The Wild West of Research Infrastructures for Language Research

Darja Fišer

EFNIL'23
13 October 2023

Ljubljana, Slovenia

CLARIN

CLARIN in eight bullets

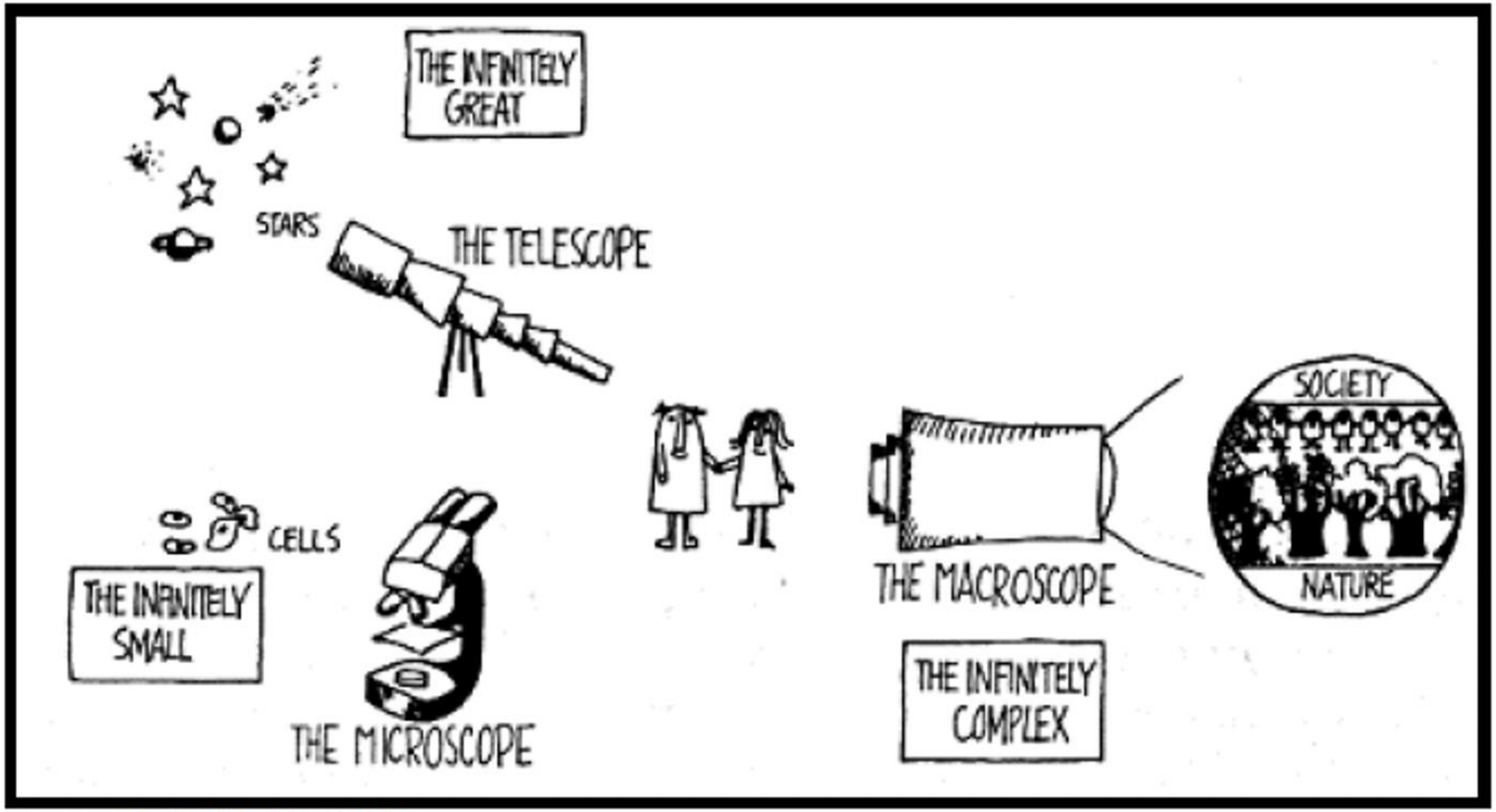
- Common Language Resources and Technology Infrastructure
- **ESFRI** roadmap 2006, **ESFRI** ERIC status 2012, Landmark 2016
- Easy and sustainable access for scholars in SSH
    - **digital language data** (written, spoken, video or multimodal)
    - **tools** to discover, analyse, combine data wherever they are located
    - **single sign-on** environment **(you all can get an account)**
- Ecosystem for **knowledge exchange**
- Some services **integrated in EOSC**

CLARIN and Open Science

- Promotion of sharing & re-use of language data through sustainable data registries
- Adherence to FAIR data principles
    - **F**indable
    - **A**ccessible
    - **I**nteroperable
    - **R**e-usable
- Enhancement & deployment of interoperability of language data & services
    - common metadata framework
    - distributed network of FAIR certified data repositories for language data

- Promotion of
    - comparative perspectives
    - multidisciplinary collaboration
    - transnational research
    - responsible data science
- Support for linguistic diversity
    - data covering many languages
    - tools for many languages
    - language resources in all modalities
    - discipline- & language-agnostic

Full data interoperability: the macroscope



Fonte: (ROSNAY, 1979)

CLARIN in countries and centres

A consortium of type ERIC

- 23 members
- 3 observers
- 1 linked party

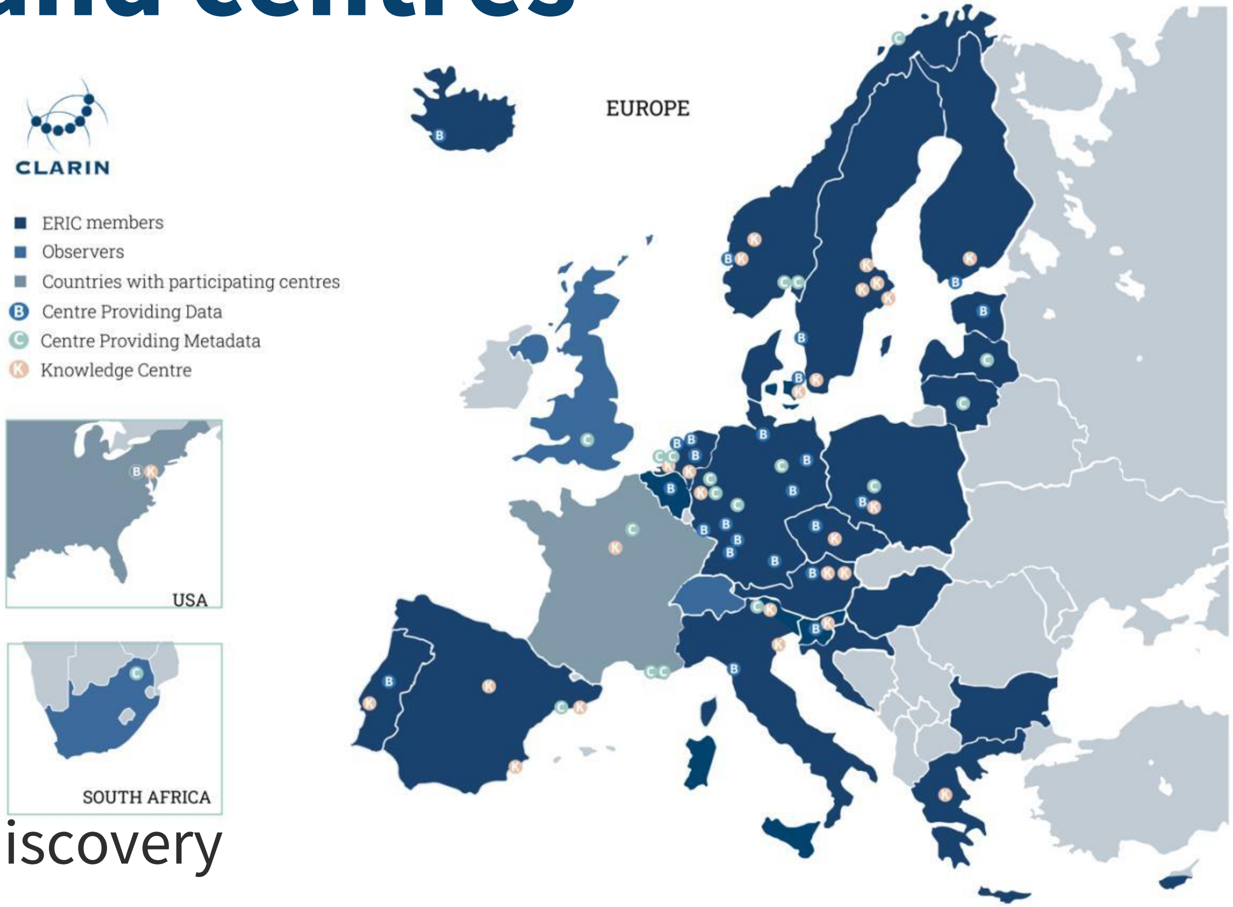
A distributed network of 70 centres

21 CTS certified data centres

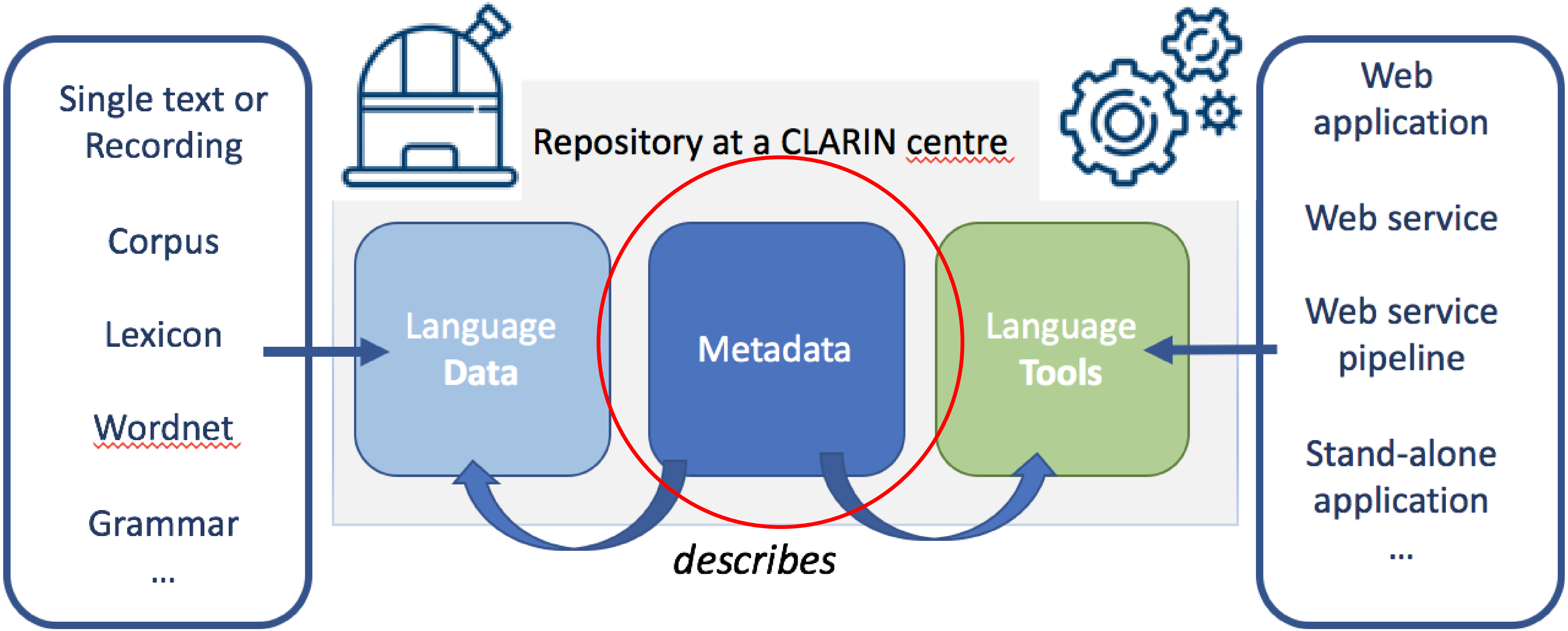
strong focus on FAIRness & interoperability

- federated login
- central metadata harvesting for easy discovery
- chained services

25 Knowledge Centres (e.g. CLASSLA, Knowledge centre for South Slavic Languages)



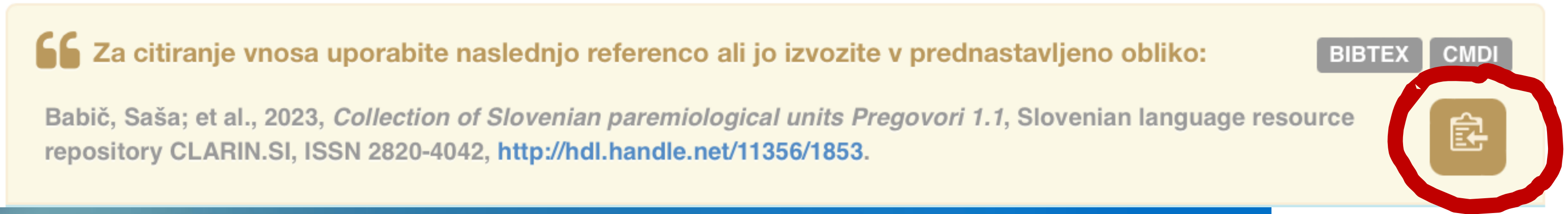
The CLARIN data architecture:
a network of distributed centre repositories



Virtual Language Observatory

https://vlo.clarin.eu

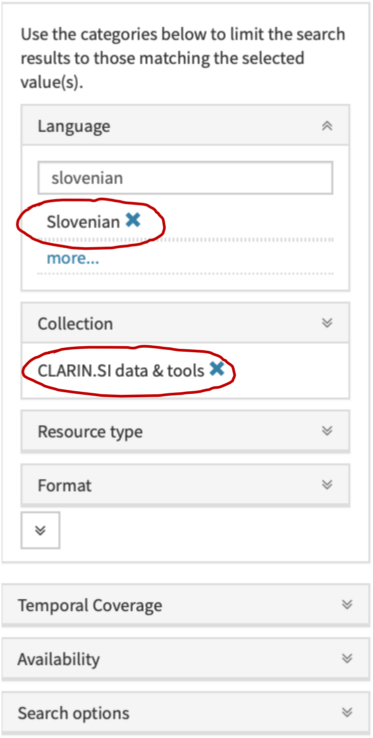
- facet search
- links to landing pages
- download options
- details on licences
- details on technical features
- overview of tools that match the data
- info on how to cite:



Za citiranje vnosa uporabite naslednjo referenco ali jo izvozite v prednastavljeno obliko:

BIBTEX CMDI

Babič, Saša; et al., 2023, *Collection of Slovenian paremiological units Pregovori 1.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1853.



Use the categories below to limit the search results to those matching the selected value(s).

Language

slovenian

Slovenian ✖

more...

Collection

CLARIN.SI data & tools ✖

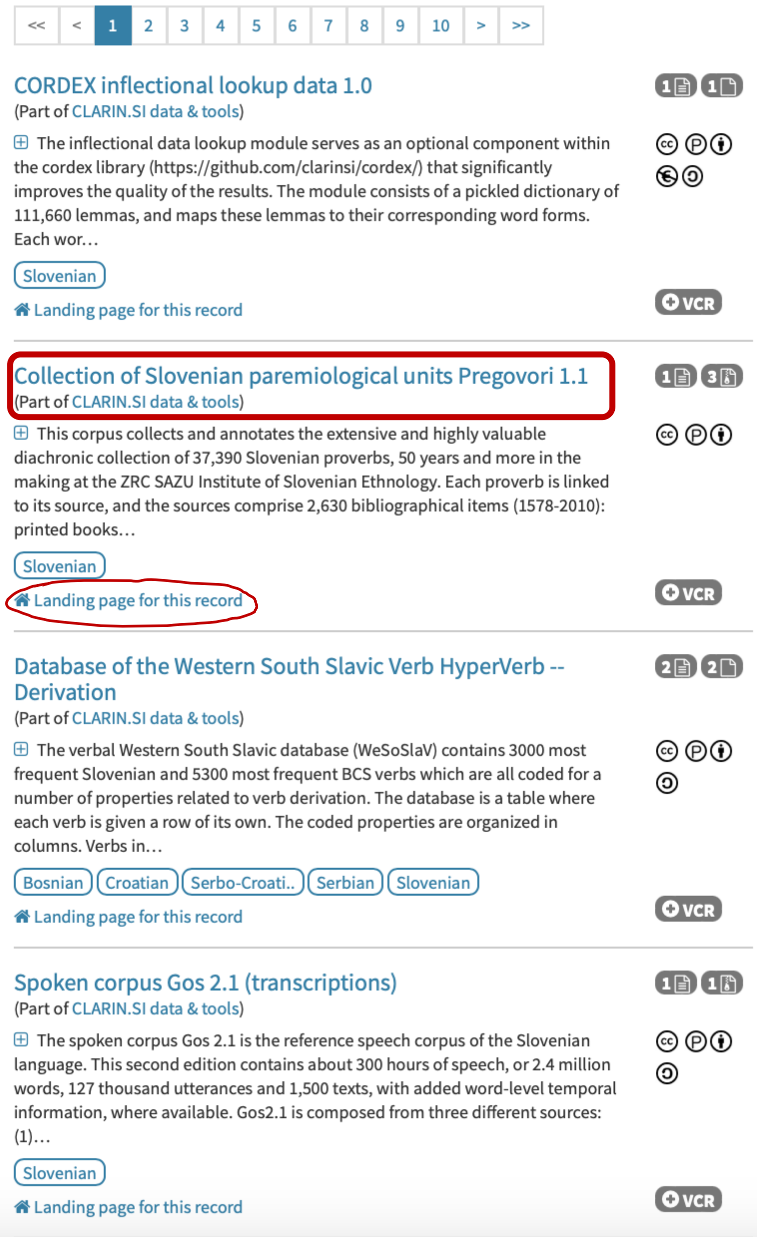
Resource type

Format

Temporal Coverage

Availability

Search options



CORDEX inflectional lookup data 1.0
(Part of CLARIN.SI data & tools)

⊞ The inflectional data lookup module serves as an optional component within the cordex library (https://github.com/clarinsi/cordex/) that significantly improves the quality of the results. The module consists of a pickled dictionary of 111,660 lemmas, and maps these lemmas to their corresponding word forms. Each wor…

Slovenian

Landing page for this record

VCR

Collection of Slovenian paremiological units Pregovori 1.1
(Part of CLARIN.SI data & tools)

⊞ This corpus collects and annotates the extensive and highly valuable diachronic collection of 37,390 Slovenian proverbs, 50 years and more in the making at the ZRC SAZU Institute of Slovenian Ethnology. Each proverb is linked to its source, and the sources comprise 2,630 bibliographical items (1578-2010): printed books…

Slovenian

Landing page for this record

VCR

Database of the Western South Slavic Verb HyperVerb -- Derivation
(Part of CLARIN.SI data & tools)

⊞ The verbal Western South Slavic database (WeSoSlaV) contains 3000 most frequent Slovenian and 5300 most frequent BCS verbs which are all coded for a number of properties related to verb derivation. The database is a table where each verb is given a row of its own. The coded properties are organized in columns. Verbs in…

Bosnian Croatian Serbo-Croati.. Serbian Slovenian

Landing page for this record

VCR

Spoken corpus Gos 2.1 (transcriptions)
(Part of CLARIN.SI data & tools)

⊞ The spoken corpus Gos 2.1 is the reference speech corpus of the Slovenian language. This second edition contains about 300 hours of speech, or 2.4 million words, 127 thousand utterances and 1,500 texts, with added word-level temporal information, where available. Gos2.1 is composed from three different sources: (1)…

Slovenian

Landing page for this record

VCR

ParlaMint

ParlaMint II releases

Tomaž Erjavec et al. (2023)

Multilingual comparable corpora of parliamentary debates ParlaMint 3.0

http://hdl.handle.net/11356/1486

Tomaž Erjavec et al. (2023)

Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0

http://hdl.handle.net/11356/1488

Taja Kuzman et al. (2023)

Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 3.0

http://hdl.handle.net/11356/1810

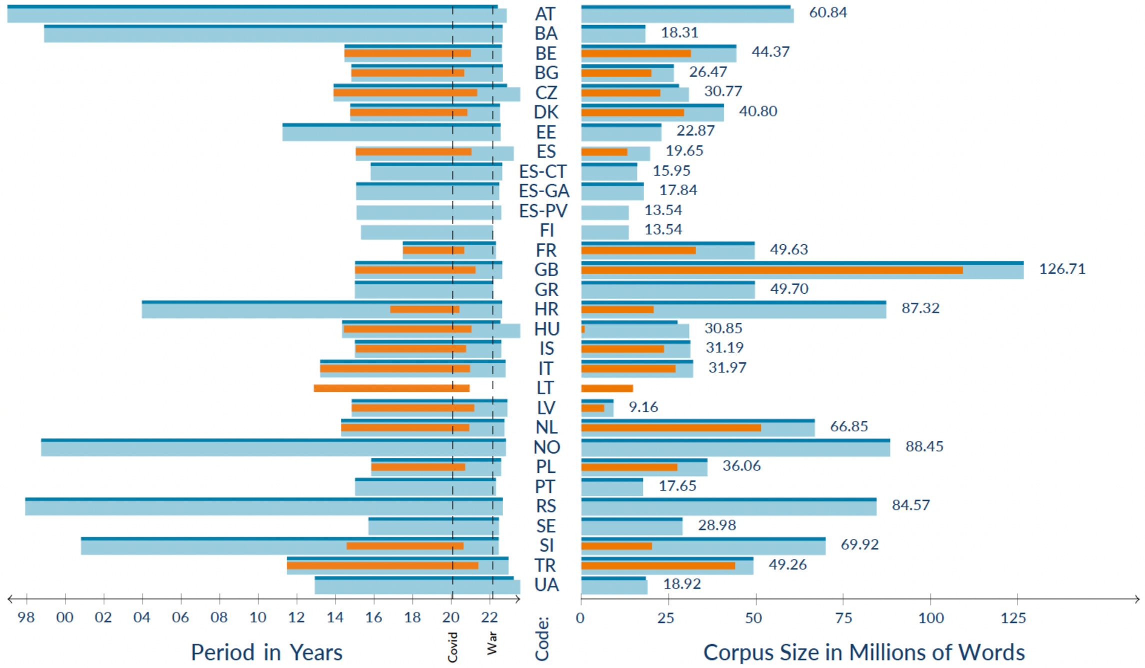
Nikola Ljubešić et al. (2022)

ASR training dataset for Croatian ParlaSpeech-HR v1.0

http://hdl.handle.net/11356/1494

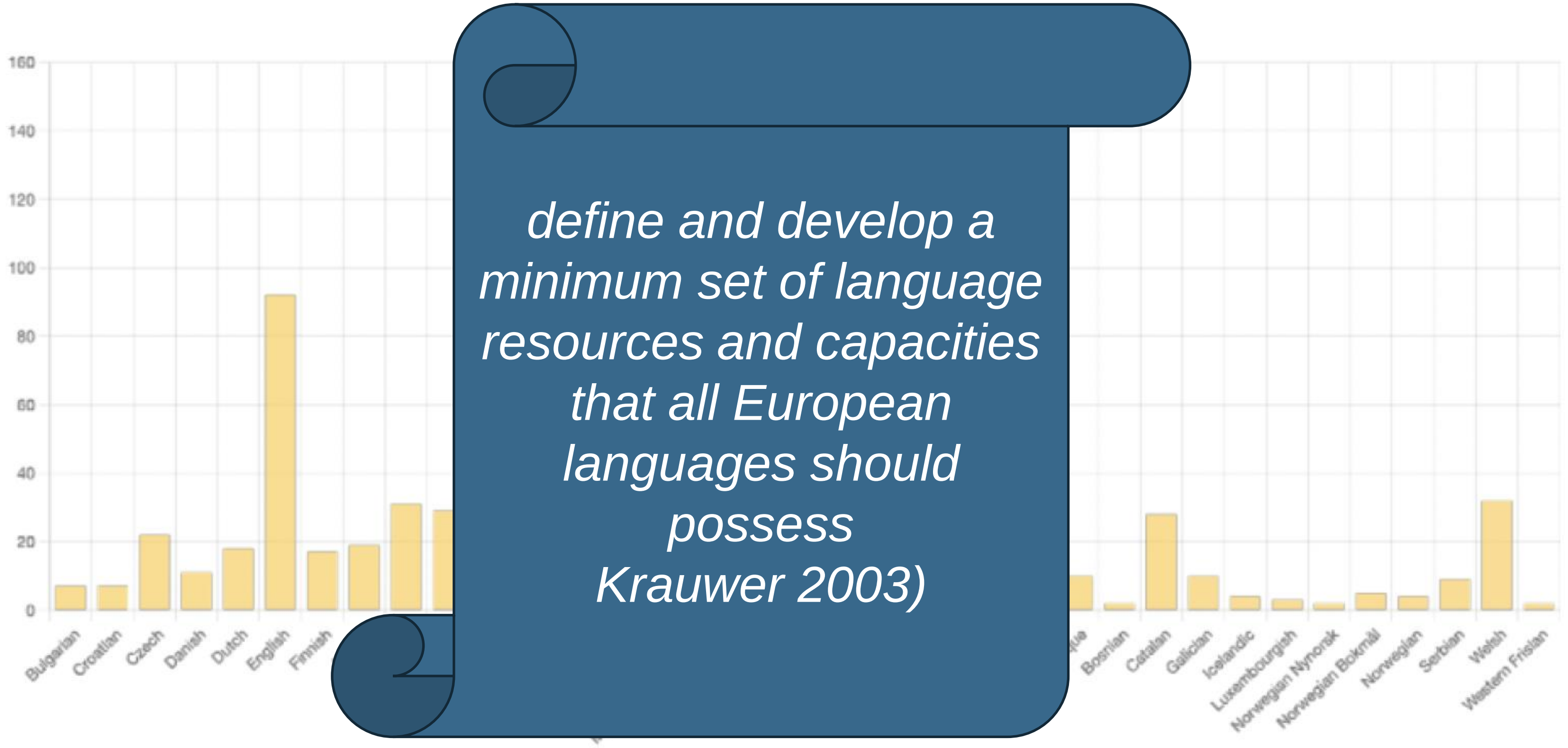
ParlaMint corpora

- ParlaMint I (v2.1)
- ParlaMint II (v3.0)
- ParlaMint II (v4.0)



Data Support

- data gaps: large language models, multimodal data, sign-language video data, domain-specific data



define and develop a minimum set of language resources and capacities that all European languages should possess
Krauwer 2003)

Fig. 2 Number of language models available in the catalogue of the European Language Grid for the EU official languages and for some indicative non-EU official ones (as of January 2023)

(Giagkou et al. 2023)

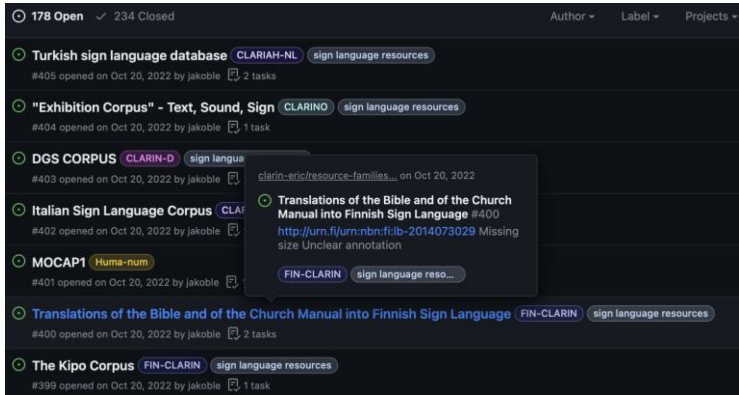
Life isn‘t always FAIR

- *Findable*
- *Accessible*
- *Interoperable*
- *Re-usable*

Surveyed 47 bilingual and 39 multilingual parallel corpora

- 38% cannot be found in the VLO
- 6% available for download & querying
- 12% completely unavailable
- 13% no info on size
- 31% no info on level of alignment
- 7% no licence info
- 12% not in CLARIN

(Fišer and Lenardič 2020)



Turkish sign language database CLARIAH-NL sign language resources
#405 opened on Oct 20, 2022 by jakoble 2 tasks

"Exhibition Corpus" - Text, Sound, Sign CLARINO sign language resources
#404 opened on Oct 20, 2022 by jakoble 1 task

DGS CORPUS CLARIN-D sign langua
#403 opened on Oct 20, 2022 by jakoble

Italian Sign Language Corpus CLAI
#402 opened on Oct 20, 2022 by jakoble

MOCAP1 Huma-num
#401 opened on Oct 20, 2022 by jakoble

clarin-eric/resource-families... on Oct 20, 2022

Translations of the Bible and of the Church Manual into Finnish Sign Language #400
http://urn.fi/urn:nbn:fi:lb-2014073029 Missing size Unclear annotation
FIN-CLARIN sign language reso...

Translations of the Bible and of the Church Manual into Finnish Sign Language FIN-CLARIN sign language resources
#400 opened on Oct 20, 2022 by jakoble 2 tasks

The Kipo Corpus FIN-CLARIN sign language resources
#399 opened on Oct 20, 2022 by jakoble 1 task

Credit when credit's due

- Surveyed the landscape of linguistic data citation in 6 leading Slovene journals and in proceedings of 2 conferences focused on linguistics for 2013-2019 (Lenardič et al. 2020)
- 26% of the papers involve the use of language resources
- The resources deposited in a data catalogue virtually always incorrectly/incompletely cited
- Citation practices largely dependent on the instructions for authors of the particular publication
- Broader adoption of the Austin principles needed by publishers (Berez-Kroeker et al. 2017)
- Recognition of effort for resource and technology development needed within COARA Coalition

Close the loop

Deposit your own

- *ELG*
- *ELRC-SHARE*
- *ELRA*
- *CLARIN*
- *META-SHARE*
- *LCD*

- 125 existing but still uncatalogued corpora
  - 60% in LREC2020 (10% of proceedings)
  - 17% in LRE journal 2016-2020 (22 issues)
  - 23% corpora found through other channels
- 68% belong to existing CLARIN Resource Families
- 52% available for download (GitHub or personal websites)
- 13% available through online concordancers (corpus-specific)
- 8% available upon request
- 6% still in preparation at the time of publication of the paper
- 20% unclear availability

(Lenardič and Fišer 2022)

If You Deposit, You Better Deposit Very Well. When You Have To Cite, Cite.

Questions & comments
on how to better support you and your work
welcome!