Introduction to CLARIN Resource Families and Parthenos Training Module

Darja Fišer

#ParthenosCEE

Sofia, 7 October 2019

CLARIN PARTHENOS

CLARIN Resource Families

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

Intro

- Aim:
  - provide a user-friendly overview of the available language resources in the CLARIN infrastructure
- Included:
  - overviews are organized according to the types of data in the resources and include listings sorted by language
  - listings include the most important metadata and brief descriptions
    - resource size, text sources, time periods, annotations, licences, links to download pages and concordancers
  - hyperlinks to other relevant materials such as the thematic CLARIN workshops and tutorials and their accompanying videolectures
  - list of key publications on the resources surveyed

Current state

- Corpora
  - Computer-mediated communication corpora
  - Historical corpora
  - L2 learner corpora
  - Literary corpora
  - Newspaper corpora
  - Parallel corpora
  - Manually annotated corpora
  - Parliamentary corpora
  - Spoken corpora
- Lexical resources
  - Lexica
  - Dictionaries
  - Conceptual Resources
  - Glossaries
  - Wordlists

To do in 2020:
Tools
Showcases

Parliamentary corpora in CLARIN Resource Families

- 18 parliamentary corpora
  - 15 national parliaments in 14 languages
  - the corpus of the European parliament proceedings in 21 languages
  - most are richly linguistically annotated
- access
  - directly downloaded
  - queried through online search environments
  - available under open licences
- characteristics
  - UK's Hansard Corpus the largest (1.6 billion tokens, 7.6 million speeches made by around 40,000 different speakers) and spanning the longest time period (1803-2005)
  - corpora from other countries are significantly smaller (10 - 100 million tokens) and cover shorter periods (mostly from the 1970s onwards)

Using parliamentary corpora in CLARIN

Hansard corpus **Size:** 1.6 billion tokens **Annotation:** tokenised, PoS-tagged, lemmatised, semantic tagging	English	The corpus contains British parliamentary debates from 1803 to 2005. It is semantically tagged with the USAS semantic tagger and the Historical Thesaurus Semantic Tagger (HTST). The corpus is available through a dedicated concordancer. For the relevant publication, see Rayson et al. (2015).	Concordancer

DISPLAY

LIST CHART KWIC COMPARE

SEARCH STRING

WORD(S) *ism

COLLOCATES

POS LIST

SEMANTIC CATEGORIES | WORDS

RANDOM SEARCH RESET

SHOW DECADE SPEAKER

1 --IGNORE-- 2000 1990 1980 1970 1960 1950

2 --IGNORE-- 2000 1990 1980 1970 1960 1950

SEE CONTEXT: CLICK ON WORD OR SELECT WORDS + [CONTEXT] [HELP...]

		CONTEXT	FREQ
1		CRITICISM	108031
2		TERRORISM	35255
3		MECHANISM	26700
4		TOURISM	21084
5		SOCIALISM	15755
6		PATRIOTISM	9165
7		COMMUNISM	8618
8		OPTIMISM	8002
9		CAPITALISM	6734

OVERVIEW: URSA BAJZELJ

Help / information / contact

DISPLAY

LIST CHART KWIC COMPARE

SEARCH STRING

WORD(S) terrorism

COLLOCATES

POS LIST

SEMANTIC CATEGORIES | WORDS

RANDOM SEARCH RESET

SHOW DECADE SPEAKER

CLICK ON BARS FOR CONTEXT

CLICK ON COLUMN HEADINGS FOR FREQUENCY IN SUB

SECTION	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
FREQ	0	0	0	17	7	6	38	99	415	76	93	85	203	374	270	621	363	4856	8991	7994	10747
PER MIL	0.00	0.00	0.00	0.61	0.23	0.18	1.11	2.67	6.91	1.49	1.44	1.06	2.83	3.93	2.85	5.13	2.39	29.73	48.94	45.13	121.50

SEE ALL SUB-SECTIONS AT ONCE

SECTION: 1970 **(4,856)**

PAGE: << < 1 / 49 > >>
SAMPLE: 100 200 500 1000

CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST ---------- CREATE NEW LIST [?]

1	C-1970	----- (C)	A	B	C	Paper of 1969 says: Although an unconventional threat already exists in the form of **terrorism**, the possibility of a conventional attack is not excluded: The organisation, training
2	C-1970	Colin_M (C)	A	B	C	There are the problems of what is described in these trendy times as " urban **terrorism** " — and I have been a bit of an urban terrorist too in
3	C-1970	Walker_Smith (C)	A	B	C	words, which will be widely interpreted as an incentive to violence and sympathy with **terrorism**: I believe that that part of his speech will be keenly resented and widely
4	C-1970	Walker_Smith (C)	A	B	C	use of force and the increasing number of terrorists in Central Africa, and that **terrorism** condoned by others; when we see that the victims of sanctions in Rhodesia are
5	C-1970	Lyon (C)	A	B	C	hon: Gentleman and The Times between them can set against those who have approved of **terrorism** is that it is the argument of pessimism, that it indicates that there is
6	C-1971	Orr (C)	A	B	C	some way or another, of the attitude of the Southern Irish Government to bring **terrorism** to an end: What conceivable changes could be made as a result of such
7	C-1971	Thomson (C)	A	B	C	; or he may now find himself dealing with the new military problem of urban **terrorism**, such as that in Northern Ireland: That again is bound to raise new
8	C-1971	Lyon (C)	A	B	C	comments by the head of the Security Services that he had been arrested under the **Terrorism** Act, which allows indefinite detention without trial and without charges, and that no
9	C-1971	----- (C)	A	B	C	the Northern Ireland Premier asking for the support of the entire community in stamping out **terrorism** and assuring those who give information that it will be treated in full secrecy:
10	C-1971	----- (C)	A	B	C	a charge universally recognised as criminal:] As the Dean was arrested under the **Terrorism** Act, which denies fundamental human rights, and charged under the Suppression of Co
11	C-1971	----- (C)	A	B	C	prove that the new Government of Northern Ireland are a good deal more tough on **terrorism** than their predecessors: That is not only necessary in military terms but for the
12	C-1971	----- (C)	A	B	C	the problem there, he is mistaken: If he thinks that he can meet **terrorism**, however mistaken that terrorism is, by the polite terrorism of statements in the

PARTHENOS

https://training.parthenos-project.eu

Background & Motivation

- from data scarcity to data flooding
  - robust, verifiable and reproducible quantitative, automated approaches, techniques and methods to process, visualize and combine research data
  - research on a much larger scale, sparking entirely novel research questions
- the new resources, tools and techniques need to be well understood
  - data formatting, corpus encoding and corpus annotation for different types of specialized discourse
  - research data and metadata harvested from the source vs. their enrichments obtained from DH tools
  - corpus querying, visualization and interpretation to tackle research questions in the same dataset from a broad range of research disciplines
- goals
  - raise awareness of the societal, institutional and technological circumstances that shape the content, structure and language of a particular type of specialized discourse
  - highlight the importance, potential, peculiarities and implications that these factors have for research in DH
- target audience
  - Digital Humanities or Cultural Heritage researcher/practitioner
  - Computer Scientists researcher or Data Steward

Overview of the training modules

- Introduction to Research Infrastructures
- Management Challenges in Research Infrastructures
- Introduction to Collaborations in Research Infrastructures
- Manage, Improve and Open Up Your Research Data
- Formal Ontologies: A Complete Novice's Guide
- **Digital Humanities Research Questions and Methods**
  - **Collections of Parliamentary Records (CLARIN-ERIC)**
  - Collections of Computer-Mediated Communication (CLARIN-ERIC)
  - Collections of Digitised Newspapers as Historical Sources (Impresso)
  - Digital Humanities and Heritage Research Infrastructures (E-RIHS)
- Citizen Science in the (Digital) arts and Humanities
- eHeritage Webinars

Preliminaries

- dedicated to developing research questions in the Digital Humanities (DH):s
  - finding, working with, contributing data to digital collections and using digital Research Infrastructures (RIs)
  - examples of how DH and eHeritage approaches in cooperation with relevant RIs can lead to innovative research questions and methods
- target audiences:
  - Digital Humanities and Cultural Heritage (CH) researchers and practitioners who wish to learn how to best benefit from and work with digital collections using digital RIs
  - computer scientists and researchers/practitioners in data centres who want to gain insight into more humanities & cultural heritage related aspects of digital research infrastructures
- skill level:
  - researchers and practitioners that are not yet users of RIs
  - researchers already aware of DH and CH Research Infrastructures who need more knowledge how to engage with them and benefit from them

By the end of this module, you should be able to…

- Have an overview of relevant Research Infrastructures, especially their collections and research communities and how to connect with and through them
- Understand how Digital Humanities and eHeritage collections and Research Infrastructures can be employed in the initial research design
- Be aware of methodological opportunities and challenges related to specific data types

Parliamentary proceedings in the Digital Age

- The parliament
  - old
    - the most important political institution in a democratic society
    - journalists, members of the civil society, and humanities and social sciences scholars pay close attention to parliamentary debates
  - new
    - increasingly decisive role
    - changing institutional relations with the public, the mass media, the executive branch and international organizations
- Parliamentary proceedings
  - Freedom of Information Act: free and transparent access
  - benefits
    - informed participation by the public
    - improves effective functioning of democratic systems
    - makes the datasets more readily available for researchers with heterogeneous backgrounds, including diachronic and transnational comparative analyses

Parliamentary debates for cross-disciplinary research

- Parliamentary discourse
  - governed by strict rules and conventions
    - specific institutional discursive features
  - motivated by a range of communicative goals & ideology / party line
    - role-based commitments and confrontations
    - constant awareness of a multi-layered audience
- Parliamentary records
  - unique content, structure and language, records
  - important resource for a wide range of research questions from many disciplines
    - political science, communication studies, history, sociolinguistics, multilinguality
- Interdisciplinary scope very new
  - technological and methodological developments enabled cross-fertilization of disciplines
  - text mining, social network analysis, geospatial analysis, data visualization >> structure, search, mine, manipulate, visualize, share, and combine parliamentary data
  - apply interpretative traditions of the humanities and social sciences to data on a very large scale
  - allow them to address new research questions and develop novel techniques for tackling complex social phenomena (migrant crisis, Euroscepticism, populist movements)
  - more comprehensive exploration of various socio-political research questions
- Empirical research and development of integrative analytical tools needed
  - a better understanding of commonalities and specificities of parliamentary discourse
  - its wider societal impact, in particular with studies of diverse parts of society (women, minorities, marginalized groups) and cross-cultural studies

Parliamentary proceedings as a research dataset

- key characteristics
  - essentially transcriptions of spoken language produced in highly controlled and regulated settings
  - rich in invaluable (sociodemographic) metadata
  - rich in links to other research data (e.g. legislation)
- key requirements
  - easily findable and accessible
  - encoded according to international standards / recommendations
  - equipped with rich and reliable annotations and metadata
- issues
  - Corpus development efforts not well co-ordinated
  - Corpora not uniformly sampled, annotated, formatted or documented
- goals
  - Promote comparability and reproducibility of research results
  - Foster interdisciplinary, trans-national and cross-cultural studies
  - Discuss how they could be made more readily available to the heterogeneous research community

Case study 1: the War in Parliament project

- data curation project
  - made proceedings of the Dutch parliament available as a semi-structured corpus compliant with CLARIN standards
  - corpus accessible through PoliticalMashup
  - advanced search engine tailored to historical and social science research
  - illustrative case study that showed how a corpus-based approach to the analysis of parliamentary proceedings unveiled certain aspects of the political past that had before remained only as vaguely remembered events in a nation's collective memory
- Piersma et al. (2012): systematically check how the Dutch Boerenpartij (Farmers' Party) was associated with National Socialism between the years 1958 and 1982
  - the Boerenpartij was implicitly criticised for their right-wing political stance throughout this entire period
  - BUT: actually only once directly accused of fostering National Socialist ideas, and this was when Hendrik Adams, a member of the party, was singled out for being a supporter of the German occupier during World War II

Case study 2: Gender in the Danish parliament

- Hansen et al. (2018) investigated gender differences in the revised transcripts of speeches from the Danish Parliament 2009-2017
  - active participation of women in politics is historically relatively new and women are still underrepresented in the Danish parliament
  - analysis:
    - number of the speakers, their age, party and role in the party
    - speech frequencies & speech lengths
    - topics addressed
  - results:
    - general
      - there were relatively more male spokespersons than female ones in the period covered by the corpus but the number of female MPs under 29 is larger than the number of male MPs from the same age group
      - in general, women speak less frequently and for a shorter time than male MPs in proportion to their seats in Parliament, the difference in speaking time between female and male MPs is statistically significant
    - role
      - women belonging to left-wing parties speak less frequently than women from right-wing parties compared to their seats in Parliament
      - female ministers and spokespersons speak more frequently than ordinary MPs
      - female ministers under a male prime minister give fewer speeches than female ministers under a female prime minister even though their percentages in the two periods are similar
    - topics
      - female MPs more often spoke about “softer” political areas, while in the speeches of male MPs “harder” subjects prevailed

Case study 3: The Linked Open Data of Talk of Europe

- data curation project
  - turned the multilingual proceedings of the European Parliament into a highly interactive and structured parliamentary corpus
  - enriched with Linked Open Data
    - linked to encyclopedia to provide information about the Members of Parliament
    - linked to a geographical knowledge base for European countries
  - corpus is accessible through an online search interface that enables SPARQL query language
  - ideal dataset for tackling various research questions where the gap between qualitative and quantitative approaches to parliamentary analysis needs to be bridged
- Kessels et al. (2014)
  - network analysis to explore highly complex constellations of relations between members of the EU parliament based on their interactions in plenary debates
- Mandravickaite et al. (2015)
  - stylometric analysis of the speeches of the EU members of parliament to uncover how the rhetoric of the members of the EP is similar / different to the rhetoric of the party groups they belong to and vs. of the other party groups

Veni, vidi, CLARIN!

https://www.clarin.eu/content/clarin-for-researchers
https://www.clarin.eu/content/knowledge-sharing

CLARIN PARTHENOS