamentary corpus is *Corpus of Bulgarian Political and Journalistic Speech*. It consists of cca. 10,000,000 tokens. It compiles Bulgarian parliamentary data from 2006 to 2012. 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 corpus could not be found through the VLO or on the webpage of the Bulgarian consortium; the link to the resource was provided by the Bulgarian UI coordinator. We were unable to find a publication relevant for this corpus.
The Czech Republic	The first Czech corpus is *CzechParl*. It consists of 81,874,122 tokens. It compiles Czech parliamentary data from 1993 to 2010. It is tokenised, MSD-tagged and lemmatised. The tools that were used for annotation are *desamb tagger* and *majka*. It is accessible through the Sketch Engine. The corpus could not be found through the VLO or on the webpage of the Czech consortium; the link to the resource was provided by the CZ UI coordinator. For the relevant publication, see Jakubíček and Kovář (2010).
The Czech Republic	The second Czech corpus is *Czech Parliamentary Meetings*. However, since the corpus compiles audio recordings, we did not research it in terms of its annotation and size. This corpus is only available for download. This corpus can be found both through the VLO and on the webpage of the Czech consortium LINDAT.
Denmark	The Danish parliamentary corpus is *DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinget*. It consists of 7,352,818 tokens. It compiles Danish parliamentary data from 2008 to 2010. It is tokenised, PoS-tagged and lemmatised. The tool that was used for annotation is *ePOS-DSL*. It is available for download on the webpage of DK-CLARIN. The corpus can be found both through the VLO and on the webpage of the Danish consortium. We were unable to find a publication relevant for this corpus.
Estonia	The Estonian parliamentary corpus is *Reference corpus of Estonian: Transcripts of Riigikogu (Estonian Parliament)*. It consists of cca. 13 million tokens and compiles Estonian parliamentary data from 1995 to 2001. 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 link to the corpus was provided by the Estonian UI coordinator, though the corpus can be found through the VLO if its name is used as the search keyword. We were unable to find a publication relevant for this corpus.

Finland	The Finnish parliamentary corpus is the *Eduskunta* corpus. It consists of 22,458,581 tokens and compiles Finnish parliamentary data from 2008 to 2016. 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*. The link to the corpus was provided by the Finnish UI coordinator, though the corpus can be found on the FIN-CLARIN webpage. We were unable to find a publication relevant for this corpus.
Germany	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.
Greece	The Greek parliamentary corpus is the corpus *Hellenic Parliament Sittings (2011-2015)*. It consists of 28,699,636 tokens and compiles the Greek parliamentary data from 2011 to 2015. It is unclear how the corpus is annotated. It is available for download on the CLARIN:EL repository. The link to this corpus was provided by one of the developers. We were unable to find a publication relevant for this corpus.
Italy	We were unable to find a parliamentary corpus for Italy.
Latvia	The Latvian parliamentary corpus is the *SEIMA* corpus. It compiles Latvian parliamentary data from 1993 to 2016. We were not able to find information regarding the size of the corpus and its annotation. It can be accessed through a concordancer on the corpus webpage. The link to the corpus was provided by the Latvian UI coordinator. We were unable to find a publication relevant for this corpus.
Lithuania	The Lithuanian corpus was developed in the framework of Project ASTRA (LIT-8-69). It consists of 23,908,302 tokens and compiles Lithuanian parliamentary data from 1990 to 2013. 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 corpus could not be found through the VLO or on the webpage of the Lithuanian consortium; the link to the resource was provided by the Lithuanian UI coordinator. For the relevant publication, see Kapočiūtė-Dzikienė et al. (2015).
The Netherlands	The Dutch corpus is *DutchParl*. It consists of cca. 800,000,000 tokens and compiles Dutch parliamentary data from 1814 to 2014. 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 and is also accessible online through the Political Mashup environment. The corpus could not be found through the VLO or on the webpage of the Dutch consortium; the link to the resource was provided by the Dutch UI coordinator. For the relevant publication, see Marx and Schuth (2010).

Norway	The Norwegian corpus is the *Talk of Norway*. It consists of 63,803,593 tokens and compiles Norwegian parliamentary data from 1998 to 2016. It is tokenised, PoS-tagged and lemmatised. The annotation was done with the tools *angid.py* and *OBT*. It is available for download on the corpus webpage. The corpus could not be found through the VLO or on the webpage of the Norwegian consortium; the link to the resource was provided by the Norwegian UI coordinator. We were not able to find a publication relevant for this corpus.
Norway	The second Norwegian corpus is *Proceedings of Norwegian Parliamentary Debates*. It consists of cca. 29,000,000 tokens and covers Norwegian parliamentary data from 2008 to 2015. The corpus is tokenised and is accessible through the concordancer *Corpuscle* on the corpus webpage. The corpus was found through the VLO. We were not able to find a publication relevant for this corpus.
Sweden	The Swedish corpus is a compilation of parliamentary data from the Riksdag available through the concordancer KORP. It consists of 1.25 billion tokens, covering the period from 1971 to 2016. It is tokenised and lemmatised with *Sparv*, which is the Språkbanken's corpus annotation pipeline infrastructure. The data can be downloaded here. The corpus could not be found through the VLO or on the webpage of the Swedish consortium; the link to the resource was provided by the Swedish UI and NC coordinators. For the relevant publication, see Borin et al. (2016)
Poland	The Polish corpus is *The Polish Sejm Corpus*.[4] It consists of cca. 114,000,000 tokens and compiles the parliamentary data from 1991 to 2017. 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).
Portugal	The Portuguese corpus is the *PTPARL Corpus*. It consists of 1,000,441 tokens and compiles Portuguese parliamentary data from 1970 to 2008. It is tokenised, PoS-tagged and lemmatised. The relevant tools are LX-Tokenizer, LX-Tagger, MBT, MBLEM (lemmatisation). The corpus is available for download on the corpus webpage. The corpus can be found through the VLO.

[4] A newer version, which is called *The Polish Parliamenty Corpus* and which also contains the data from the *Sejm* corpus, is in development.

	For the relevant publication, see Généreux et al. (2012).
Slovenia	The Slovene corpus is *SlovParl*. It consists of cca. 3,200,000 tokens and compiles parliamentary data from 1990 to 1992. It is tokenised, PoS-tagged and lemmatised with ToTrTaLe. It is available both for download and through the noSketch Engine. The corpus can be found through the VLO. For the relevant publication, see Pančur (2016).
Hungary	The Hungarian corpus is the *Hungarian National Corpus*, of which the parliamentary data are a subset. The parliamentary data consist of cca. 20,900,000 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. The corpus could not be found through the VLO or on the webpage of the Hungarian consortium; the link to the resource was provided by the Hungarian UI coordinator. For the relevant publication, see Oravecz et al. (2014).
The UK	The British corpus is the *Hansard corpus*. It consists of cca. 1.6 billion tokens and compiles the British parliamentary data from 1803 to 2005. 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. For the relevant publication, see Rayson et al. (2015).
Europarl	The Europarl corpus covers data from the European Parliament from 1996 to 2011. The corpus is not tokenised. The corpus is available for download. We found the corpus on the webpage of the Czech consortium. For the relevant publication, see Koehn (2005).

7 corpora appear to be available for download only. These are the Austrian, Danish, German, Portuguese and Latvian corpora, as well as the *Czech Parliament Meetings* and *Talk of Norway* corpora. 7 corpora are accessible through online search environments – these are *KORP* for the Finnish corpus, the *Sketch Engine* for *CzechParl*, the *noSketch Engine* for the Latvian corpus, and *Corpuscule* for the *Proceedings of Norwegian Parliamentary Debates* corpus, CLaRK for the Bulgarian corpus and the HNC for the Hungarian one. 5 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.

Finding the relevant corpora is not an easy task. Very few were found through the VLO through basic keyword searchers (i.e. using *parliament* or *parliamentary* as keywords) – only the Estonian,

Slovenian and *Proceedings of Norwegian Parliamentary Debates* corpora were found this way. The Portuguese and Danish corpora were found through the VLO only by using the corpus name as the keyword. The *Czech Parliament Meetings*, the Danish, Finnish, and the British corpora were found on the websites of the consortia. The links to all the other corpora (i.e. *Czechparl,* those of Austria, Bulgaria, Germany, Latvia, Lithuania, the Netherlands, Sweden, Poland and Hungary) were provided to us by the national UI coordinators.

The documentation for the surveyed corpora is inconsistent. In some cases, it is not obvious how the corpora are annotated (e.g. Estonia, Finland), or other crucial information, such as corpus size (no. of tokens) is not readily available (e.g. *Talk of Norway*).[5]

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.

6 References

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., Schumacher, A. 2016. “Sparv: Språkbanken’s corpus annotation pipeline infrastructure.” http://www8.cs.umu.se/~johanna/sltc2016/abstracts/SLTC_2016_paper_31.pdf. Last accessed on 6 April 2017.

[5] We were able to obtain the information by contacting one of the developers of the corpus.

Généreux, M., Hendrickx, I., Mendes, A. 2012. ”A Large Portuguese Corpus On-Line: Cleaning and Preprocessing.” *Conference: Computational Processing of the Portuguese Language (PROPOR)*.

Jakubíček, M. and Kovář, V. 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, J. Utka, A. Šarkutė, L.. 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.

Koehn, P. 2005. “Europarl: A Parallel Corpus for Statistical Machine Translation”. http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf. Last accessed on 9 April 2017.

Marx, M. and Schuth, A. 2010. “DutchParl: The Parliamentary Documents in Dutch.” http://politicalmashup.nl/new/uploads/2010/03/lrecfinalversionlong.pdf. Last accessed on 7 April 2017.

Ogrodniczuk, M. 2012. “The Polish Sejm Corpus.” http://www.lrec-conf.org/proceedings/lrec2012/pdf/653_Paper.pdf. Last accessed on 7 April 2017.

Oravecz C., Váradi, T., Sass, B. 2014. “The Hungarian Gigaword Corpus.” http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf. Last accessed on 6 April 2017.

Pančur, A. 2016. “Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI.” Conference on Language Technologies & Digital Humanities, Ljubljana 2016.

Rayon, P., Baron, A., Piao, S., Wattam, S. 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 April 2017.

Sippl, C., Burghardt, M., Wolff, C., Mielke, B. 2016. " Korpusbasierte Analyse österreichischer Parlamentsreden.”