CLARIN Corpora for Parliamentary Discourse Research

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

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

LREC 2018 Workshop ParlaCLARIN
7 May 2018
Miyazaki, Japan

CLARIN

Background and motivation

- **Parliamentary proceedings**
  - Unique content, structure and language
  - Big societal impact
  - Simple access
- **Old**
  - Key resource for a wide range of disciplines for the past 50 years
- **New**
  - Interdisciplinary investigations
  - Diverse parts of society (women, minorities, marginalized groups)
  - Cross-cultural dimensions
- **Needed**
  - Empirical research
  - Richly annotated corpora
  - Integrative analytical tools

- **Issues**
  - Corpus development efforts not well co-ordinated
  - Corpora not uniformly sampled, annotated, formatted or documented
  - In many cases not easily accessible
- **Goals**
  - Promote comparability and reproducibility of research results
  - Foster interdisciplinary, trans-national and cross-cultural studies
  - Give an overview of the parliamentary corpora available through CLARIN
  - Discuss how they could be made more readily available to the heterogeneous research community

Overview of CLARIN parliamentary corpora

Country	Size (mil tokens)	Period	Linguistic annotation	Availability
cz	0.5	/	Speech-text alignment	**Download & concordancer, CC-BY**
de	0.4	1998-2015	/	Download, CC-BY
dk	7.3	2008-2010	Tok, PoS, Lem	Download, Pub
ee	13	1995-2001	/	Download & concordancer, Aca
el	28.7	2011-2015	/	Download & concordancer
fi	2.2	2008-2016	/	Download, CC-BY-NC
fr	0.17	2002-2012	/	Download, CC-BY
lt	23.9	1990-2013	Tok, PoS, Lem	Download, Pub
no_1	63.8	1998-2016	Tok, PoS, Lem	Download, NLOD
no_2	29	2008-2015	/	Concordancer, NLOD
pt	1	1970-2008	Tok, PoS, Lem	Download, ELRA END USER & VAR
se	1,250	1971-2016	**Tok, PoS, Lem, Semantic**	**Download & concordancer, CC-BY**
si	10.8	1990-1992	Tok, PoS, Lem	**Download & concordancer, CC-BY**
uk_1	**1,600**	**1803-2005**	Tok, PoS, Lem	Concordancer
uk_2	0.19	1998-2015	/	Download, CC-BY
eu	588	1996-2011	Sentence alignment	Download

Recommendations towards improved visibility of CLARIN parliamentary corpora

- **Focus on intended use and users**
  - Annotated, formatted, documented and released in such a way that they are valuable for researchers from diverse research backgrounds
  - Comprehensive data and metadata inclusion policies
  - Cross-referencing corpora with external knowledge bases (e.g., biographical lexica)
  - Regular updates of corpora with new materials
- **Optimisation of user interfaces and documentation**
  - Easy access and navigation
  - Research environments that prevent interface fatigue and support comparative research
  - Good documentation needed for resource and tool criticism as well as interpreting research results
  - Advanced functionalities and tools (e.g., text mining, visualization)

Recommendations towards improved visibility of CLARIN parliamentary corpora

- **Improvement of data structure and annotation**
  - A systematic roadmap with a set of mutually agreed-upon building blocks and standards for text annotation, corpus encoding and metadata encoding
  - Procedures to guarantee and monitor the quality of not only metadata but also the quality of data and tools
- **Increasing outreach activities and knowledge sharing**
  - Systematic promotion of parliamentary corpora and relevant tools
  - Training materials and best practice use cases for parliamentary data across research disciplines
  - Consolidation of research community
  - Educational use cases should be integrated into university curricula

CLARIN RESOURCE FAMILIES

https://www.clarin.eu/resource-families/parliamentary-corpora

PARLIAMENTARY CORPORA
16 SURVEYED

1 MULTILINGUAL
15 MONOLINGUAL: 13 LANGUAGES

1 Czech
1 Danish
2 English
1 French
1 Estonian
1 Finnish
2 German
1 Greek
1 Lithuanian
2 Norwegian
1 Portuguese
1 Slovenian
1 Swedish

AVAILABILITY
3 through a concordancer
9 for download
4 both

ANNOTATION
7 PoS-tagged
7 lemmatised

SIZE
7 small (<10 million tokens)
6 medium (10–100 million tokens)
2 large (>100 million tokens)

LICENCE
8 CC-BY
1 CLARIN ACA
1 CLARIN PUB

Parliamentary corpora in the CLARIN infrastructure

Corpus	Language	Description	Availability
Czech Parliamentary Meetings **Size:** 88 hours, 0.5 million tokens **Annotation:** error correction of transcriptions, division into speech sections with speaker information **Licence:** CC-BY	Czech	The corpus contains recordings of the parliamentary sessions as well as corresponding transcriptions. The corpus is available for download from LINDAT and through the concordancer KonText.	Concordancer Download
DK-CLARIN Almensprogligt korpus - offentlig del: tekster fra Folketinge **Size:** 7.3 million tokens **Annotation:** tokenised, PoS-tagged, lemmatised **Licence**: CLARIN PUB	Danish	The corpus contains Danish parliamentary debates from 2008 to 2010. It is annotated with ePOS-DSL. The corpus is available for download from the DK-CLARIN repository.	Download
Hansard corpus	English	The corpus contains British parliamentary debates from 1803	Concordancer

Table of Contents

Introduction
Parliamentary corpora in the CLARIN infrastructure
Additional materials
Publications on the parliamentary corpora

Comments and suggestions welcome!

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

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

LREC 2018 Workshop ParlaCLARIN
7 May 2018
Miyazaki, Japan

CLARIN