CLARIN services for FAIR language resources

Franciska de Jong

executive director CLARIN ERIC

f.m.g.dejong@uu.nl

AIUCD2021 - DHs for society, 21 January 2021

CLARIN

From the call text of AIUCD 2021

Digital humanists, thanks to the knowledge and expertise acquired in recent years and experimented in numerous research activities, have been among the often silent protagonists of the change of society, pervaded by information technologies in continuous development.

Part I

DHs for society and the role of FAIR data

Part II

CLARIN's service infrastructure – a bird's eye view

Part III

Examples of CLARIN service offer for DH tracks with potential for societal impact

Part IV

Concuding remarks

DH for Society

AIUCD

- Digital Humanties as a silent protagonist of societal change in the era of IT

Open Science

- Research as potential protagonist of innovation
- Investment in research with public funding should benefit society at large
- No monopoly for commercial parties

Digital turn -> Open Science +

- Adherence to principles of responsible data science

Aspects of Open Science

Open Access

Open Source

Open Data

Aspects of Open Science

Open Access

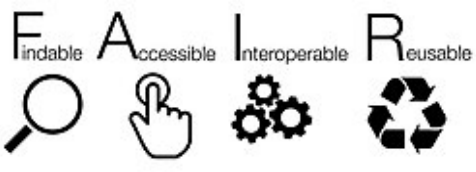
- Focus on reuse of academic publications
- Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003)
- https://openaccess.mpg.de/Berlin-Declaration

Open Source

- Focus on reuse of software
- OSS declaration (1998)

Open Data

- Focus on reuse of data resulting from research
- More recently (2016) promoted as FAIRness of data:
  Findable, **A**ccessible, **I**nteroperable, **R**eusable
- https://www.force11.org/group/fairgroup/fairprinciples



Open Science is calling for an Open Mind

- Open Science is calling for open access to data. However, in many domains ….
  (underlining that not all data can be open, and that Open Science is not to be imposed as a unified framework for all).

- Data and algorithms are surrounded by an aura of accuracy, objectivity, and truth. However, ….
  (calling for attention for values such as democratic society, fundamental rights and the rule of law, through examination of epistemic and ethical impact and regulations that promote and enable good data practices)

Open Science is calling for an Open Mind

- Open Science is calling for open access to data. However, in many domains …. (underlining that not all data can be open, and that Open Science is not to be imposed as a unified framework for all).
  Data open if it can, closed if required.

- Data and algorithms are surrounded by an aura of accuracy, objectivity, and truth. However, ….
  (calling for attention for values such as democratic society, fundamental rights and the rule of law, through examination of epistemic and ethical impact and regulations that promote and enable good data practices)

  Data-driven research is to follow the principles of Responsible Data Science, taking into account context and diversity

Europe’s multilinguality is key*

- to our understanding of how language affects identity, culture, society
- to our understanding of diversity across boundaries of time and regions
- and therefore for comparative studies

* image from https://www.coe.int/t/dg4/linguistic/jel_en.asp

Language data as pillar for data science for society

- Europe's **multilinguality** is a basis for **comparative research** of societal and cultural phenomena, that are reflected in language use
- Some examples:
  - Migration patterns
  - Intellectual history
  - Language variation across period and region
  - Dynamics in mental health conditions / speech pathology
  - Parliamentary discourse

Data reuse and sharing: disciplinary dynamics

History

- from fear for subjectivity to interest in multiperspectivity

Linguistics

- shift in attention from language proper to language in context
  - Identity shaping
  - Language variation
  - Language as vehicle for memories
  - Language as facilitating a dialogue
  - Language as part of multimodal expression

Social sciences

- Integration of surveys and interviews

CLARIN and Open Science

- Promoting the sharing and re-use of data through sustainable data registries
- All integrated datasets available in open access for research purposes
- Support for linguistic diversity
  - Data covering more than 1500 languages
  - Tools for many languages
  - Language resources in all modalities
- Adherence to FAIR data principles
  - Interoperability through a common metadata framework
- Promotion of responsible data science
- Strengthening the support for 500.000 professional SSH researchers

CLARIN: Towards FAIR and Responsible Data Science Using Language Resources." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, May 2018, 3259-3264.

Part II

A bird’s eye view on the CLARIN services

CLARIN in eight bullets

- **CLARIN** is the Common Language Resources and Technology Infrastructure
- **ESFRI** ERIC status since 2012, Landmark since 2016
- that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form)
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a **single sign-on** environment
- that serves as an ecosystem for **knowledge exchange**
- and: with some services already **integrated in EOSC** (European Open Science Cloud; link)

CLARIN ERIC in members and centres

A consortium of:

- 21 members:

  AT, BG, CY, CZ, DE,

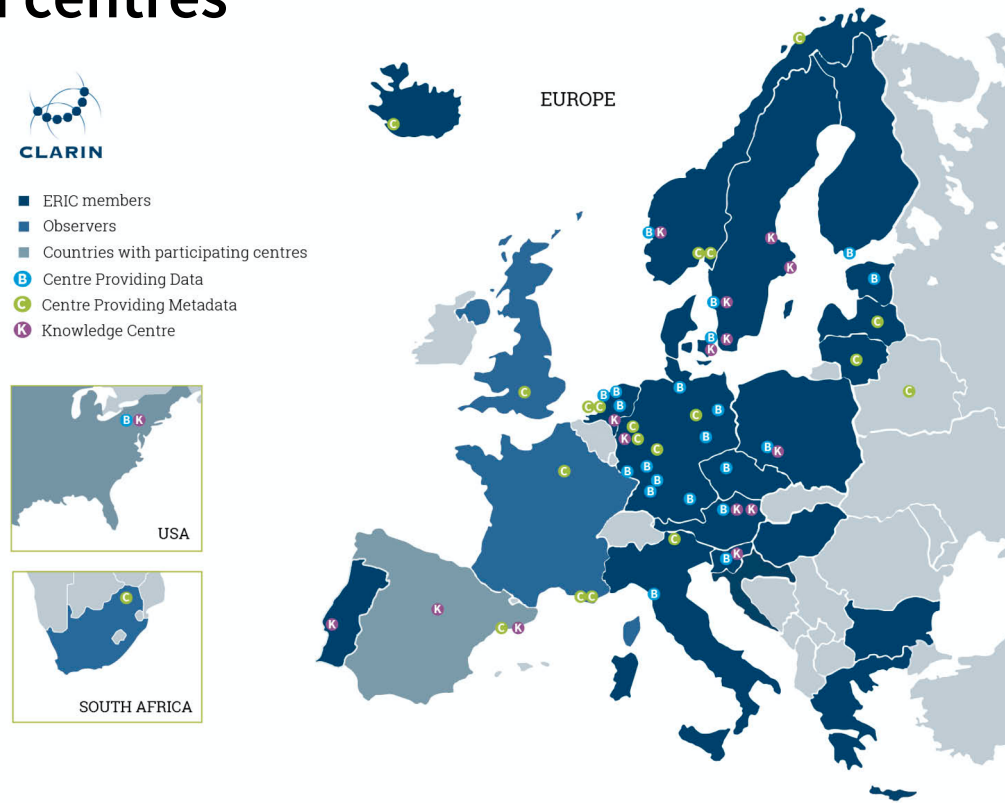
  DK, EE, FI, GR, HR, HU, IS, IT,

  LT, LV, NL, NO, PL, PT, SE, SI

- 3 observers:

  FR, UK, ZA

>64 centres, of which
>24 CTS* certified data centres



*CTS = Core Trust Seal: www.coretrustseal.org

CLARIN in data types

- Newspaper archives
- Literary texts
- Social Media data
- Parliamentary records
- Historical letters
- Oral History data
- Disciplinary libraries
- Institutional archival data
- Broadcast archives
- …

See also the info on the CLARIN Resource Families initiative: https://www.clarin.eu/resource-families

CLARIN's communities of use

- Linguistics and Philology
- Digital Humanities
- Translation and Lexicography
- Literary Studies
- History
- Political and Social Sciences
- Media Studies
- Culture heritage experts
- Speech therapy
- Teachers
- General Public
- …

CLARIN in central services

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.

Language Resource Switchboard

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 research.

Content Search (prototype)

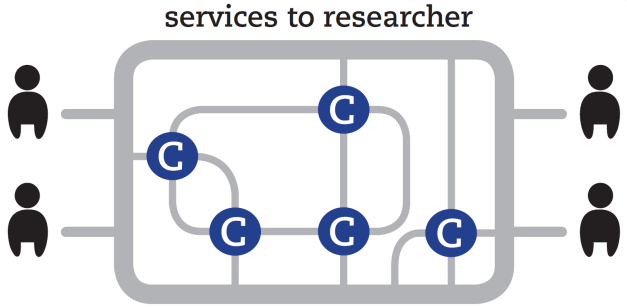
Search different corpora with a single search engine

Questions & Answers

Searching for a specific data set or application? Wondering how CLARIN can assist your research? Feel free to contact us!

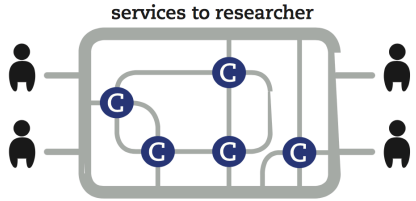
CLARIN: network of centres

- **Distributed architecture**: (http-accessible) files, web applications and web services spread all over centres in Europe and beyond
- Metadata **harvesting** to enable search and access from one central platform
- Currently:
  - 24 certified **B Centres**
  - over 60 registered centres in total
- Tools and data from different CLARIN centres are **interoperable**, so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work.

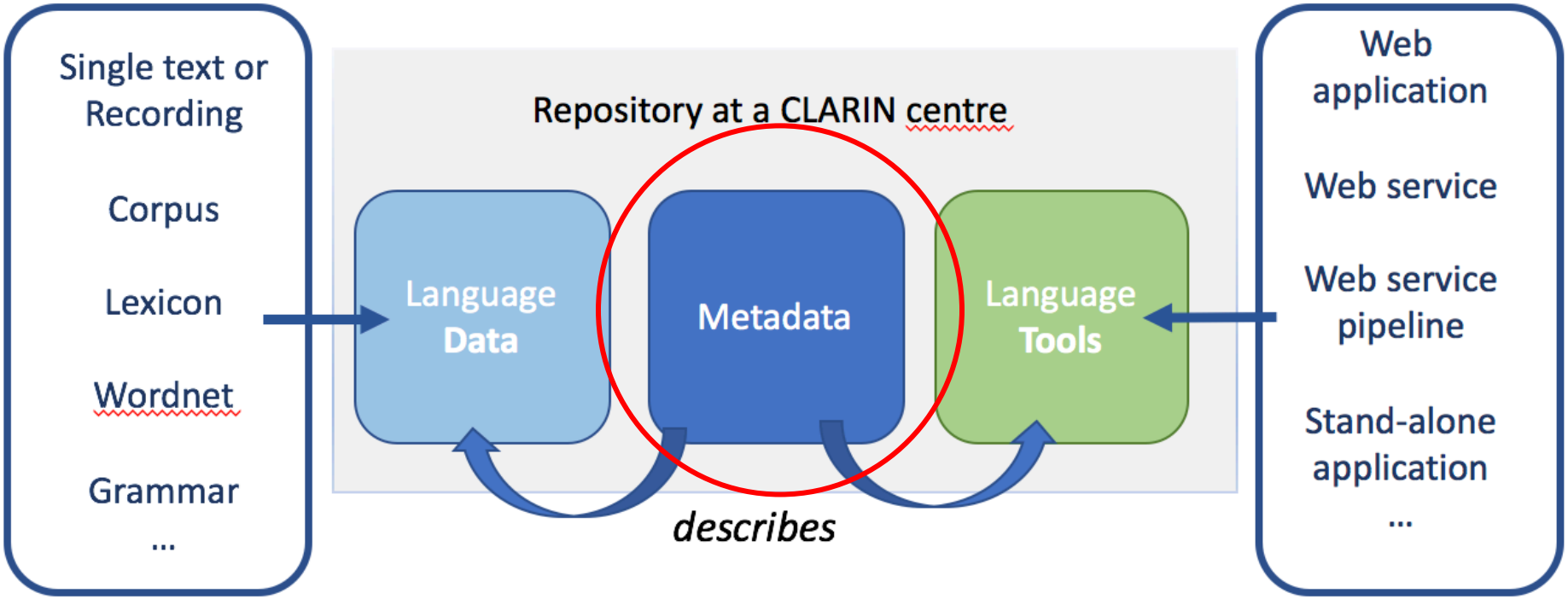


What CLARIN Centres offer

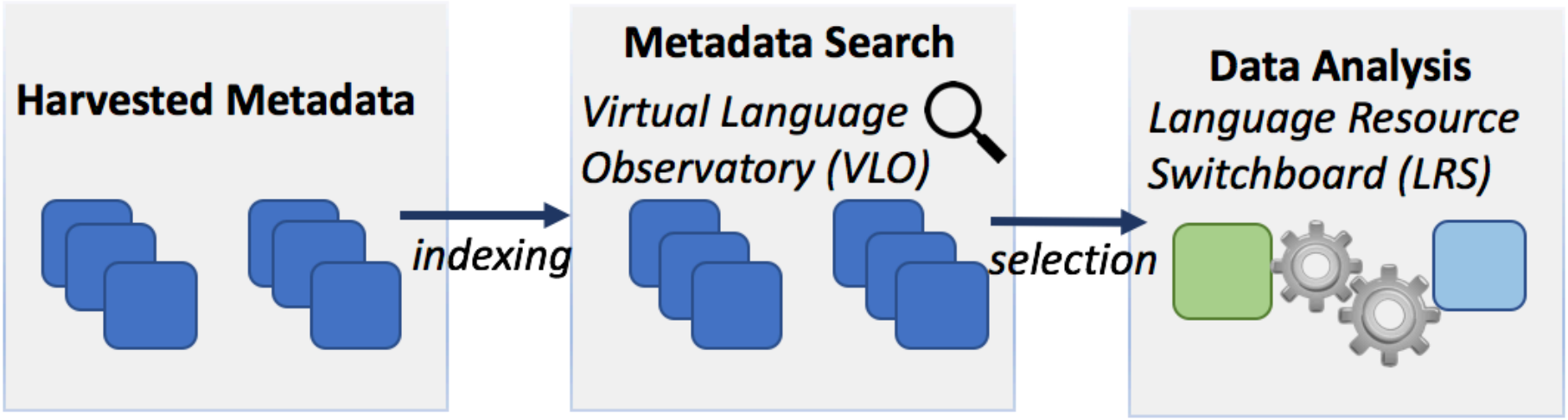
- Repository
  - library of linguistic data and tools
  - search for data and tools and easily use them online or download them
  - deposit your data and be sure it is safely stored, everyone can find it, and correctly cite it
- Federated single sign-on
  - log in once with your existing institutional credentials
  - get access to protected resources
- Metadata
  - describe content, provenance and formats of linguistic data and tools
  - facilitate preservation and dissemination of linguistic data and tools
- Persistent Identifier (PID or handle)
  - a special permanent URL that provides a permanent link to linguistic data and tools
  - will resolve correctly even if in some distant future the data is moved
  - to be used as URL in citations
- Licensing, aligned with diverse conditions
  - Public
  - Academic
  - Restricted
- Preservation
  - committed to long-term care of items in the repository



The CLARIN data architecture: *centre repositories*



The CLARIN data architecture:
central processing of metadata



Examples of what CLARIN can do for you

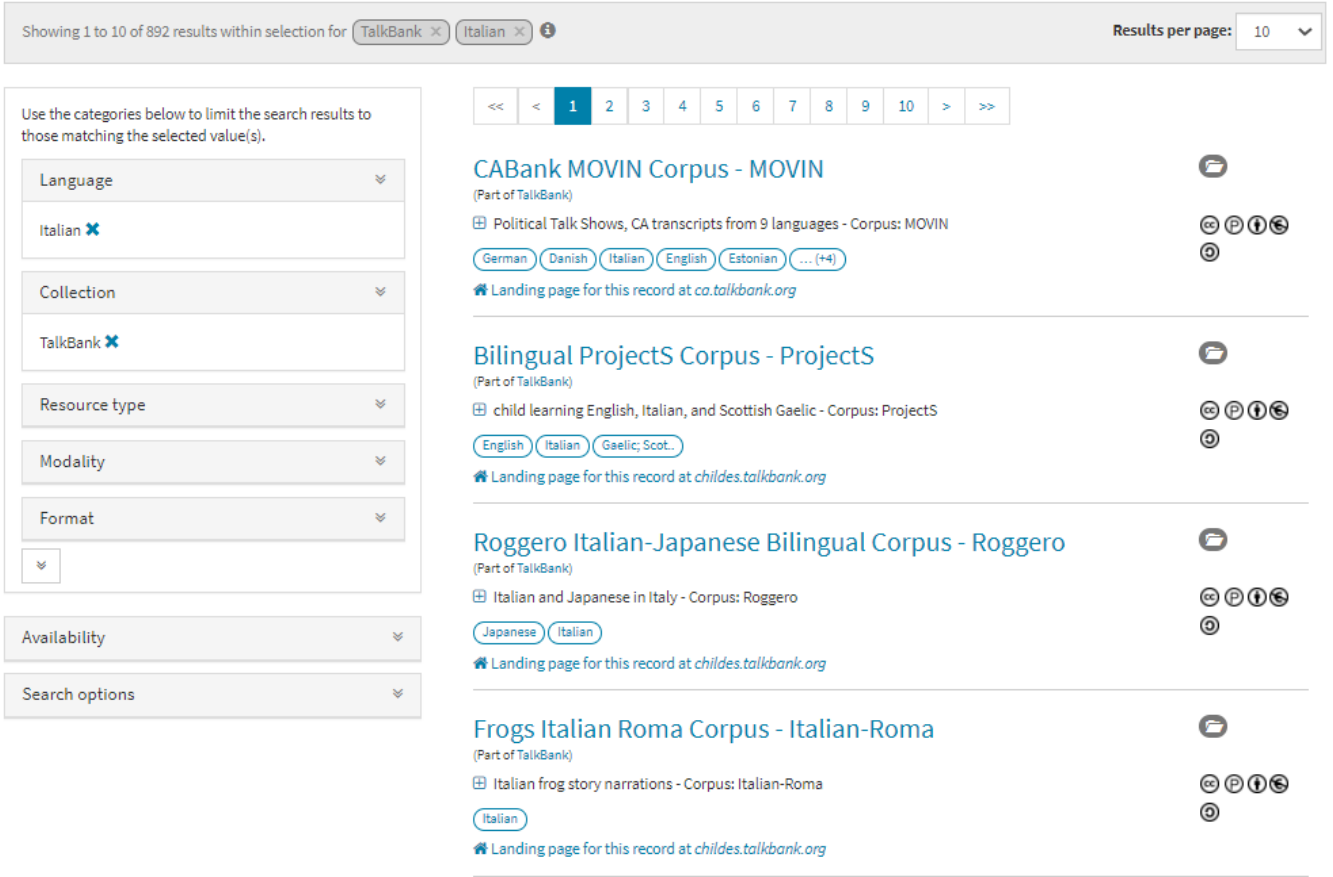
- Datasets to be found via central discovery service
- Network of open access repositories
- Interoperability through common metadata framework
- Information on rights for reuse through licence details
- Virtual Language Observatory: search based on facets
- SwitchBoard: matching tool for selecting suitable tools
- Federated Content Search: content based search across the centre network
- CLARIN Resource Families: overviews of comparable datasets in multiple languages
- Virtual Collection Registry: create, cite, share coherent set of digital objects

Examples of what CLARIN can do for you

- Datasets to be **F**ound via central discovery service
- Network of open **A**ccess repositories
- **I**nteroperability through common metadata framework
- Information on rights for **R**euse through licence details
- Virtual Language Observatory: search based on facets
- SwitchBoard: matching tool for selecting suitable tools
- Federated Content Search: content based search across the centre network
- CLARIN Resource Families: overviews of comparable datasets in multiple languages
- Virtual Collection Registry: create, cite, share coherent set of digital objects

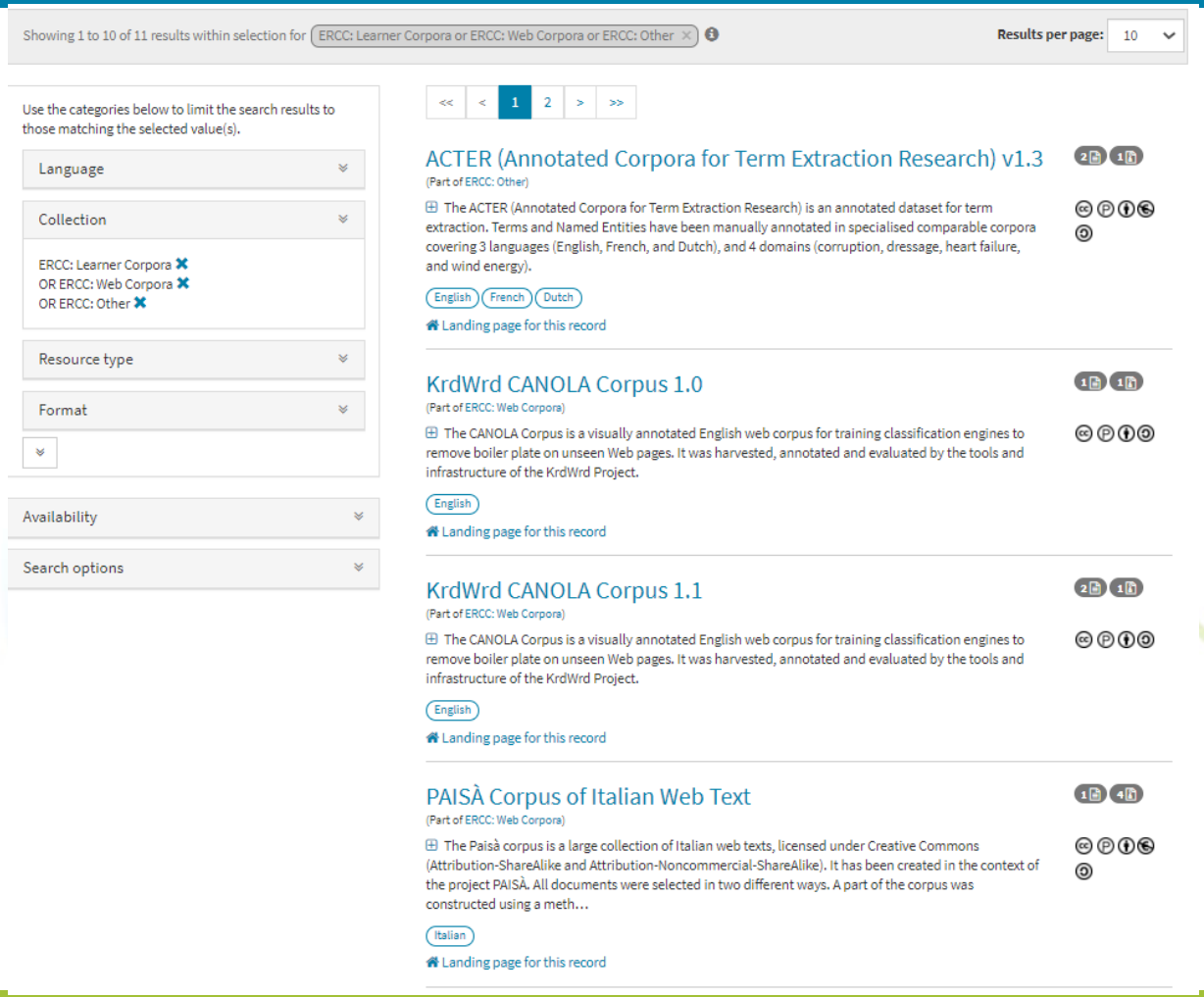
VLO facets selected:

Italian
TalkBank



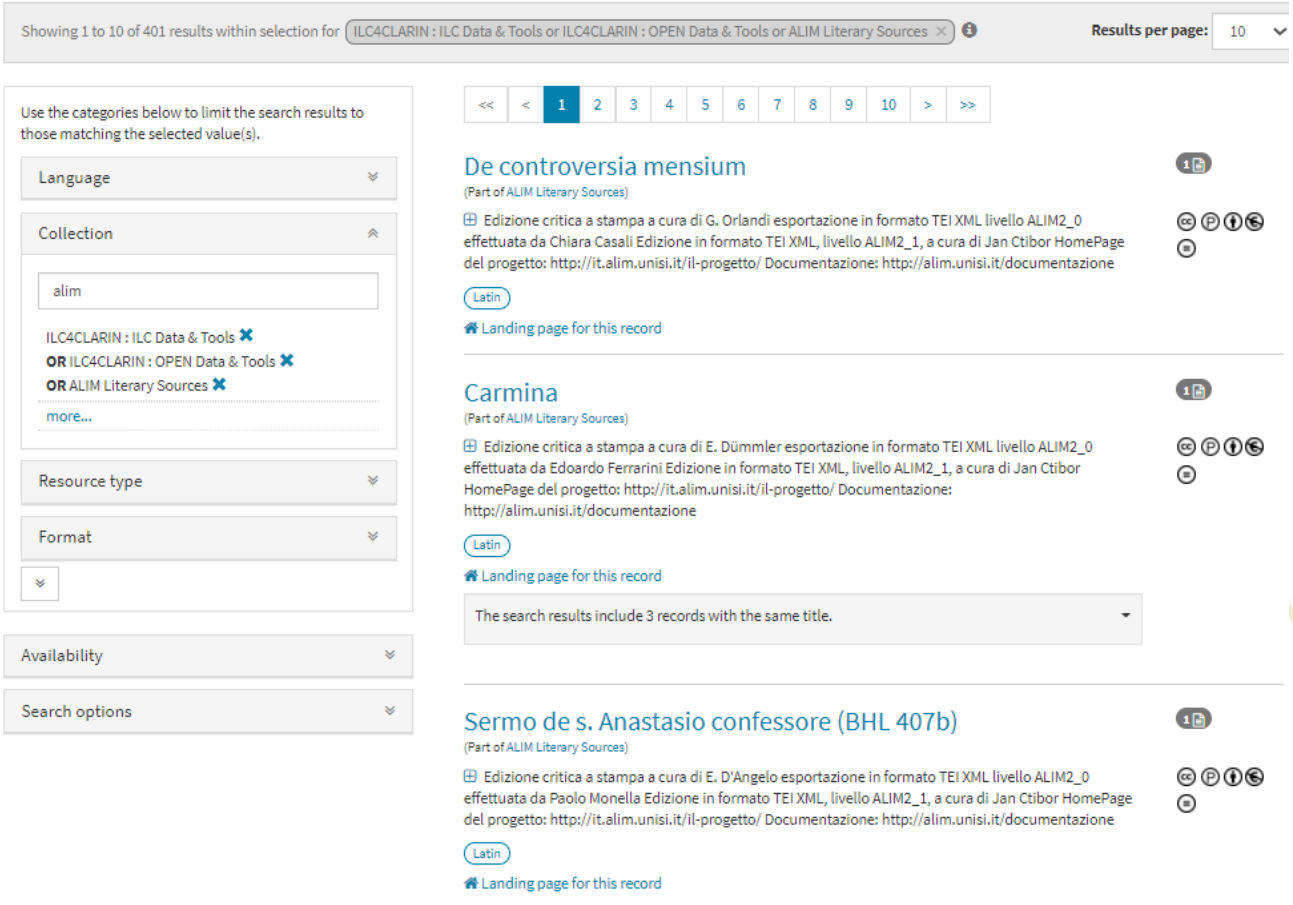
VLO facets selected:

ERCC



VLO facets selected:

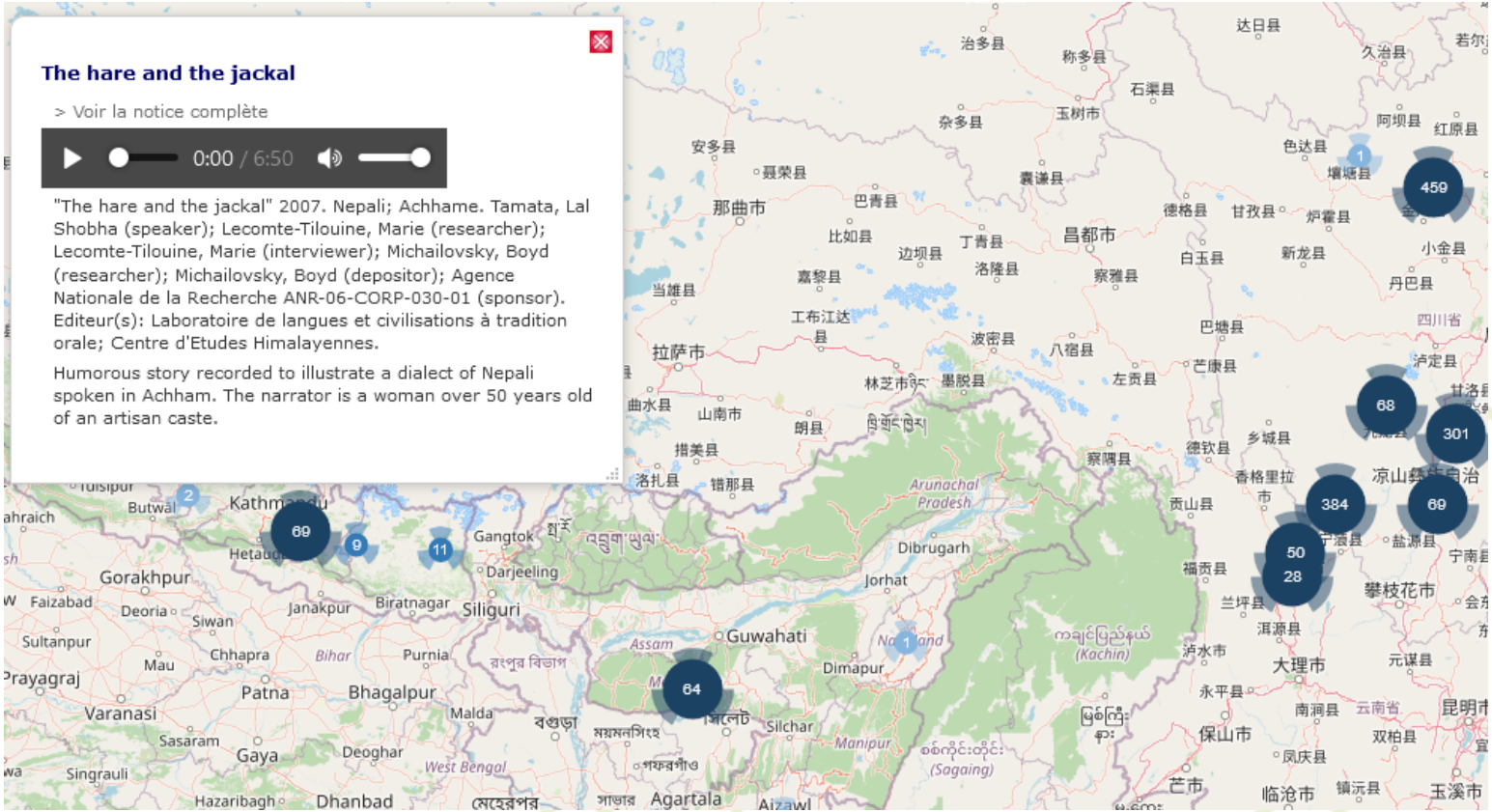
ILC4CLARIN
ALIM



ALIM	CLARIN-IT
Digital library: resources	Infrastructure: discover, search

Image copied from the flash presentation at AIUCD2021 by Marina Buzzoni on ALIM:
"A tea for two"

Cocoon – collection of stories in multiple language variants



CLARIN Resource Families (CRF)

Currently:
12 corpora families
5 families of lexical resources
4 tool families

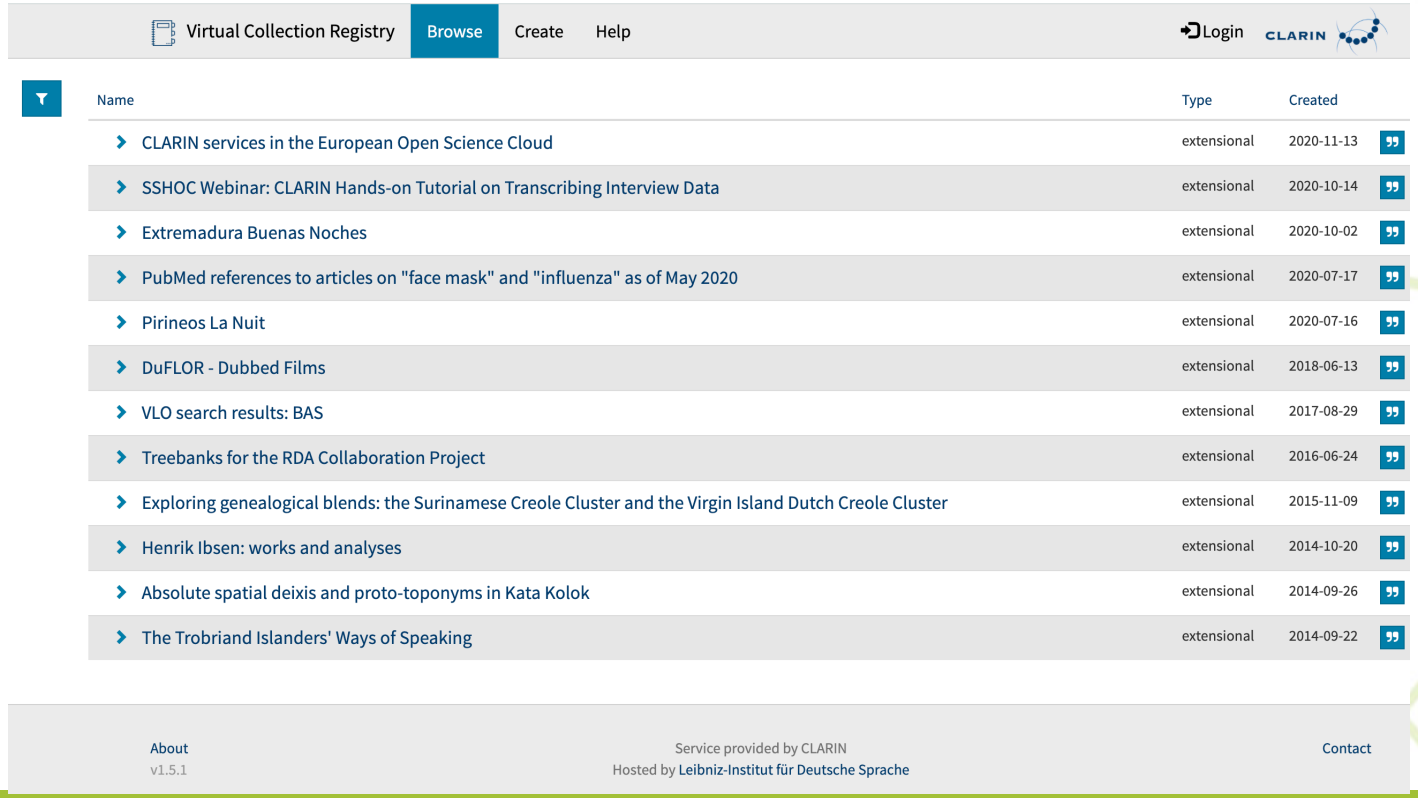
Read more :
https://www.clarin.eu/resource-families

Computer-Mediated Communication corpora (16)

SoNaR New Media **Size:** 35 million tokens **Annotation:** tokenised, PoS-tagged, lemmatised	Dutch	This corpus contains tweets, chats and SMS from 2005 to 2012. The corpus is available for download from the Dutch Language Institute and for searching online through the OpenSONAR environment. For the relevant publication, see Sanders (2012).	Concordancer Download
NTAP English **Size:** 660,798,199 tokens	English	This corpus contains blog posts that are related to climate change issues across science, politics, and the environment. The vast majority of the posts are from 2005 onwards. The corpus is available for searching online through the Corpuscle concordancer. For the relevant publication, see Salway et al. (2016).	Concordancer
DIDI - The DiDi Corpus of South Tyrolean CMC 1.0.0 **Size:** 600,000 tokens **Annotation:** tokenised, PoS-tagged, lemmatised **Licence:** ACA-BY-NC-NORED 1.0	English, German, Italian, Ladino	The corpus consists of Facebook posts gathered from 136 Facebook users from South Tyrol. All texts are anonymised. The corpus is available for download from the EURAC Research CLARIN repository. For the relevant publication, see Frey et al. (2016).	Download
The Mixed Corpus: New Media **Size:** 25 million tokens **Annotaiton:** tokenised	Estonian	This corpus contains chat room messages, forum posts and news comments from 2000 to 2008. The corpus is available for download from a dedicated webpage associated with CLARIN Estonia and through a dedicated concordancer.	Concordancer Download

Virtual Collection Registry

- https://collections.clarin.eu/public?0



CLARIN and service integration into EOSC portal

- The European Open Science Cloud is emerging (www.eosc.eu)
- An EOSC portal is up and running already.
- More and more discipline-specific services are being integrated (registered)
- At the launch of the EOSC portal the integration of the VLO and the Language Resource Switchboard in EOSC was demonstrated: https://www.clarin.eu/eosc
- 135,000 cultural heritage objects from Europeana (news papers) visible through VLO and therefore also in EOSC:https://www.clarin.eu/tags/europeana

Part III

- CLARIN service offer for DH work with potential for societal impact

CLARIN Resource Family: Parliamentary corpora

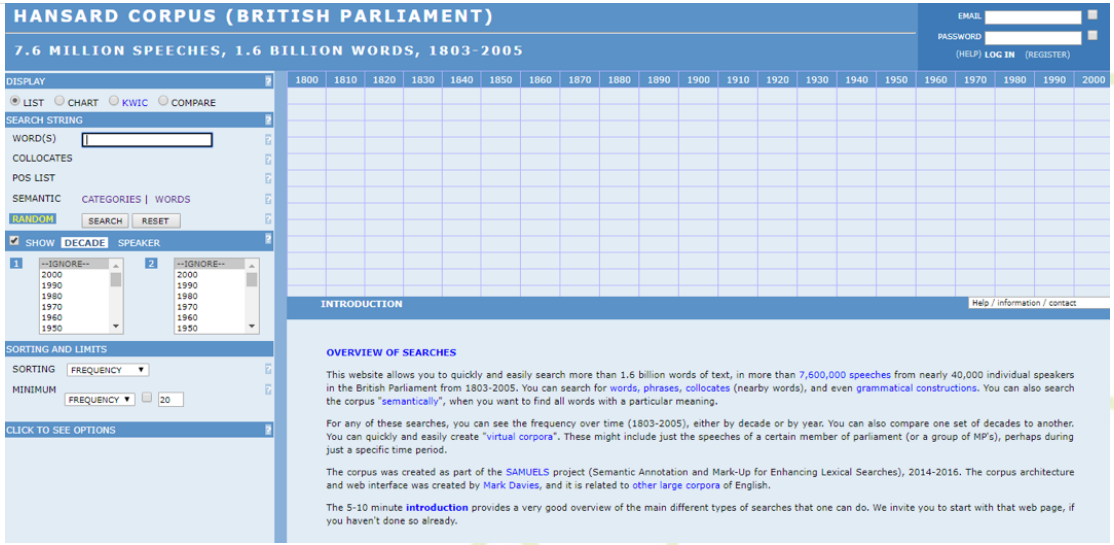
- Characteristics
  - essentially transcriptions of spoken language produced in highly controlled and regulated settings
  - rich in invaluable (sociodemographic) metadata
  - rich in links to other research data (e.g., legislation, data from news channels)
- Requirements
  - easily findable and accessible
  - encoded according to international standards / recommendations
  - equipped with rich and reliable annotations and metadata
- Scenarios of scholarly and public use and benefits
  - fuels multidisciplinary research, diachronic and transnational comparative analyses
  - enables more informed participation in public debate
  - supports effective functioning of democratic systems
  - facilitates focused comparative studies on public debates on global crises

Indication of diversity of corpora integrated in CLARIN

- UK’s Hansard Corpus is the largest corpus (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)
- 25 European national parliaments
- The corpus of the European parliament proceedings comes in 21 languages

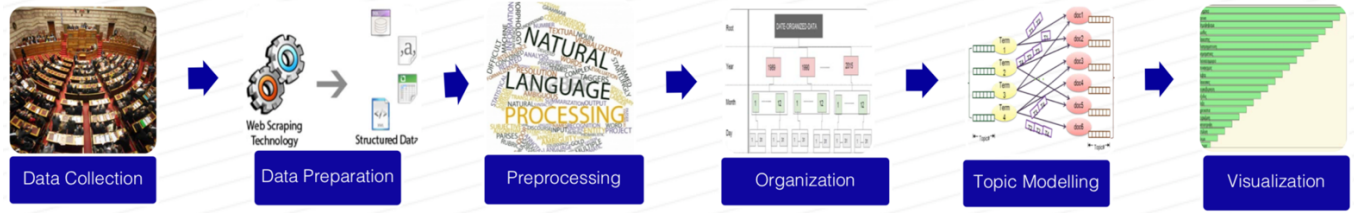
Parliamentary corpora as shown in CLARIN

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



Encoding and analysis of parliamentary data

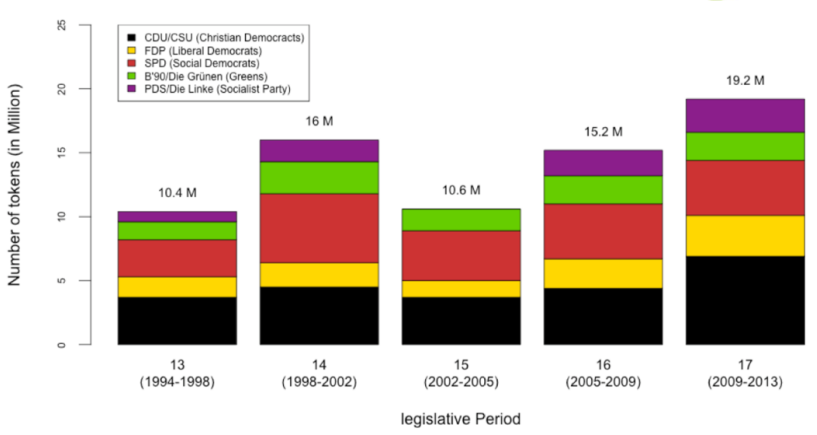
- format, quality and structure of the source files are essential
- traditionally, only transcriptions of parliamentary sessions have been made available; now increasingly released in audio and video as well
- key issues in the encoding of parliamentary data
  1. Source files
  2. Structural elements
  3. Encoding standards
  4. Metadata
  5. Linguistic annotations
  6. Text enrichment
- example analysis workflow



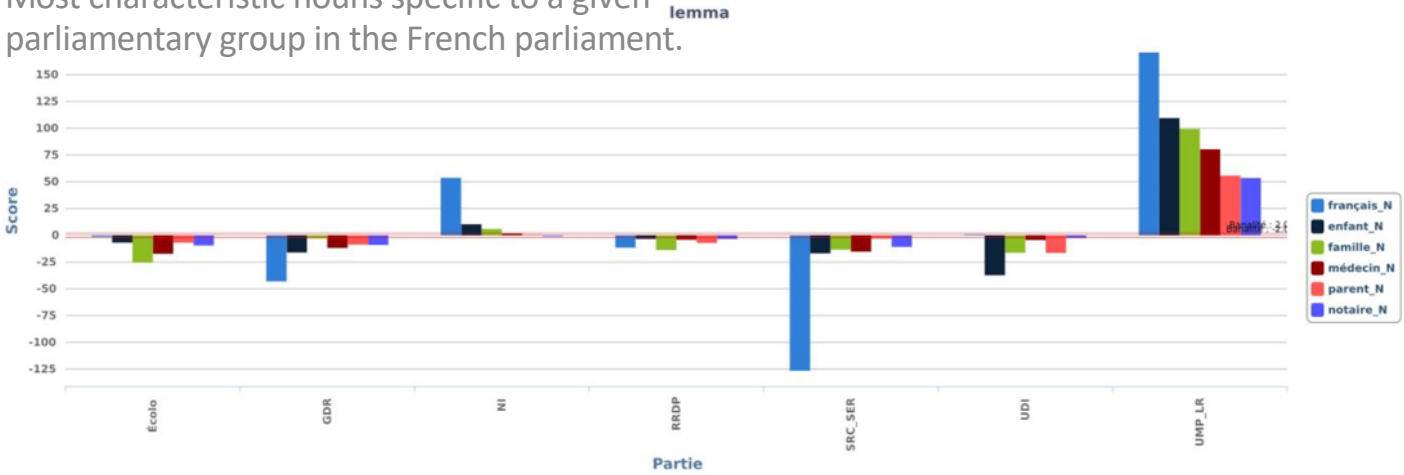
Metadata already tell a story

- Parliamentary records are rich in extralinguistic markup/metadata
  - on the speech level
    - parliamentary session
    - date
    - meeting item
    - speaking time
  - on the speaker level
    - speaker's name
    - date of birth
    - gender, education
    - party affiliation)
- that can be used as variables in analysis or for fine-grained filtering of the research data

Text production in the German parliament over time by political party



Most characteristic nouns specific to a given parliamentary group in the French parliament.



Representation and text production of female members of the Danish parliament

Age	% Female MPs	% Female speeches
20-29	56.1	52.2
30-39	47.9	47.0
40-49	42.7	35.4
50-59	29.8	22.0
60-69	24.7	15.0
70-79	33.3	30.8
Total	38.7	34.4

Case study: 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

Links CLARIN activities and publications for parliamentary data

- Workshop 2017: https://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records
- ParlaCLARIN workshop I @LREC 2018: https://www.clarin.eu/ParlaCLARIN
- ParlaFormat: https://www.clarin.eu/event/2019/parlaformat-workshop
- ParlaCLARIN workshop II @LREC 2020: https://www.clarin.eu/ParlaCLARIN-II
- ParlaMint project (2020-2021): https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora
- CLARIN Café on ParlaMint: https://www.clarin.eu/event/2020/clarin-cafe-join-our-parliamentary-flavoured-coffee-parlamint

CLARIN Resource Family: CMC corpora from Italy

- **DIDI - The DiDi Corpus of South Tyrolean CMC 1.0.0** (Facebook comments in English, German, Italian, and the minority language "Ladino")
- **PAISÀ Corpus** (Italian web text)

Aim DIDI: to provide scientific answers to current issues of language and education policy as well as to economic and social questions at both the local and international level" for South Tyrol, where "multilingualism involves geographic, institutional, social and personal aspects".

CLARIN Resource Family: Learner corpora from Italy

- **Italian PAROLE corpus**, in TalkBank

Aim: the PAROLE corpus presents storytelling-type tasks involving 23 learners of Italian (all French-L1; average age 19), and the transcriptions are marked up for prosodic phenomena relevant for L2-language learning ("unbroken sequences of hesitation phenomena, such as silent pauses, filled pauses, and certain paralinguistic noises")

- **MERLIN Written Learner Corpus** for Czech, German, Italian

Aim: support language learning for speakers of Italian as an L1 language

CLARIN as Knowledge Infrastucture 1

Promotion of data harmonization and methodology development

- Activities related to CLARIN Resource Families
- TwinTalk workshops at DH events
- CLARIN Hackaton on the detection of misinformation related to Covid-19:

Coordination and support of collaborative work with potential for societal impact

- Workshops and webinars aimed at a the devopment of a support chain for creating and processing oral history data.

  Impact: providing instruments for minorities to voice their perspective on society
- Workshop series on how to share and deposit speech with communication disorders (aka DELAD).
  Impact: speech pathology, early signalling of Alzheimer disease

CLARIN as Knowledge Infrastucture 2:

- Aims
  - increase the visibility of the national consortia and centres
  - reveal the richness of the CLARIN landscape
  - display the full range of activities throughout the network

- Highlights per chapter
  - consortium / centre
  - resource
  - tool
  - UI event
  - researcher

- Links:
  - https://www.clarin.eu/Tour-de-CLARIN
  - https://www.clarin.eu/sites/default/files/tour-de-clarin-italy.pdf (vol II)

Part IV
Concluding remarks

Open Science and multidisciplinarity

Language as a relevant resource for all SSH disciplines

- *Social survey data*: interviews, questionnaires
- *Social media data*: content + metadata
- *Psychology data*: oral history, interviews, ego documents
- *Humanities data*: archives, libraries, web content
- *Digital cultural heritage*: free text metadata, content related to artefacts, oral reports documenting excavations, etc.

European research infrastructures rooted in SSH disciplines: CESSDA, CLARIN, DARIAH, ESS, SHARE, ERIHS;
all involved in collaborative projects (H2020 funding) towards better service alignment

- SSHOC (=SSH Open Cloud): www.sshopencloud.eu
  SSH cluster activities aiming at integration of practices and services into the European Open Science Cloud
- TRIPLE: www.triple.eu
  innovative platform and practices for multidisciplinary exploration of SSH artefacts

CLARIN TwinTalk workshops at DH events

Dynamics in Research Assessment: DORA

San Francisco Declaration on Research Assessment (DORA)

- Promotes recognition of the need to improve the ways in which researchers and the outputs of scholarly research are evaluated
- https://sfdora.org/
- >2000 organisations signed
- >600 times signed from Italy

Recognition aspects

- Publications in other channels than journals
- Data publications
- Multiliguality
- Impact beyond research

Links

- Previous events: https://www.clarin.eu/events
- Funding for mobility, development of training, CRF projects, event organisation: www.clarin.eu/funding
- Monthly Newsletter: link
- Teaching and training materials:
  - Recorded lectures: http://videolectures.net/clarin/
  - DH Course Registry: https://dhcr.clarin-dariah.eu/
  - more to come soon
- CLARIN2021 – Annual Conference, 27-29 September 2021
  (PC chair: Monica Monachini, Institute of Computational Linguistics, Pisa)
  https://www.clarin.eu/event/2021/clarin-annual-conference-2021

see you @

www.clarin.eu
or
f.m.g.dejong@uu.nl

Acknowledgement

Several colleagues generously provided ideas and slides for this presentation:
Darja Fiser, Francesca Frontini, Jakob Leonardic, Dieter Van Uytvanck