



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)

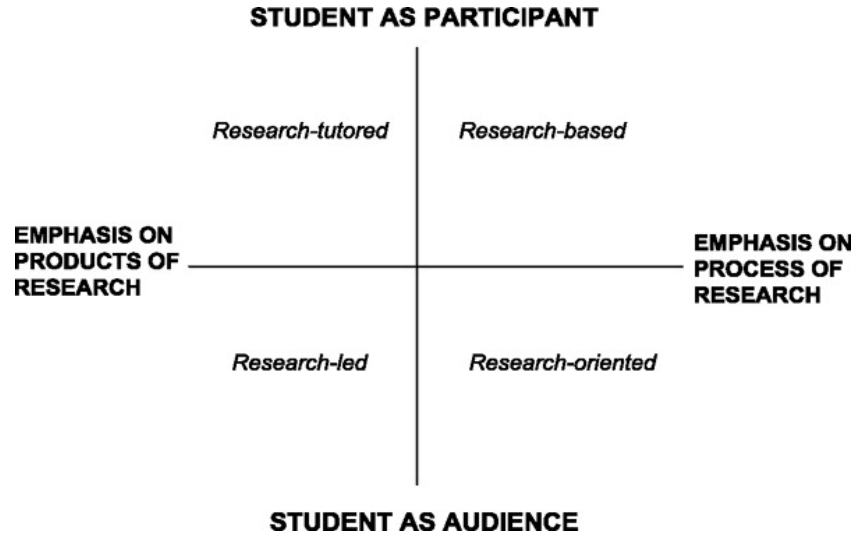
Graz Multiplier Event, 5 July 2022

Overview

Research-Based Teaching: Guidelines and Best Practices

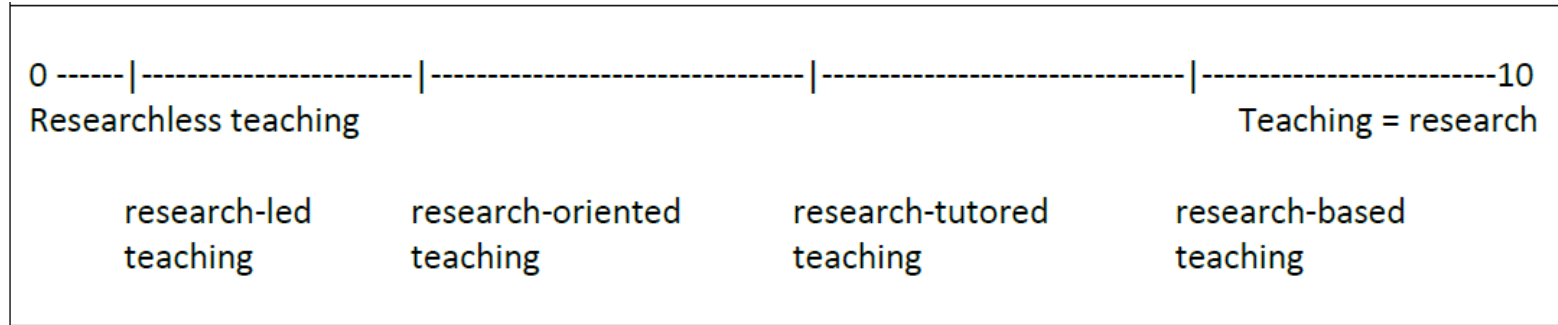
Integrating Research Infrastructures into Teaching: The Case of CLARIN

What is research-based teaching?



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?



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 solving**
- practice **forward thinking**
- 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

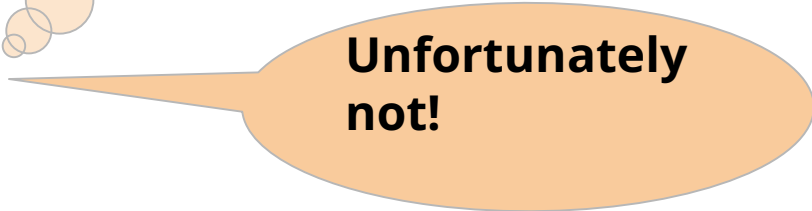
- direct **preparation for a**
- stimulate **curiosity**
- practice **problem solving**
- practice **forward thinking**
- gain **independence**
- gain **experience**



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



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



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



Research-based Teaching Guidelines

Table of contents	
Executive Summary	2
List of abbreviations	3
PART ONE: Research-based teaching and UPSKILLS	4
Introduction	4
RBT in UPSKILLS	9
Rationale and layout of an UPSKILLS RBT course	11
Choosing/developing a course subject	14
Learning outcomes of RBT courses	16
PART TWO: GUIDELINES	18
Making a course outline	18
Instructions for students	19
Organising and supervising the work	20
Evaluation and grading	22
After the course	24
PART THREE: EXAMPLES OF RBT COURSES	26
Course description template	26
Course Description 1: Current trends in Phonology	27
Course Description 2: Secondary imperfectivisation in Bosnian/Croatian/Serbian	34
Course Description 3: Multilingualism: Obtaining, representing and analysing empirical data	41
Course Description 4: Correlates of thematic roles in Bosnian/Croatian/Montenegrin/Serbian	48
Course Description 5: Deverbal derivations	55
References	62
Annex 1 - Detailed overview of general topics in RBT courses	63
Annex 2 - The structure of a research report	65
Annex 3 - Detailed overview of learning outcomes per topic block in RBT courses	67
Annex 4 - Assessment survey	69

Guidelines for integrating research into teaching

Featuring:

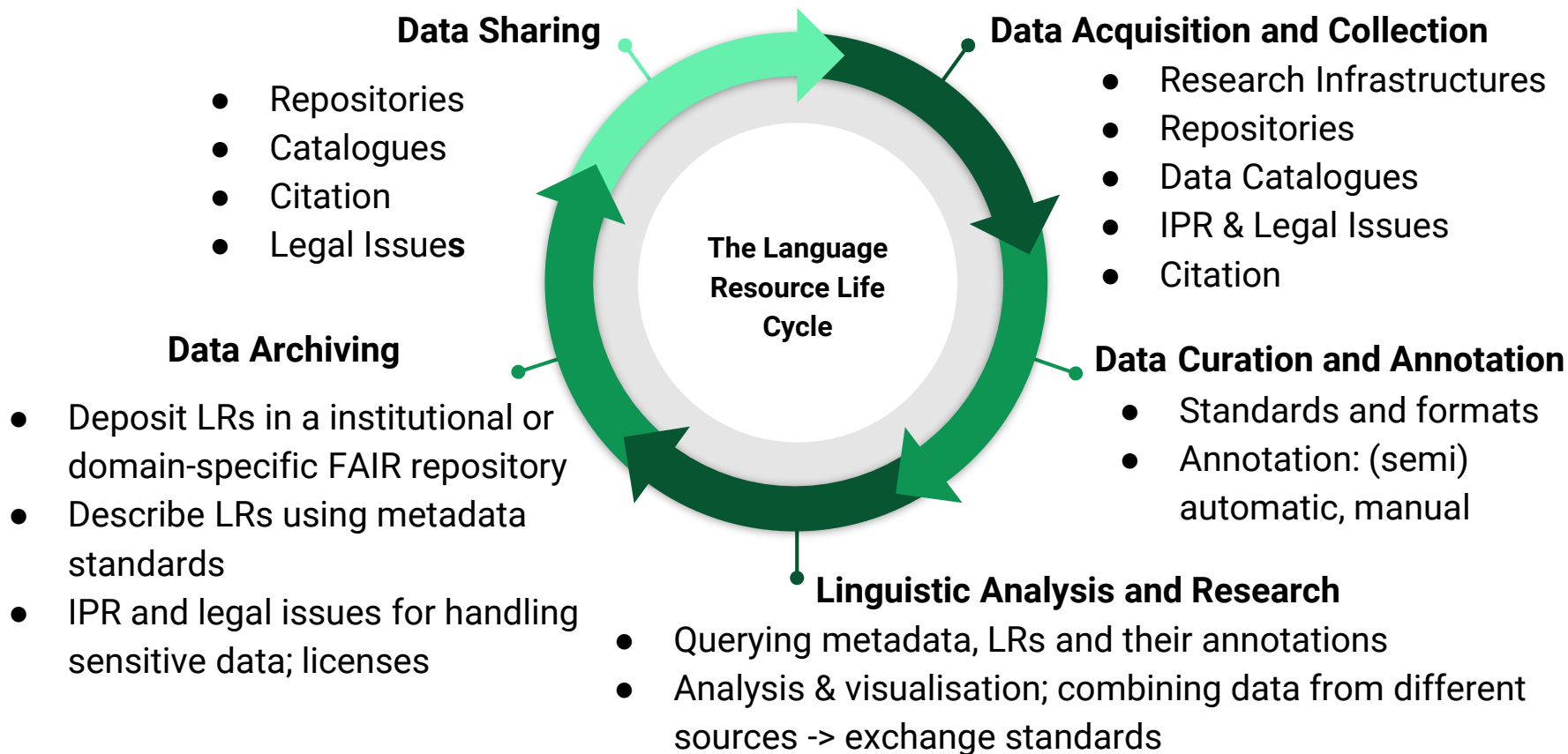
- Instructions for choosing/developing a course subject
- Detailed pick-and-choose list of research-related topics
- Detailed pick-and-choose list of research-related learning outcomes
- Making instructions for students
- Organising and supervising the work
- Evaluation and grading
- After the course
- 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



A Guide and a Course for Teachers

1. About	8
1.1. Context and Motivation	8
1.2. Aims of this Guide	9
1.3. How to Use this Guide	9
2. What are European Research Infrastructures?	10
3. What is CLARIN?	12
3.1. Technical Infrastructure	13
3.2. Knowledge Infrastructure	14
4. Integrating CLARIN in the University Curricula	15
4.1. Designing a University Course	18
4.1. Reusing the UPSKILLS Courses	20
4.1.1. Introduction to Language Data: Standards and Repositories	20
4.1.2. UPSKILLS Course: Automatic Speech Recognition and Forced Alignment	23
5. Other Training Resources	25
5.1. Teaching with CLARIN	25
5.2. Teaching with Parliamentary Datasets	26
5.3. Using Jupyter Notebooks and CLARIN NLP Tools for Full-Text Processing	28
5.4. The SSH Open Marketplace and the Training Discovery Toolkit	28
5.6. CLARIN Cafés	29
5.7. The Digital Humanities Course Registry	30
5.8. CLARIN Video Channels	31
5.9. The CLARIN Annual Conference	32
6. Examples of Courses and Teaching Practices	33
6.1. Building and Maintaining Online Courses in Digital Research Methods at the University of Helsinki/FIN-CLARIN	33

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

The screenshot shows the CLARIN Virtual Language Observatory (VLO) search results page. The search criteria are 'tmx corpora' and 'Italian'. The results are filtered to show 4 items. The first result is 'Parallel Global Voices', which is a set of parallel and monolingual corpora generated from the Global Voices multilingual group of websites. The second result is 'ECDC Translation Memory', which is a collection of sentences and their professionally produced translations in twenty-five languages. The third result is 'DGT-TM-2016', which is a multilingual Translation Memory for the Acquis Communautaire. The fourth result is 'DGT-Translation Memory', which is a multilingual Translation Memory for the Acquis Communautaire. The page also includes a sidebar with filters for Language, Collection, Resource type, Keyword, Temporal Coverage, Availability, and Search options.

Virtual Language Observatory Search Contributors Help

VLO / Faceted search / Search results

tmx corpora

Showing 4 results within selection for **tmx corpora** **Italian** Results per page: 10

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

Language

Italian

Italian **X**

more...

Collection

Resource type

Type to filter or search for more

corpus (4)

Keyword

Temporal Coverage

Availability

Search options

Parallel Global Voices.
(Part of CLARIN 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...

Amharic Arabic Aymara Bengali; Ban. Bulgarian ... (+35)

Landing page for this record

ECDC Translation Memory
(Part of CLARIN 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 ...

Bulgarian Czech Danish Spanish; Cas. Estonian ... (+20)

Landing page for this record

DGT-TM-2016
(Part of CLARIN 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...

Bulgarian Czech Danish Spanish; Cas. Estonian ... (+19)

Landing page for this record

DGT-Translation Memory
(Part of CLARIN 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

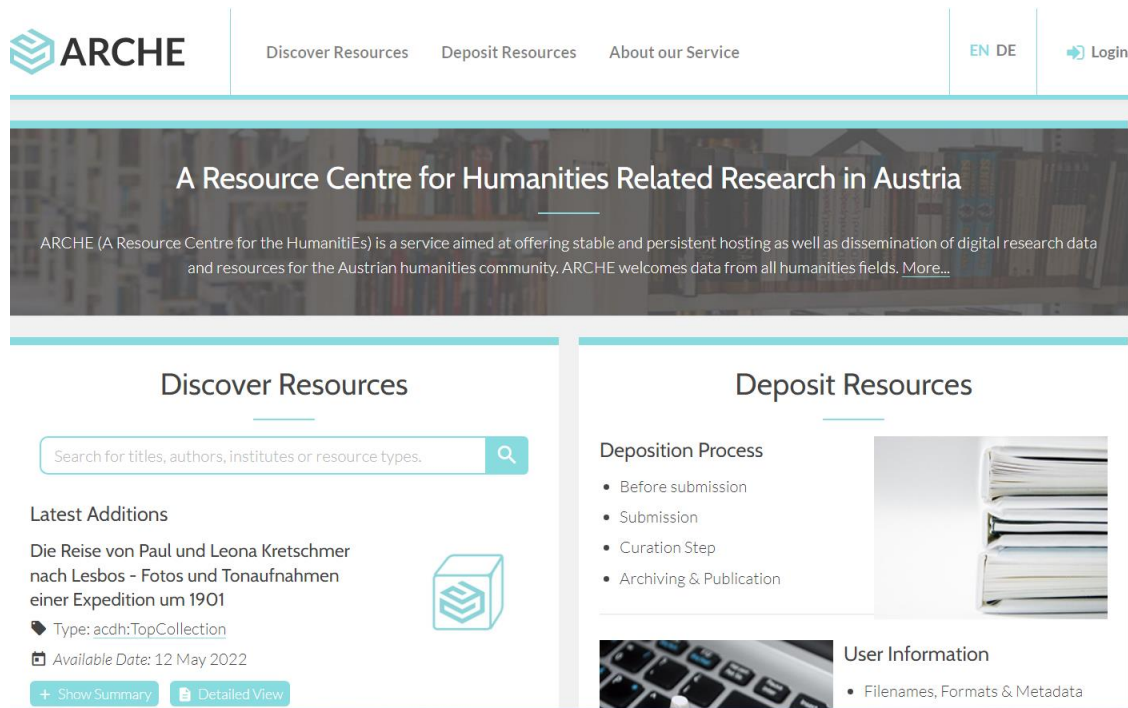
- To search for specific patterns across corpora at the same time
- The search results from a specific corpus can be downloaded in various formats and processed in other tools for further analysis
- Possible to retrieve the citation information
- Useful as a first step to discover where interesting lg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

The screenshot displays the 'Content Search' interface. At the top, there are tabs for 'Content Search', 'Aggregator', and 'Help'. Below the tabs is a search bar with the text 'elephant' and a magnifying glass icon. Underneath the search bar, there are filters: 'Search for' with a dropdown menu set to 'English', 'Text layer Contextual Query Language (CQL)' with a dropdown menu, 'in' followed by a dropdown menu set to '5 selected collections', and 'and show up to' followed by a dropdown menu set to '10' and 'hits per endpoint'. Below the filters, it says '32 matching collections found'. To the right of this text are two buttons: 'Display as Key Word In Context' (checked) and 'Download'. Below this, there are three search results, each with a title, a description, and a 'View' button. The first result is 'Tagesspiegel – Berlin-Brandenburg Academy of Sciences and Humanities'. The second result is 'Corpus LVK2018 (12M tokens) – CLARIN Centre of Latvian language resources and tools'. The third result is 'Berliner Zeitung – Berlin-Brandenburg Academy of Sciences and Humanities'. The 'Berliner Zeitung' result is expanded, showing a list of text snippets with the word 'elephant' highlighted in blue. The snippets are: 'Niemand kann sich erinnern, warum der Pub am historischen Marktplatz von Dover ausgerechnet * elephant und 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 Kameval 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 Elephant and Castle * trägt.', 'Fhemann_sitzt wie so oft während dieser öffentlichen Anhörung zu'.

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](#)



The screenshot shows the ARCHE website. The header includes the ARCHE logo and navigation links: "Discover Resources", "Deposit Resources", "About our Service", "EN DE", and a "Login" button. The main banner reads "A Resource Centre for Humanities Related Research in Austria" and provides a brief description of the service. Below the banner, there are two main sections: "Discover Resources" and "Deposit Resources". The "Discover Resources" section features a search bar with the placeholder text "Search for titles, authors, institutes or resource types." and a magnifying glass icon. Below the search bar, it lists "Latest Additions" with a specific example: "Die Reise von Paul und Leona Kretschmer nach Lesbos - Fotos und Tonaufnahmen einer Expedition um 1901". It also includes the type "acdih:TopCollection" and the available date "12 May 2022". There are two buttons: "+ Show Summary" and "Detailed View". The "Deposit Resources" section includes a "Deposition Process" list: "Before submission", "Submission", "Curation Step", and "Archiving & Publication". It also features a "User Information" section with the text "FileNames, Formats & Metadata".

<https://arche.acdh.oew.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	Lexical Resources	Tools
<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• Lexica• Dictionaries• Conceptual Resources• Glossaries• Wordlists	<ul style="list-style-type: none">• 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 \(2016\)](#)

KonText

noSketch

Download

User guide

[Query the ReLDI corpus via NoSketch engine on CLARIN.SI](#)

[ReLDI-hr \(manually tagged Croatian tweets\)](#)

ReLDI-hr (manually tagged Croatian tweets)

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

Counts	General info	Lexicon sizes	Tags legend
Tokens 89,104	Corpus description Document	word 27,289	Noun N.*
Words 71,768	Language Croatian	norm 25,395	Noun proper Np.*
Sentences 7,939	Encoding UTF-8	lempos 17,270	Noun common Nc.*
Documents 3,871	Compiled 09/11/2019 15:41:23	tag 694	Verb V.*
	Tagset Description	ud_pos 90	Adjective A.*
		ud_feats 764	Pronoun P.*
		diff 5	Adverb R.*
		lc 25,219	Preposition S.*
		lemma 16,675	Conjunction C.*
		lemma_lc 16,020	

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

The screenshot shows the Virtual Collection Registry (VLO) interface. At the top, there's a navigation bar with 'Virtual Collection Registry', 'Browse', 'Create', and 'Help'. On the right, there are links for 'Login' and 'CLARIN'. The main content area displays 'VLO search results: Brontë'. Below this, there's a 'General' tab with the following details:

- 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: 10.34733/vc-1082
- Persistent Identifier: 10.34733/vc-1082
- Keywords: Brontë, text/plain, text, Brontë, text/plain, text

Below the 'General' tab, there are sections for 'Creators' and 'Resources'. The 'Resources' section is expanded, showing two entries:

- The Life of Charlotte Brontë**
DOI: 10.500.12024/3094@format=cmidi
Original data catalogue: <https://vlo.clarin.eu>
- The tenant of Wildfell Hall**
DOI: 10.500.12024/3073@format=cmidi
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.

The screenshot displays the Virtual Language Observatory (VLO) interface. The main header includes 'Virtual Language Observatory', 'Search', 'Contributors', and 'Help'. The breadcrumb trail shows 'VLO / Faceted search / Search results / Record: Jane Eyre / by Charlotte Bronte [sic]'. The record is identified as 'Record 1 of 3' and titled 'Jane Eyre / by Charlotte Bronte [sic]'. Below the title are tabs for 'Record details', 'Links (5)', 'Availability', 'All metadata', and 'Technical Details'. A table lists the following items:

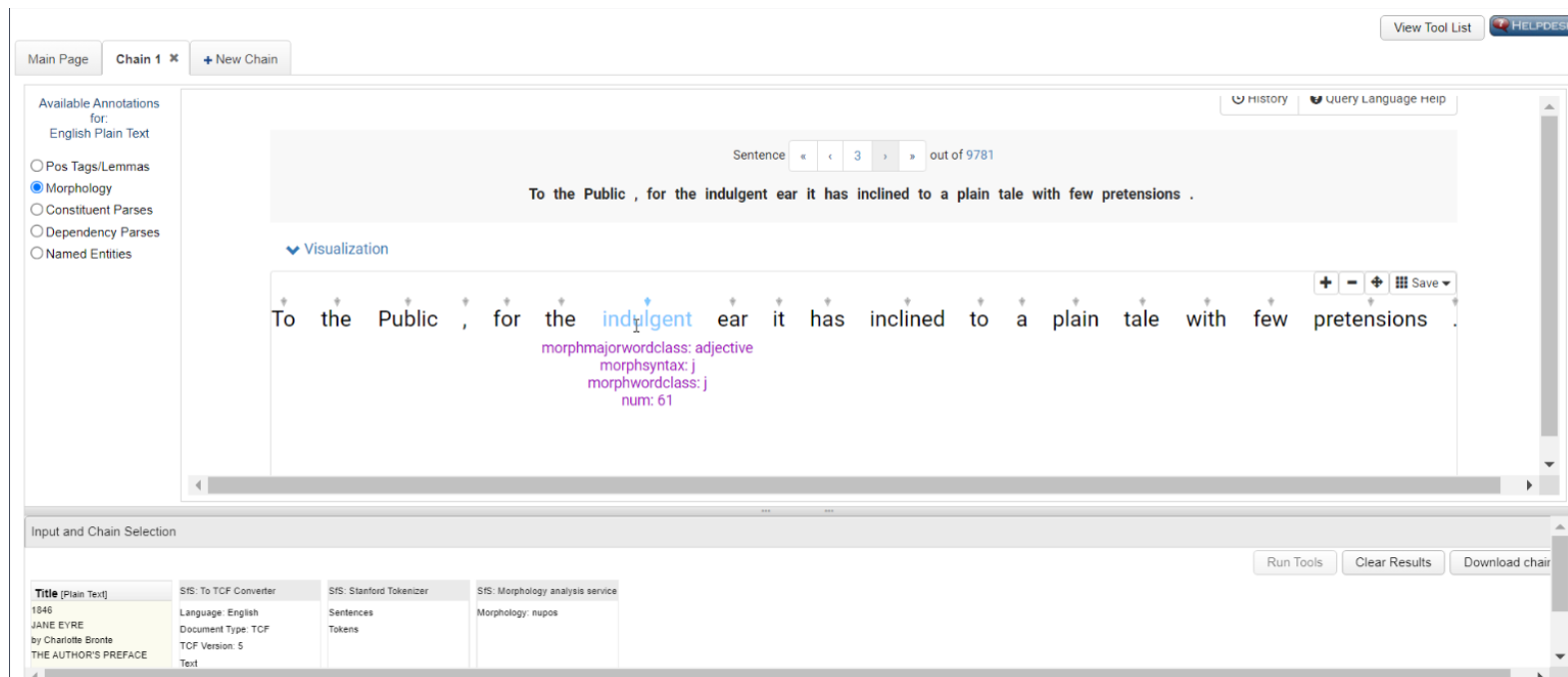
Name	Type
HDL_2001	landing page
dublin_core.xml	XML
metadata_local.xml	XML
header2001.xml	XML
jeyre_2001.txt	Plain Text

A modal window titled 'Switchboard' is open, showing details for 'jeyre_2001.txt' (1007.02 KiB). It includes a 'Show content' button, dropdown menus for 'Mediatype' (text/plain) and 'Language' (English), and a section for 'Matching Tools'. The 'Matching Tools' section includes a 'Group by task' checkbox and a 'Search for tool' input field. The following tools are listed:

- Constituency Parsing
 - WebLight Const Parsing EN (Open button, Requires authentication)
- Dependency Parsing
 - UDPipe (Open button)

Processing Text Collections

Example of morphological analysis in [WebLicht](#)



The screenshot displays the WebLicht web application interface. At the top, there are tabs for 'Main Page', 'Chain 1', and '+ New Chain'. On the right, there are links for 'View Tool List' and 'HELPDESK'. The left sidebar, titled 'Available Annotations for: English Plain Text', contains radio buttons for 'Pos Tags/Lemmas', 'Morphology' (which is selected), 'Constituent Parses', 'Dependency Parses', and 'Named Entities'. The main content area shows a sentence: 'To the Public, for the indulgent ear it has inclined to a plain tale with few pretensions .'. Above the sentence, it says 'Sentence 3 out of 9781'. Below the sentence, there is a 'Visualization' section. The word 'indulgent' is highlighted in blue, and a tooltip shows its morphological analysis: 'morphmajorwordclass: adjective', 'morphsyntax: j', 'morphwordclass: j', and 'num: 61'. At the bottom, there is an 'Input and Chain Selection' section with a table showing the input text and the tools used for processing.

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

Buttons at the bottom right: Run Tools, Clear Results, Download chain

Archiving Language Resources

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

Item submission



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:

BIBTEX

CMDI

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>.



Share:



Guidance on the Use of Standards and Formats

CLARIN Standards Information System

- Data deposition formats
- Language-technology related standards

Home
Centres
Format Recommendations
Data Deposition Formats
Functional Domains
File Extensions
Media Types
Statistics
Popular Formats
Sanity Check
Standards and Specifications
Standard Bodies
Topics
Search
API
Contact
Login
Register

> Home

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.

corpus encoding format spoken language data image compression image format data lossless compression document format audio file format container format statistical data video file format binary format CAD format SPSS GIS workspace dataset format portable format storage format AutoCAD ArcGIS world file interchange format georeference SAS

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