Parliamentary data, corpora and tools: An overview

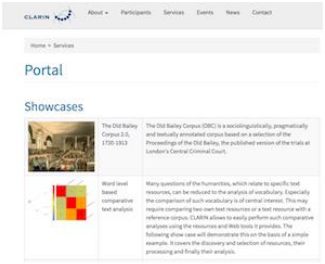
Darja Fišer
Director for User Involvement CLARIN ERIC
Darja.Fiser@ff.uni-lj.si
Jakob Lenardič
Assistant to Director for User Involvement CLARIN ERIC
jakob.lenardic@ff.uni-lj.si

CLARIN-PLUS Workshop "Working with Parliamentary Records"

Sofia, Bulgaria

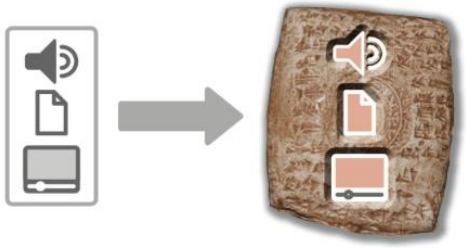
28 March 2017

CLARIN



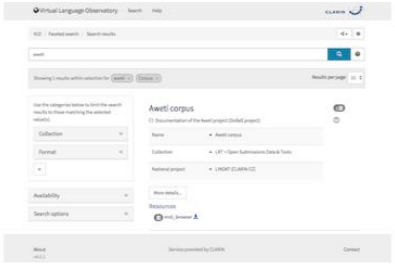
CLARIN portal

Get an example-based impression of what's currently available



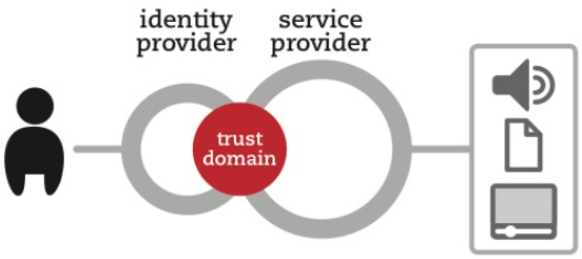
Depositing services

Store language resources in a sustainable repository at a CLARIN centre



Virtual Language Observatory

Discover language resources using a faceted browser or a map

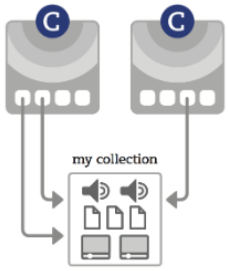


Easy access to protected resources

Get easy access to protected resources, with your institutional username and password.

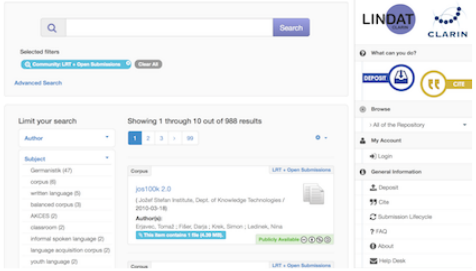
Web services and applications

Explore and analyze language data with a wide variety of tools



Virtual Collections

Create your own digital bookmarks, ideal for citing data sets.



Language Resource Inventory

Submit and access information about language resources relevant to your

Content Search (prototype)

Search different corpora with a single search engine

Consulting Services

Searching for a specific data set or application? Wondering how CLARIN can

CLARIN 4 U

- While CLARIN is based on a federation of long-term archives and services, it is not a static infrastructure
- This workshop is an excellent opportunity for us to:
  - listen to your needs
  - learn from your experiences
  - create additional bridges between the research infrastructure and potential users and data/service providers

Parliamentary records

- Availability
  - Accessible from the parliamentary websites
  - Text: Available for all countries except Estonia and Poland
  - Videos: Available for Germany, Hungary, Finland, Greece, the Netherlands, Norway, Europarl
  - Audio: Not found
- Period
  - Historical
    - The Netherlands (1814-1995)
  - Diachronic
    - Norway (1814-), Portugal (1821-), UK (1807-), Latvia (1920s-), Austria (1920-)
  - Contemporary
    - Greece (1990s-), Lithuania (1990s-), Sweden (1971-), Slovenia (1990-), Hungary (1990-), Bulgaria (2001-), the Czech Republic (2013-), Denmark (cca. 2000-), Finland (unknown), sGermany (2013-), the Netherlands (2014-)
- Formats
  - pdf: Italy, Germany, Portugal
  - 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

Corpora

- Coverage
  - exist for 17 countries; not for Italy & possibly Greece
  - the Czech Republic & Norway have 2 each
- Size (in tokens)
  - largest: UK (1.6 billion)
  - smallest: Portuguese (1 million)
- Availability
  - For download (7):
    - at, cz [CPM], dk, de [sample only], no [ToN], pt, lv
    - 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)
  - Both for download and on-line searching (5):
    - Dutch (Political Mashup)
    - Estonian (Keeleveeb)
    - Swedish (KORP)
    - Slovenian (noSketchEngine)
    - Polish (NKJP)

Problems

- Finding corpora is hard:
  - very few found via keyword search on VLO:
    - Estonian, Slovenian & Norwegian (Proceedings of Norwegian Parliamentary Debates)
  - very few found via corpus name on VLO
    - Portuguese, Danish
  - Found on websites of the consortia
    - Czech (CPM), Danish, Finnish, UK, outdated version of Europarl on LINDAT
- Finding documentation about the corpora is hard:
  - no info on corpus size (Talk of Norway)
  - no info on the annotation tools used (Hansard, Bulgarian, Finnish)
  - No info on the annotation (Finnish)

Overview of results (1/2)

Country	Corpus	Ann0	Size	Period	Avail.	Found
Austria	✓	PoS	X	2013-2015	D	✉
Bulgaria	✓	PoS,L,T	10m	2016-2012	C	✉
Czech Rep (CzechParl)	✓	L,T MSD	81m	1993-2010	C	✉
Czech Rep (CPM)	✓	X	0.8m	X	D	LINDAT
Denmark	✓	PoS,L,T	7.3m	2008-2010	D	VLO, DK-CLARIN
Estonia	✓	X	13m	1995-2001	D + C	VLO
Finland	✓	X	22.4m	2008-2016	C	FIN-CLARIN,✉
Germany	✓	X	X	X	D (sample)	✉
Greece	?	?	?	?	?	X
Italy	X	X	X	X	X	X
Latvia	✓	X	X	1993-2016	C	✉

Overview of results (2/2)

Country	Corpus	Anno	Size	Period	Avail.	Found
Netherlands	✓	PoS, L,T	800m	1814-2014	D + C	✉
Norway (ToN)	✓	PoS, L,T	63.8m	1998-2016	D	Google
Norway (PoNPD)	✓	T	29m	2008-2015	C	VLO
Sweden	✓	PoS, L, T	1.25bn	1971-2016	D + C	✉
Poland	✓	L,T	114m	1991-2011	D + C	✉
Portugal	✓	PoS,L,T	1m	1970-2008	D	VLO
Slovenia	✓	PoS, L, T	3.2m	1990-1992	D + C	VLO
UK	✓	PoS,L,T	1.6bn	1803-2005	C	CLARIN-UK, ✉
Hungary	✓	PoS,L,T	20.9m	X	C	✉
EU	✓	Sent. Align.	X	1996-2011	D	LINDAT Google

Feedback welcome

darja.fiser@ff.uni-lj.si

CLARIN-PLUS Workshop “Working with Parliamentary Records”

Sofia, Bulgaria

28 March 2017

CLARIN