Parliamentary corpora in the CLARIN infrastructure

Darja Fišer
Director of User Involvement CLARIN ERIC
darja.diser@ff.uni-lj.si

Jakob Lenardič
Assistant to Director of User Involvement CLARIN ERIC
jakob.lenardic@ff.uni-lj.si

CLARIN Annual Conference 2017
19 September 2017
Budapest, Hungary

CLARIN

Background and motivation

- Background
  - Quintessential resource for numerous SSH research questions
    - *Eduskunta corpus* for critical discourse analysis (Eero Voutilainen)
    - *Greek Parliament Sittings* for analysis of verbal aggressivness (Marianthi Georgalidou)
  - Made easily available under the Freedom of Information acts in over 100 countries all around the world to enable informed participation by the public and improve effective functioning of democratic systems

Motivation

- We wanted to make an overview of the existing parliamentary corpora for all CLARIN members and observers
- We wanted to analyze to what extent they are integrated with the CLARIN infrastructure

Parliamentary records

- Availability
  - Accessible from the parliamentary websites
  - Text: available for all countries except Estonia
  - Video: available for Germany, Hungary, Finland, Greece, the Netherlands, Norway, Europarl
  - Audio: not found
- Period
  - Diachronic
    - Norway (1814-), Portugal (1821-), UK (1807-), Latvia (1920s-), Austria (1920-), the Netherlands (1814-)
  - Contemporary
    - Greece (1990s-), Lithuania (1990s-), Sweden (1971-), Slovenia (1990-), Hungary (1990-), Bulgaria (2001-), the Czech Republic (2013-), Denmark (cca. 2000-), Finland (unknown), Germany (2013-), the Netherlands (2014-)

Parliamentary records

- Formats
  - pdf: Italy, Germany, Portugal, Poland
  - html: Czech Republic, Hungary, Norway, Slovenia
  - pdf and html: Austria, Denmark, the Netherlands, UK, Finland, Sweden, Europarl, Latvia
  - pdf and xls: Bulgaria
  - pdf and docx: Greece, Lithuania

Parliamentary corpora 1/2

NC	Corpus	Size	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	D
dk	DK-CLARIN Almensprogligt korpus	7.3	2008-2010	T,PoS,L	VLO	D
ee	Transcripts of Riigikogu	13	1995-2001	/	VLO	D,C
fi	Eduskunta Corpus	22	2008-2016	/	FIN-CLARIN	C
de	PolMine Sample Corpus	/	/	/	E-Mail	D
It	/	/	/	/	/	/
el	Hellenic Parliament Sittings	29	2011-2015	/	CLARIN:EL	D
lv	SAEIMA	/	1993-2016	/	E-Mail	C

Parliamentary corpora 2/2

NC	Corpus	Size	Period	Anno	Found	Avail
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	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

Summary

- Coverage
  - 20 corpora identified in total; exist for 17 countries; not for Italy
  - the Czech Republic & Norway have 2 each
- Size (in tokens)
  - largest: UK (1.6 billion), Riksdag's Open Data (1.25 billion)
  - smallest: *Czech Parliament Meetings* (0.5 million), Portuguese (1 million)
  - Generally most between 1 million and 100 million tokens.
- Periods covered by the corpus
  - Generally second part of 20th century and 21st century, Dutch and British corpora from beginning of 19th century.

Availability

- For download (8):
  - Austrian, Czech [CPM], Danish, German [sample only], Norwegian[ToN], Portuguese, Lithuanian, Greek
  - Note: no info if available on concordancers as well
- For on-line searching (7):
  - Finnish (KORP)
  - CzechParl (SketchEngine)
  - Latvian (noSketchEngine)
  - Bulgarian (CLaRK)
  - Hungarian (HNC, registration required)
  - Proceedings of Norwegian Parliamentary Debates (Corpuscle)
  - British
- Both for download and on-line searching (5):
  - Dutch (Political Mashup)
  - Estonian (Keeleveeb)
  - Swedish (KORP)
  - Slovenian (noSketchEngine)
  - Polish (NKJP)

State of the infrastructure

- **7 corpora are available via VLO:**
  - Norwegian (PNPD), Czech (CPM), Danish, Estonian, Portuguese, Slovenian, Hungarian
- 5 corpora are available in CLARIN centres, but not in VLO:
  - Finnish ***Eduskunta Corpus***
  - Greek ***Hellenic Parliament Sittings***
  - Swedish ***Riksdag's Open Data***
  - British ***Hansard Corpus***
  - EU ***Europarl Corpus*** but not the latest version
- 9 corpora are not yet part of CLARIN infrastructure:
  - Austrian, Bulgarian, Czech [CzechParl], German [PolMine], Latvian, Lithuanian, Dutch, Norwegian and Polish

Issues identified

- Corpora indexed by the VLO cannot be found with keywords like *parliament*,* even worse for phrase *parliamentary corpus* or adjective *parliamentary;*
  - DK-CLARIN Almensprogligt korpus
  - PTPARL Corpus
  - SlovParl
- Missing metadata:
  - Unknown size and annotation for Latvian *SAEIMA*
  - Unknown annotation for Estonian *Transcripts of Riigikogu* and Finnish *Eduskunta Corpus*
  - Clear license info only for three corpora: *SlovParl* and *Czech Parliament Meetings, Eduskunta Corpus*

Suggestions for corpora developers

- Improving the metadata for corpora already in the infrastructure
  - Using *parliament(ary)* in the title of the metadata file
    - Provide descriptions in English
  - Using terms like *parliament(ary)* in description → better ranking
  - Using distinctive titles
    - e.g. 148 hits for the generic “Flemish parliamentary debate”
  - Including licensing information
- Adding the metadata of the 5 corpora available in CLARIN centres but not VLO to the LRI

Suggestions for corpora developers

- Adding the following exiting corpora to a CLARIN centre:
  - Austrian ***Korpusbasierte Analyse österreichischer Parlamentsreden***
  - Bulgarian ***Corpus of Bulgarian Political and Journalistic Speech***
  - Czech ***CzechParl***
  - German ***PolMine*** **corpus**
  - Latvian **SAEIMA**
  - Lithuanian ***Project Astra STENOGRAMOS_INDV***
  - Norwegian ***Talk of Norway***
  - Dutch ***DutchParl***

Further endeavours – overviews of other resource families

- CMC corpora, datasets and tools
  - 15 corpora (8 of which part of CLARIN infrastructure)
  - 14 specialized datasets
  - 13 tools (10 in CLARIN)
- Parallel corpora
  - 77 corpora identified
  - 51 part of CLARIN infrastructure, mostly LINDAT;
  - A lot of missing metadata (unclear language direction, alignment)
- Newspaper corpora and likely tools
  - Work in progress

Feedback welcome