Co-funded by the
Erasmus+ Programme
of the European Union

Integrating Research & Research Infrastructures into Teaching

Marko Simonović (Uni Graz)
Iulianna van der Lek (CLARIN ERIC)

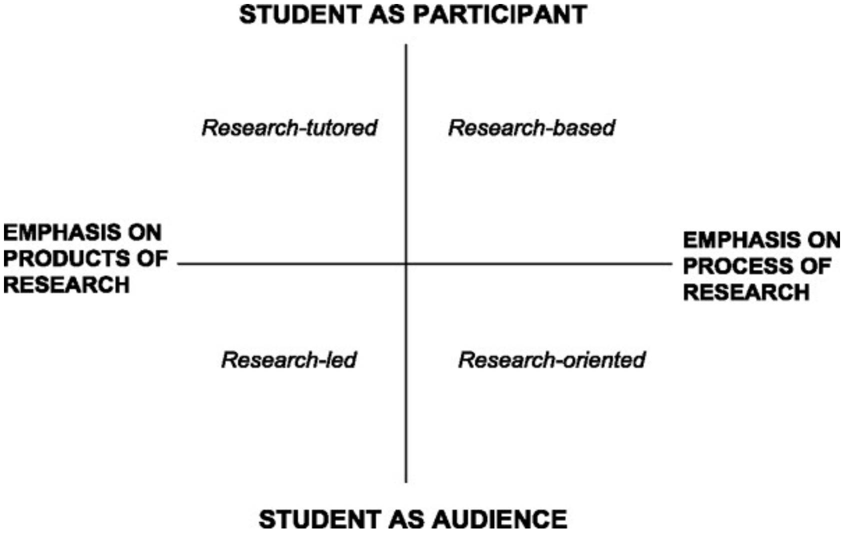
Graz Multiplier Event, 5 July 2022

Overview

Research-Based Teaching: Guidelines and Best Practices

Integrating Research Infrastructures into Teaching: The Case of CLARIN

What is research-based teaching?



STUDENT AS PARTICIPANT

Research-tutored

Research-based

EMPHASIS ON PRODUCTS OF RESEARCH

EMPHASIS ON PROCESS OF RESEARCH

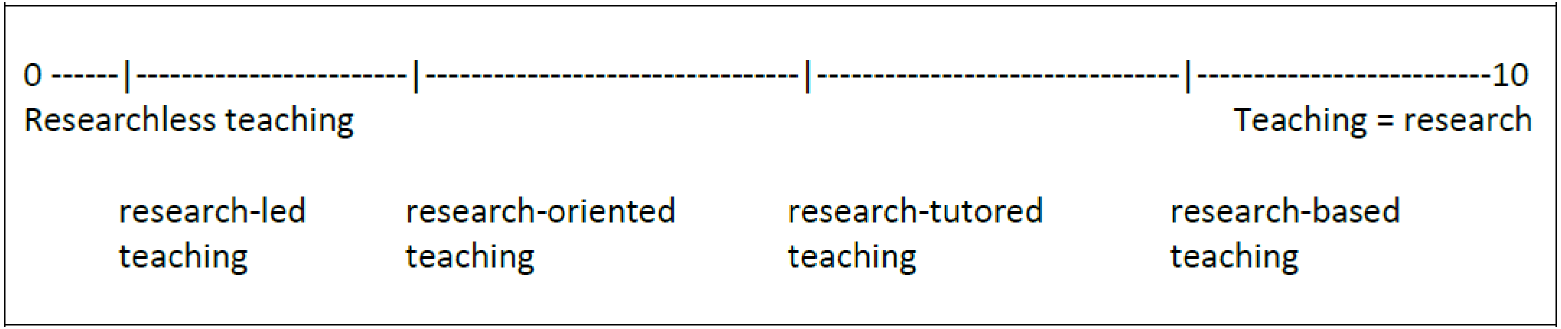
Research-led

Research-oriented

STUDENT AS AUDIENCE

Healey’s (2005, Jenkins et. al 2007) adapted model of the research-teaching nexus (from Visser-Wijnveen et al. 2010).

What is research-based teaching?



0 ------|------------------------|---------------------------------|--------------------------------|--------------------------10

Researchless teaching

Teaching = research

research-led teaching

research-oriented teaching

research-tutored teaching

research-based teaching

Dekker & Wolf (2016)'s scale of the research-teaching nexus.

Why research-based teaching?

Advantages for for students

- direct **preparation for a research career**, but also
- stimulate **curiosity**
- practice **problem solving**
- practice **forward thinking**
- gain **independence**
- gain **experience**

Advantages for lecturers

- integrate **research and teaching**
- **new insights** about own research
- **more interaction** with students

Why research-based teaching?

Advantages for for students

- direct **preparation for a**
- stimulate **curiosity**
- practice **problem s**
- practice **forward thi**
- gain **independence**
- gain **experience**

Isn't this what university lecturers do all the time?

Advantages for lecturers

- integrate **research and teaching**
- **new insights** about own research
- **more interaction** with students

Why research-based teaching?

Advantages for for students

- direct **preparation for a**
- stimulate **curiosity**
- practice **problem s**
- practice **forward th**
- gain **independence**
- gain **experience**

Isn't this what university lecturers do all the time?

Advantages for lecturers

- integrate **research and teaching**
- **new insights** about own research
- **more interaction** with students

Unfortunately not!

Research-teaching nexus

Teaching and **research** are typically planned, performed and evaluated separately.

The connection between them, usually discussed in the literature under the rubric of the **research-teaching nexus**, receives little attention in the reality of most academic institutions (for a notable exception, and a rich collection of experiences and ideas, see Bastiaens, van Tilburg & van Merriënboer 2017).

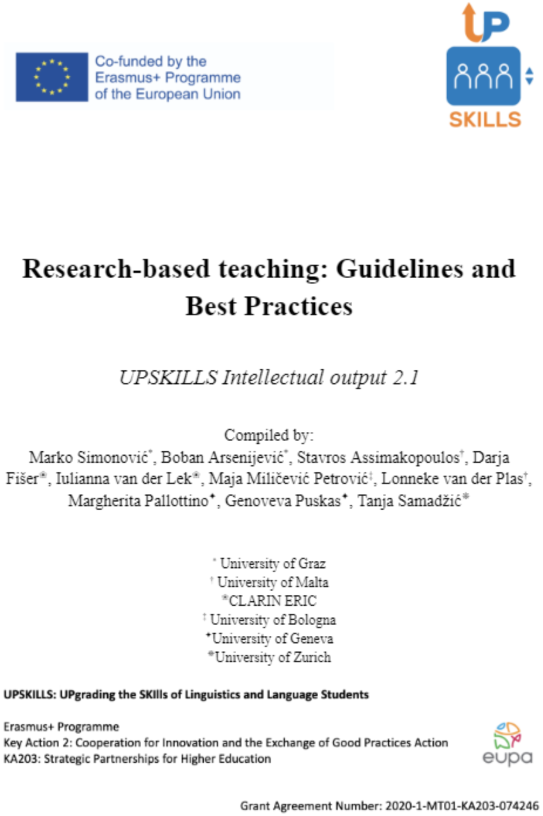
Visser-Wijnveen (2009: 141): “academics’ conceptions of the research-teaching nexus are related to their conceptions of teaching and **not to their conceptions of research and knowledge**”.

Why we don’t do more research-based teaching?

- The results of first research-based courses may be viewed as unsatisfactory by the lecturer.
    - Course design turns out to be time consuming.
    - Teaching on own research turns out to be more challenging than originally assumed.

Still, as transpired from the survey we held within the preliminary needs analysis of our project, a bit over 50% of lecturers indicated that they integrate research into teaching, while around 75% indicated that they would be happy to follow dedicated training focusing on this.

Filling the gap: Our guidelines and best practices



Co-funded by the Erasmus+ Programme of the European Union

SKILLS

Research-based teaching: Guidelines and Best Practices

UPSKILLS Intellectual output 2.1

Compiled by:
Marko Simonović[*], Boban Arsenijević[*], Stavros Assimakopoulos[†], Darja Fišer[⁎], Iulianna van der Lek[⁎], Maja Miličević Petrović[‡], Lonneke van der Plas[†], Margherita Pallottino[✦], Genoveva Puskas[✦], Tanja Samadžić[⁑]

[*] University of Graz
[†] University of Malta
[⁎] CLARIN ERIC
[‡] University of Bologna
[✦] University of Geneva
[⁑] University of Zurich

UPSKILLS: UPgrading the SKIlls of Linguistics and Language Students

Erasmus+ Programme
Key Action 2: Cooperation for Innovation and the Exchange of Good Practices Action
KA203: Strategic Partnerships for Higher Education

eupa

Grant Agreement Number: 2020-1-MT01-KA203-074246

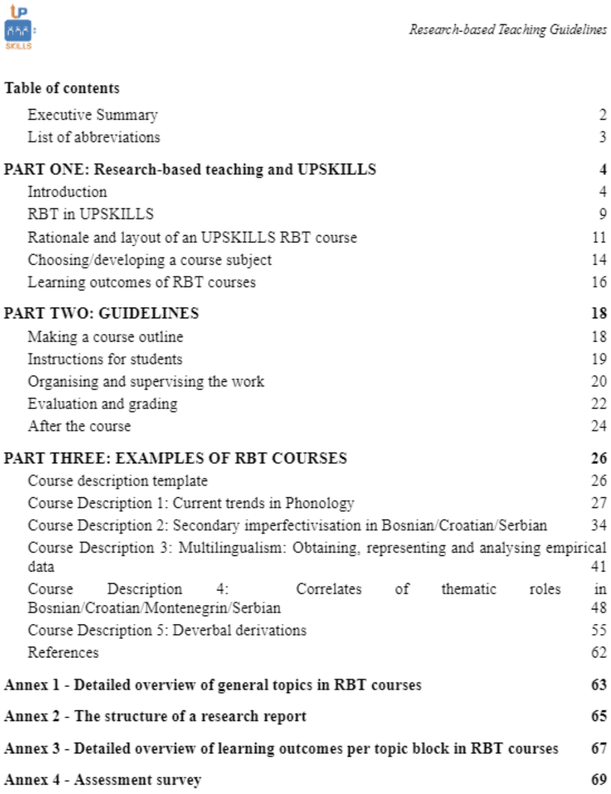


Table of contents

Guidelines for integrating research into teaching

Featuring:

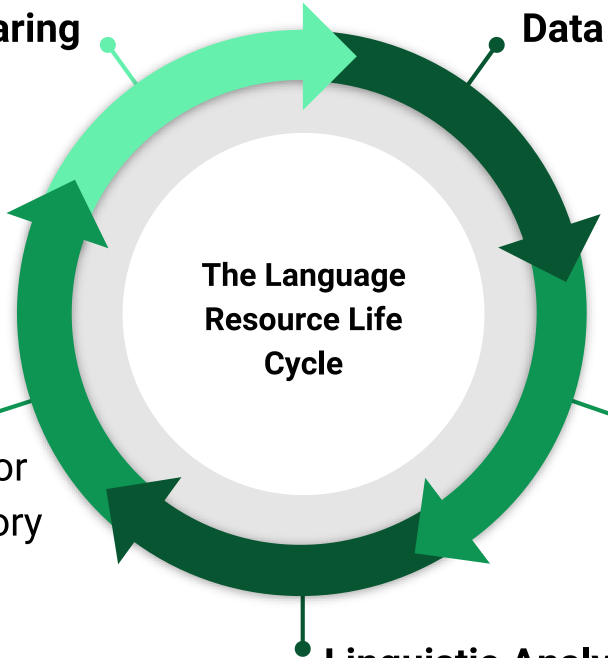
- Instructions for choosing/developing a course subject
- Detailed pick-and-choose list of research-related topics
- Detailed pick-and-choose list of research-related learning outcomes
- Making instructions for students
- Organising and supervising the work
- Evaluation and grading
- After the course
- Course description template
- Course Examples
  - Current trends in Phonology
  - Multilingualism
  - Deverbal nominalisation in West South Slavic
- Annexes
  - The structure of a research report
  - Assessment survey

Integrating Research Infrastructures into Teaching

Iulianna van der Lek

The Case of CLARIN

The Language Resource Life Cycle



Data Acquisition and Collection

- Research Infrastructures
- Repositories
- Data Catalogues
- IPR & Legal Issues
- Citation

Data Curation and Annotation

- Standards and formats
- Annotation: (semi) automatic, manual

Linguistic Analysis and Research

- Querying metadata, LRs and their annotations
- Analysis & visualisation; combining data from different sources -> exchange standards

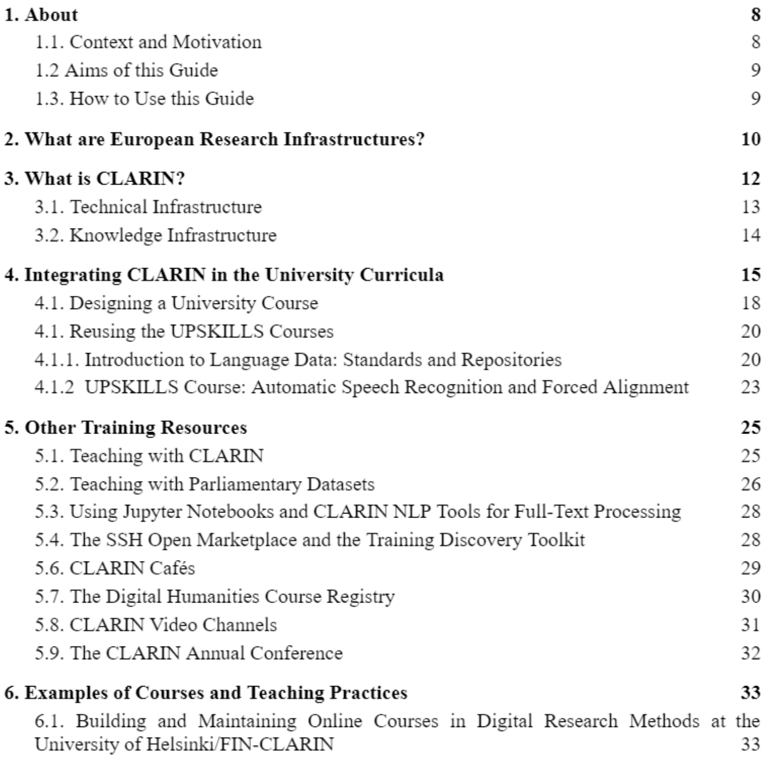
Data Archiving

- Deposit LRs in a institutional or domain-specific FAIR repository
- Describe LRs using metadata standards
- IPR and legal issues for handling sensitive data; licenses

Data Sharing

- Repositories
- Catalogues
- Citation
- Legal Issues

A Guide and a Course for Teachers



Course: Introduction to Language Data: Standards & Repositories

- 4-5 ECTS

Unit 1: Intro to Language Resources and Infrastructures for Language Resources and Technologies

Unit 2: Finding, accessing and using LRs

Unit 3: Tools for Linguistic Processing

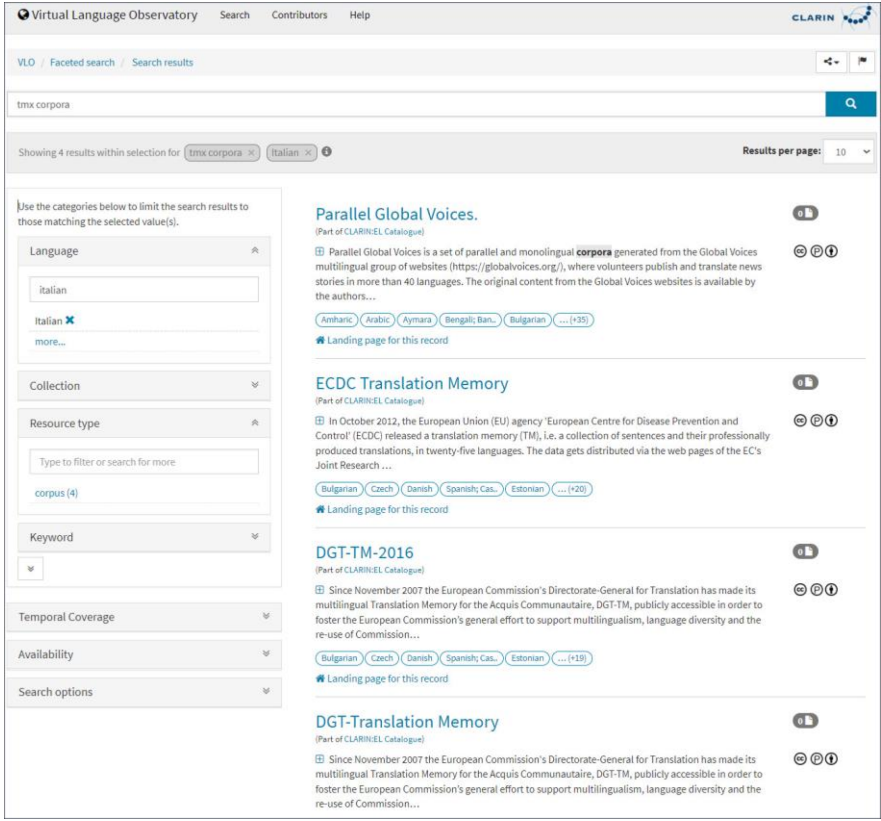
Unit 4: Archiving and Sharing LRs

Available for Piloting in September 2022

Finding Language Resources

CLARIN Virtual Language Observatory

- Metadata search in the CLARIN language repositories across Europe
- Does not contain data, only refers to it using PIDs
- Useful as a first step to identify which language resources are available



Showing 4 results within selection for tmx corpora × Italian ×

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

Parallel Global Voices.
(Part of CLARIN:EL Catalogue)
⊞ Parallel Global Voices is a set of parallel and monolingual **corpora** generated from the Global Voices multilingual group of websites (https://globalvoices.org/), where volunteers publish and translate news stories in more than 40 languages. The original content from the Global Voices websites is available by the authors…
Landing page for this record

ECDC Translation Memory
(Part of CLARIN:EL Catalogue)
⊞ In October 2012, the European Union (EU) agency 'European Centre for Disease Prevention and Control' (ECDC) released a translation memory (TM), i.e. a collection of sentences and their professionally produced translations, in twenty-five languages. The data gets distributed via the web pages of the EC's Joint Research …
Landing page for this record

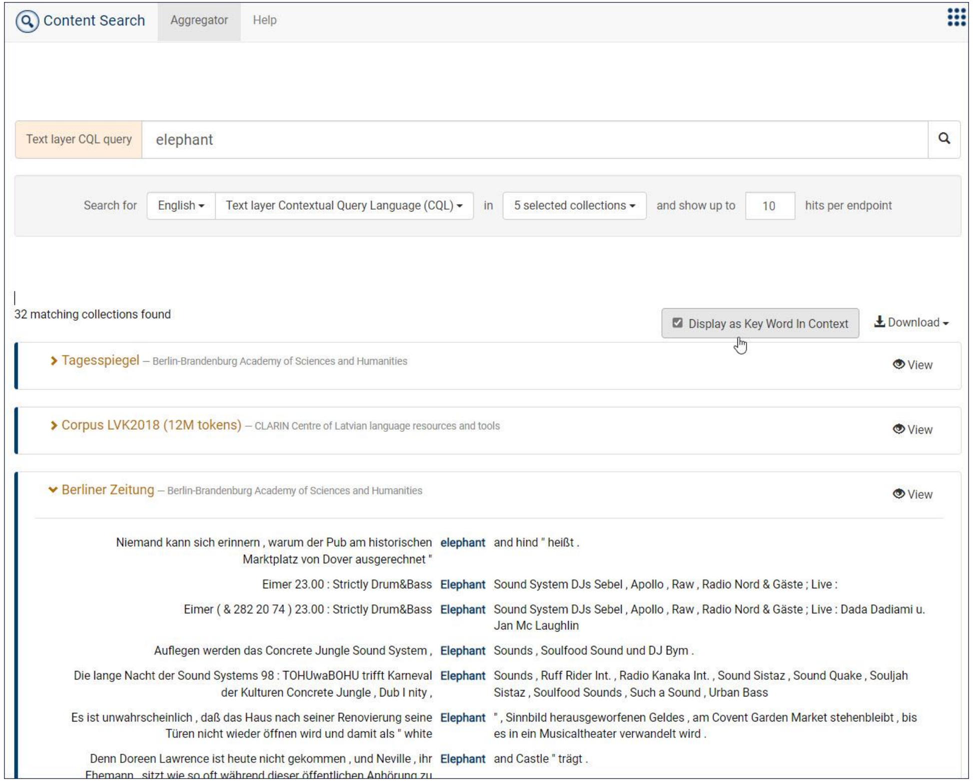
DGT-TM-2016
(Part of CLARIN:EL Catalogue)
⊞ Since November 2007 the European Commission's Directorate-General for Translation has made its multilingual Translation Memory for the Acquis Communautaire, DGT-TM, publicly accessible in order to foster the European Commission's general effort to support multilingualism, language diversity and the re-use of Commission…
Landing page for this record

DGT-Translation Memory
(Part of CLARIN:EL Catalogue)
⊞ Since November 2007 the European Commission's Directorate-General for Translation has made its multilingual Translation Memory for the Acquis Communautaire, DGT-TM, publicly accessible in order to foster the European Commission's general effort to support multilingualism, language diversity and the re-use of Commission…

Finding Language Resources

Content Search

- To search for specific patterns across corpora at the same time
- The search results from a specific corpus can be downloaded in various formats and processed in other tools for further analysis
- Possible to retrieve the citation information
- Useful as a first step to discover where interesting lg resources are hosted



Text layer CQL query: elephant

Search for English ▾ Text layer Contextual Query Language (CQL) ▾ in 5 selected collections ▾ and show up to 10 hits per endpoint

32 matching collections found

☑ Display as Key Word In Context

› Tagesspiegel – Berlin-Brandenburg Academy of Sciences and Humanities

› Corpus LVK2018 (12M tokens) – CLARIN Centre of Latvian language resources and tools

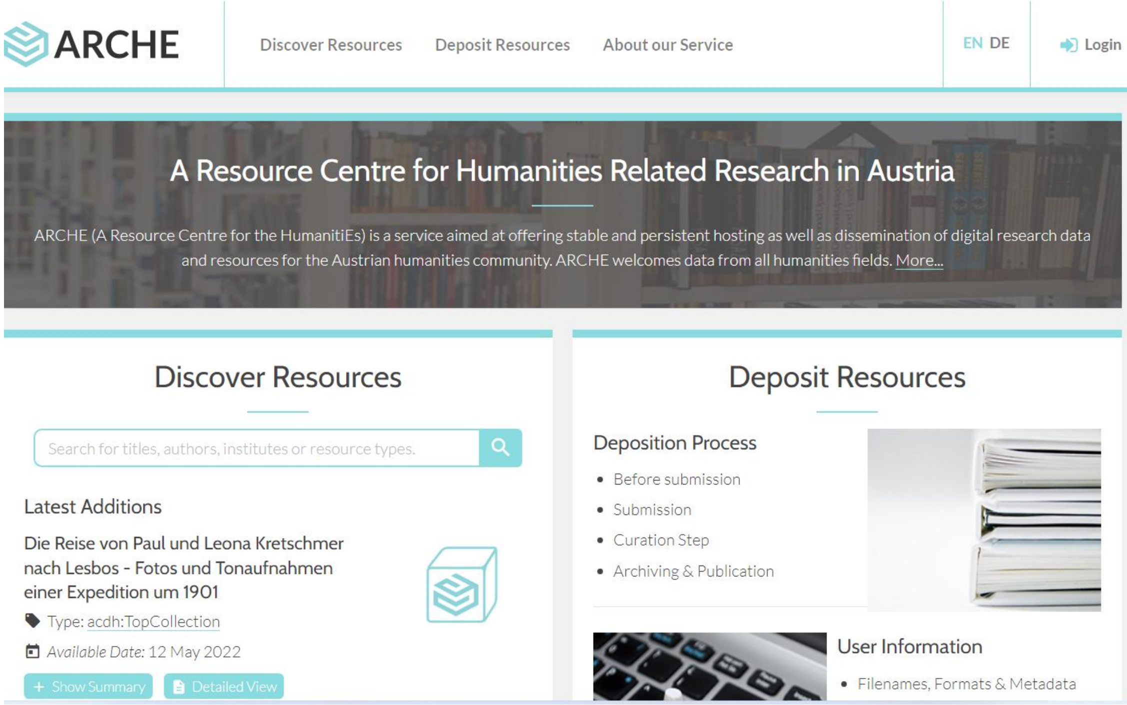
⌄ Berliner Zeitung – Berlin-Brandenburg Academy of Sciences and Humanities

Niemand kann sich erinnern , warum der Pub am historischen Marktplatz von Dover ausgerechnet "	elephant	and hind " heißt .
Eimer 23.00 : Strictly Drum&Bass	Elephant	Sound System DJs Sebel , Apollo , Raw , Radio Nord & Gäste ; Live :
Eimer (& 282 20 74) 23.00 : Strictly Drum&Bass	Elephant	Sound System DJs Sebel , Apollo , Raw , Radio Nord & Gäste ; Live : Dada Dadiami u. Jan Mc Laughlin
Auflegen werden das Concrete Jungle Sound System ,	Elephant	Sounds , Soulfood Sound und DJ Bym .
Die lange Nacht der Sound Systems 98 : TOHUwaBOHU trifft Karneval der Kulturen Concrete Jungle , Dub I nity ,	Elephant	Sounds , Ruff Rider Int. , Radio Kanaka Int. , Sound Sistaz , Sound Quake , Souljah Sistaz , Soulfood Sounds , Such a Sound , Urban Bass
Es ist unwahrscheinlich , daß das Haus nach seiner Renovierung seine Türen nicht wieder öffnen wird und damit als " white	Elephant	" , Sinnbild herausgeworfenen Geldes , am Covent Garden Market stehenbleibt , bis es in ein Musicaltheater verwandelt wird .
Denn Doreen Lawrence ist heute nicht gekommen , und Neville , ihr Ehemann , sitzt wie so oft während dieser öffentlichen Anhörung zu	Elephant	and Castle " trägt .

Finding Language Resources

Discipline-specific repositories

- Access to curated LRs
- Datasets can be cited via PIDs
- E.g. CLARIN B-Centres, The Open Language Archive,Meta-Share



A Resource Centre for Humanities Related Research in Austria

ARCHE (A Resource Centre for the HumanitiEs) is a service aimed at offering stable and persistent hosting as well as dissemination of digital research data and resources for the Austrian humanities community. ARCHE welcomes data from all humanities fields. More...

Discover Resources

Search for titles, authors, institutes or resource types.

Latest Additions

Die Reise von Paul und Leona Kretschmer nach Lesbos - Fotos und Tonaufnahmen einer Expedition um 1901

Type: acdh:TopCollection

Available Date: 12 May 2022

Deposit Resources

Deposition Process

- Before submission
- Submission
- Curation Step
- Archiving & Publication

User Information

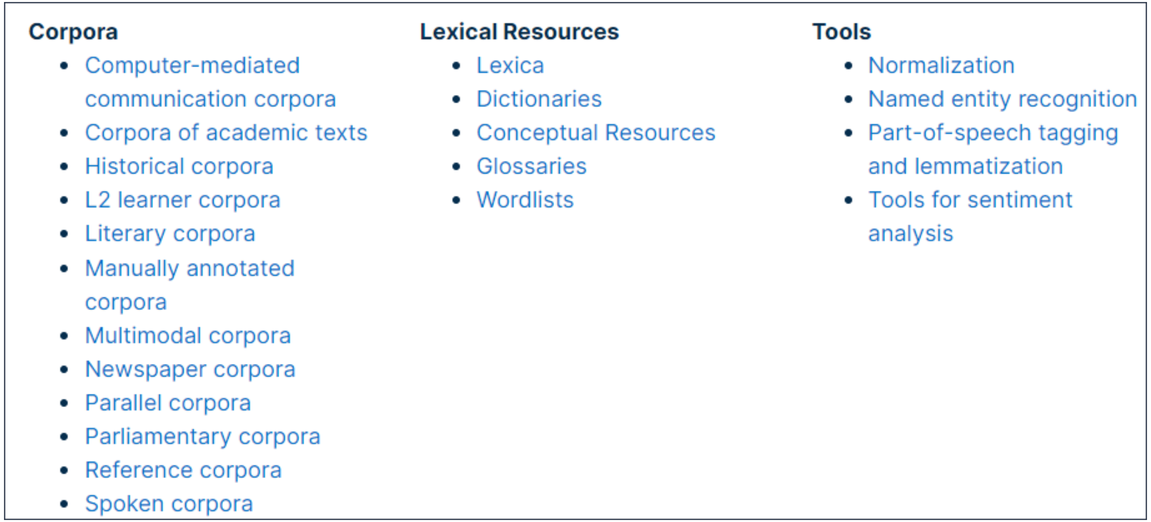
- Filenames, Formats & Metadata

https://arche.acdh.oeaw.ac.at/browser/

Find and Query Large Collections of Corpora

CLARIN Resource Families

- User-friendly overviews of well-curated corpora and tools
- Download or query with concordancers, e.g. Korp, Corpuscle and KonText



Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

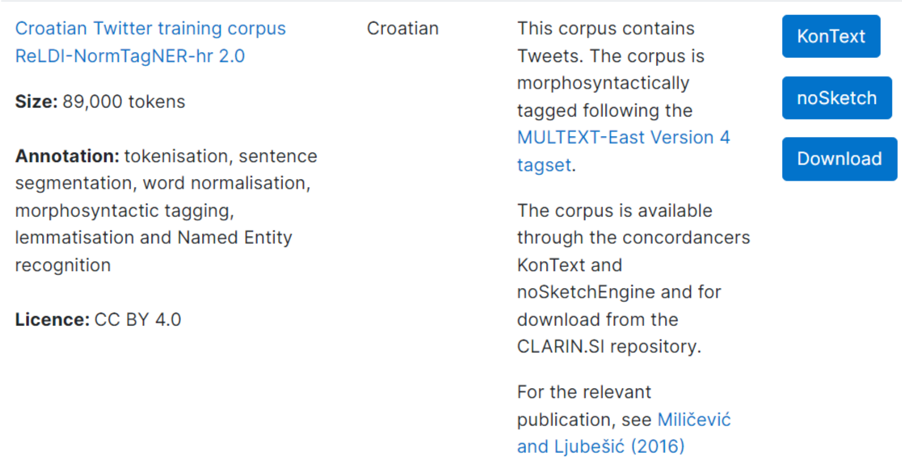
Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

Find and Query Large Collections of Corpora



Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0

Size: 89,000 tokens

Annotation: tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging, lemmatisation and Named Entity recognition

Licence: CC BY 4.0

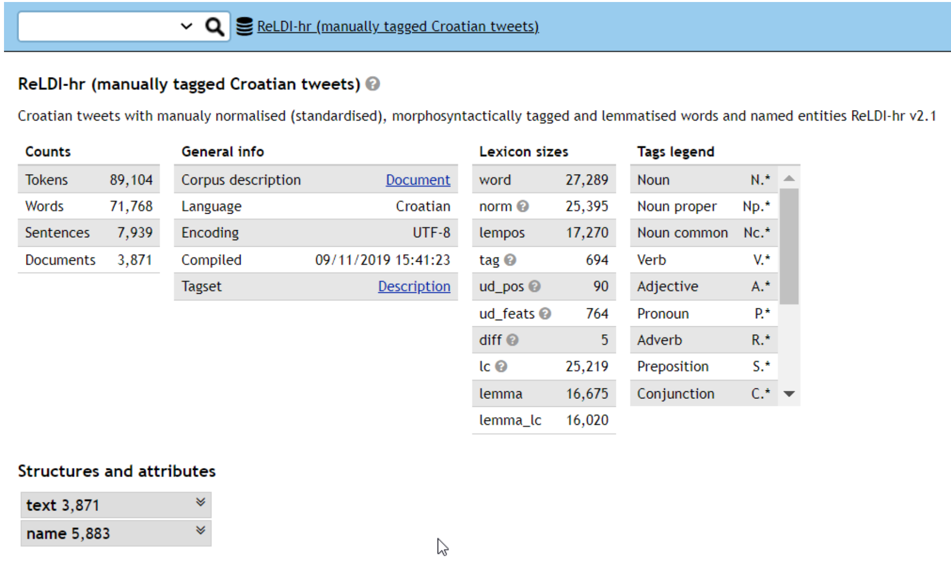
Croatian

This corpus contains Tweets. The corpus is morphosyntactically tagged following the MULTEXT-East Version 4 tagset.

The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.

For the relevant publication, see Miličević and Ljubešić (2016)

Query the ReLDI corpus via NoSketch engine on CLARIN.SI



ReLDI-hr (manually tagged Croatian tweets)

Croatian tweets with manualy normalised (standardised), morphosyntactically tagged and lemmatised words and named entities ReLDI-hr v2.1

Counts	
Tokens	89,104
Words	71,768
Sentences	7,939
Documents	3,871

General info	
Corpus description	Document
Language	Croatian
Encoding	UTF-8
Compiled	09/11/2019 15:41:23
Tagset	Description

Lexicon sizes	
word	27,289
norm	25,395
lempos	17,270
tag	694
ud_pos	90
ud_feats	764
diff	5
lc	25,219
lemma	16,675
lemma_lc	16,020

Tags legend	
Noun	N.*
Noun proper	Np.*
Noun common	Nc.*
Verb	V.*
Adjective	A.*
Pronoun	P.*
Adverb	R.*
Preposition	S.*
Conjunction	C.*

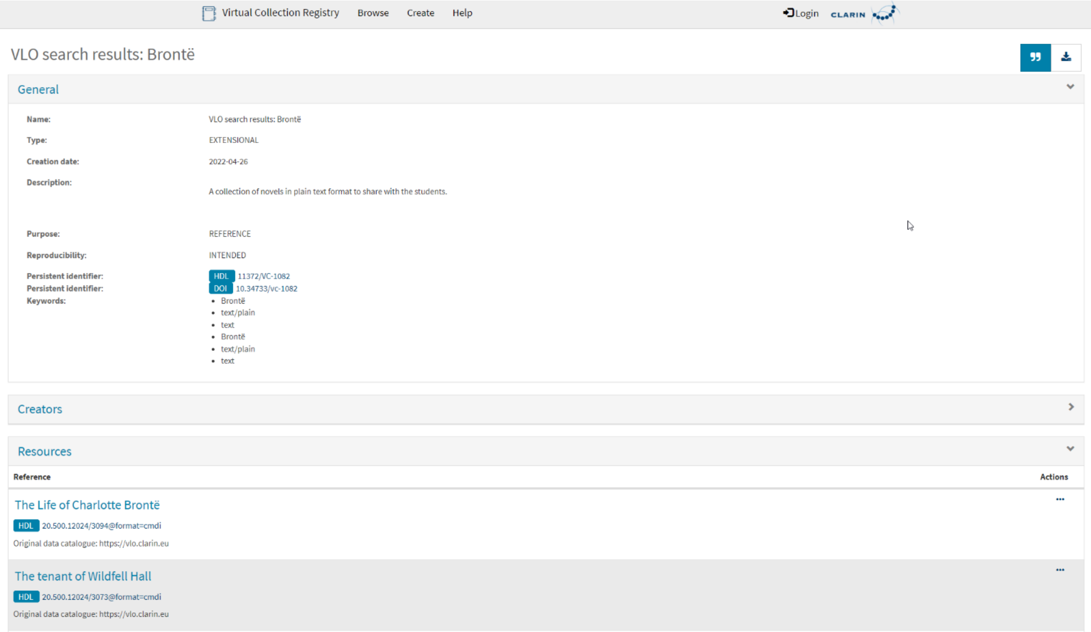
Structures and attributes

text	3,871
name	5,883

Collecting and Citing Language Resources

Virtual Collection Registry

- Collect the datasets discovered in the VLO or other data catalogues into a Virtual Collection
- Share the collection with the others
- Cite the collection
- Process the collection with Switchboard tools



VLO search results: Brontë

General

Name:	VLO search results: Brontë
Type:	EXTENSIONAL
Creation date:	2022-04-26
Description:	A collection of novels in plain text format to share with the students.
Purpose:	REFERENCE
Reproducibility:	INTENDED
Persistent identifier:	HDL 11372/VC-1082
Persistent identifier:	DOI 10.34733/vc-1082
Keywords:	• Brontë • text/plain • text • Brontë • text/plain • text

Creators

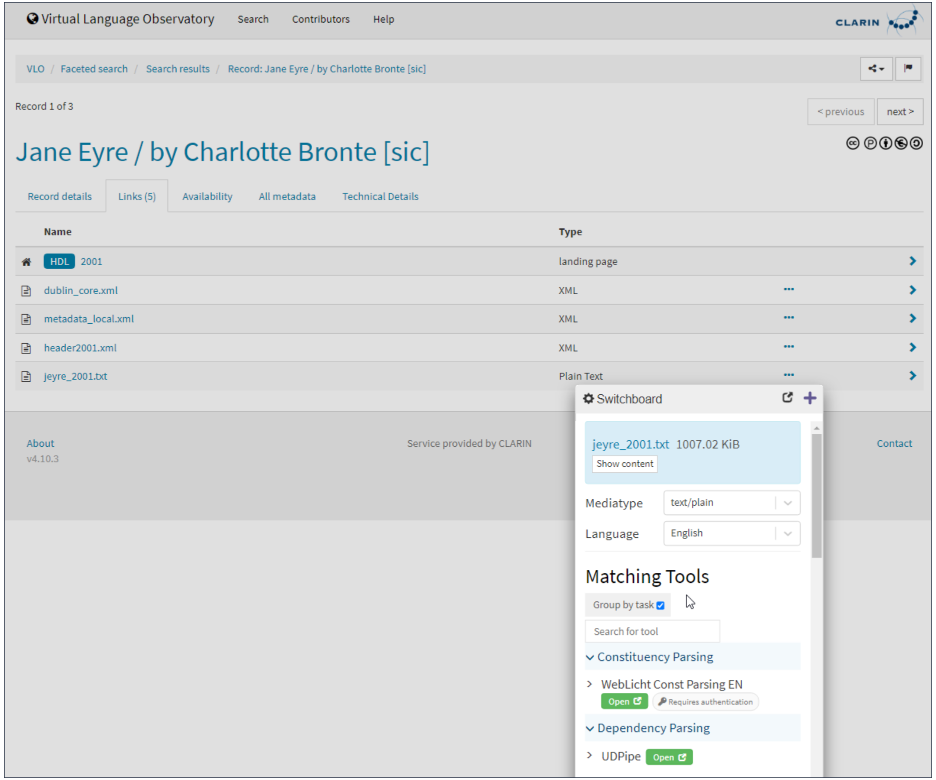
Resources

Reference	Actions
The Life of Charlotte Brontë HDL 20.500.12024/3094@format=cmdi Original data catalogue: https://vlo.clarin.eu	···
The tenant of Wildfell Hall HDL 20.500.12024/3073@format=cmdi Original data catalogue: https://vlo.clarin.eu	···

Processing Text Collections

The Language Resource Switchboard

- A service that allows you to find a matching tool to analyse plain text files
- Taggers, lemmatizers, named entity recognizers, chunking tools etc.



Record 1 of 3

Jane Eyre / by Charlotte Bronte [sic]

Name	Type
HDL 2001	landing page
dublin_core.xml	XML
metadata_local.xml	XML
header2001.xml	XML
jeyre_2001.txt	Plain Text

jeyre_2001.txt 1007.02 KiB

Mediatype: text/plain

Language: English

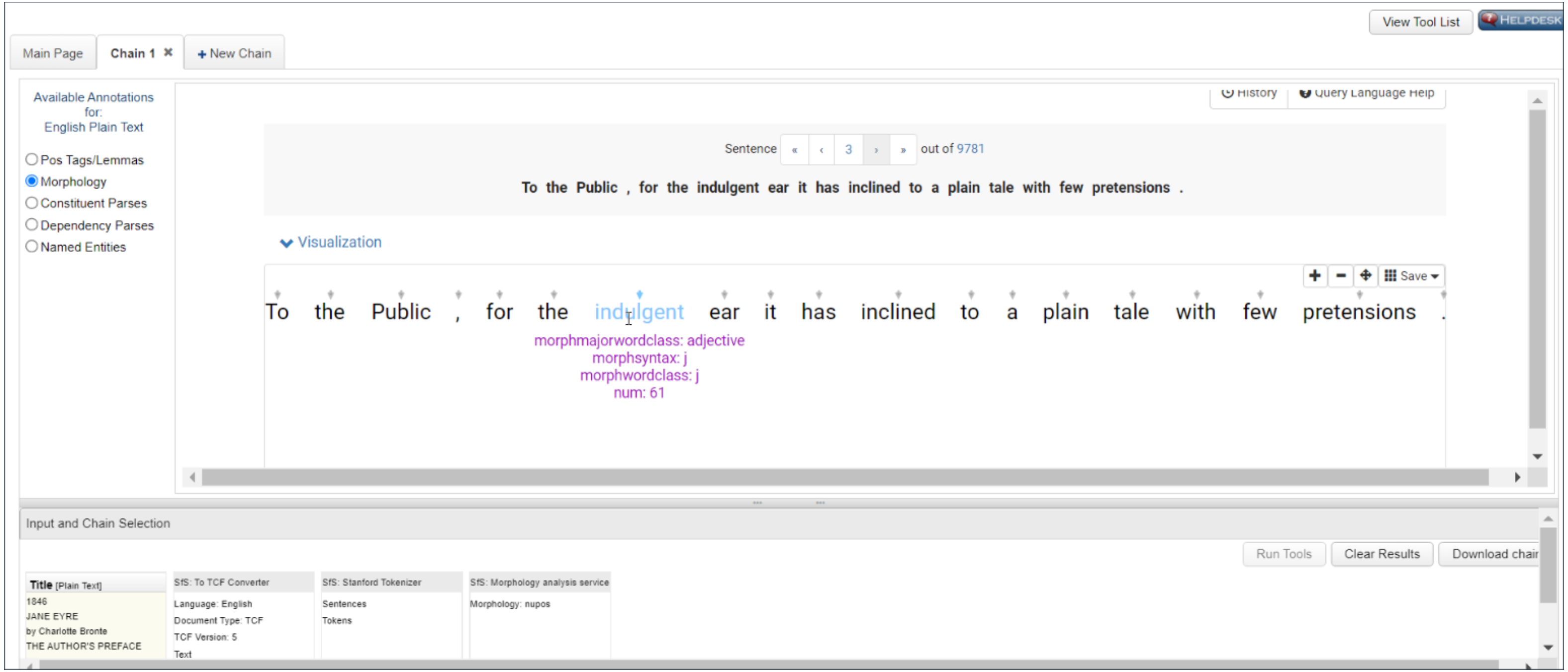
Matching Tools

Constituency Parsing
- WebLicht Const Parsing EN

Dependency Parsing
- UDPipe

Processing Text Collections

Example of morphological analysis in WebLicht



Available Annotations for: English Plain Text

○ Pos Tags/Lemmas
● Morphology
○ Constituent Parses
○ Dependency Parses
○ Named Entities

Sentence « ‹ 3 › » out of 9781

To the Public , for the indulgent ear it has inclined to a plain tale with few pretensions .

To the Public , for the indulgent ear it has inclined to a plain tale with few pretensions .

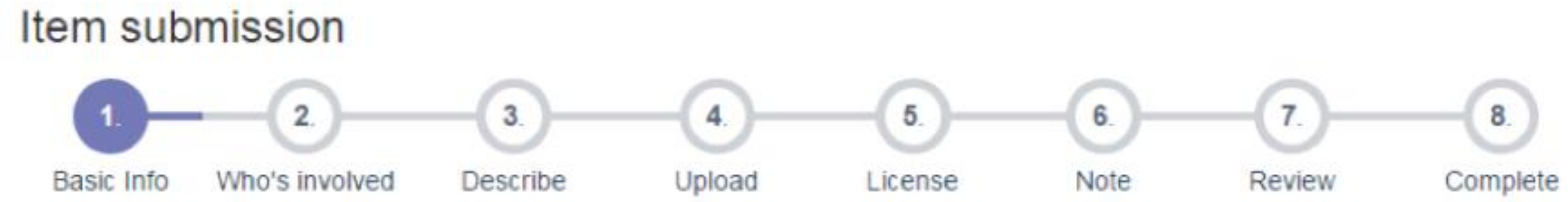
morphmajorwordclass: adjective
morphsyntax: j
morphwordclass: j
num: 61

Input and Chain Selection

Title [Plain Text]	SfS: To TCF Converter	SfS: Stanford Tokenizer	SfS: Morphology analysis service
1846 JANE EYRE by Charlotte Bronte THE AUTHOR'S PREFACE	Language: English Document Type: TCF TCF Version: 5 Text	Sentences Tokens	Morphology: nupos

Archiving Language Resources

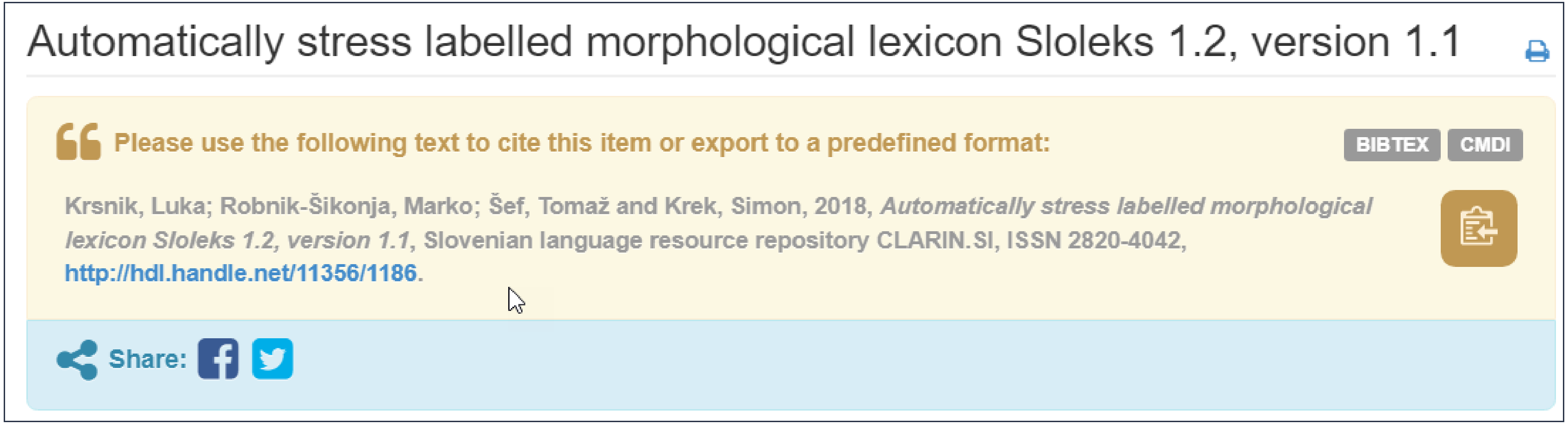
- Institutional repository
- Domain-specific repository
- Infrastructural repository: Depositing Services | CLARIN ERIC
- Support with data preparation, license selection, depositing process



Example of data submission lifecycle in CLARIN.SI

Sharing Language Resources

- Select appropriate licenses
- The repository assigns unique identifiers
- Others can cite the resources



Automatically stress labelled morphological lexicon Sloleks 1.2, version 1.1

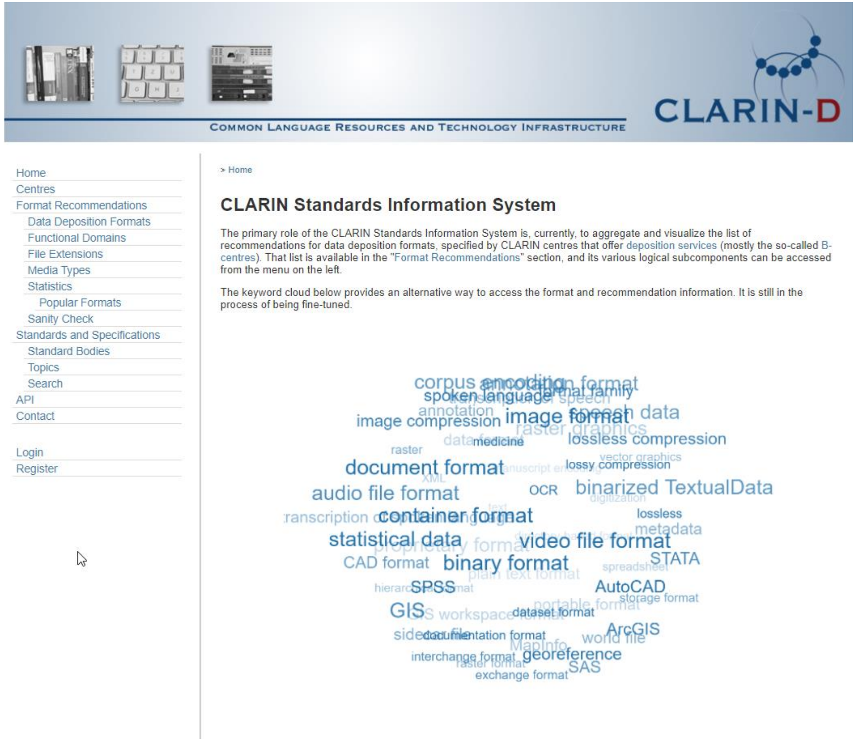
Please use the following text to cite this item or export to a predefined format:

Krsnik, Luka; Robnik-Šikonja, Marko; Šef, Tomaž and Krek, Simon, 2018, *Automatically stress labelled morphological lexicon Sloleks 1.2, version 1.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1186.

Guidance on the Use of Standards and Formats

CLARIN Standards Information System

- Data deposition formats
- Language-technology related standards



CLARIN Standards Information System

The primary role of the CLARIN Standards Information System is, currently, to aggregate and visualize the list of recommendations for data deposition formats, specified by CLARIN centres that offer deposition services (mostly the so-called B-centres). That list is available in the "Format Recommendations" section, and its various logical subcomponents can be accessed from the menu on the left.

The keyword cloud below provides an alternative way to access the format and recommendation information. It is still in the process of being fine-tuned.

Guidance on Licenses and Legal Issues in Data Reuse

On the CLARIN Legal Information Platform, you will find useful information about:

- Introduction to Copyright and Related Rights
- Licensing Practice
- Personal Data Protection

Questions?

Email: iulianna@clarin.eu