CLARIN in the Classroom

Moderator: Francesca Frontini

Wednesday 7 October, 13.45 - 14.45
Introduction

For CLARIN:
• building a strong community of teachers and lecturers
• progressing towards the creation of a CLARIN Training Suite

For you:
• more usable tools and resources, easier to integrate in your syllabus
• possibilities for collaboration and support
Theme 1 - Corpora and Linguistics
Mietta Lennes, University of Helsinki / FIN-CLARIN

Building and maintaining online courses in digital research methods

Corpus Linguistics and Statistical Methods (5 cr)

Introduction to Speech Analysis (5 cr)

Data Clinic (5 cr)

https://www.kielipankki.fi/support/training
Example course: Data Clinic (5 ECTS)

- Support in language data management for students during their MA or PhD projects
- Each student composes a Data Management Plan (DMP)
Which components could be shared as Learning Objects?

Stages of the Data Clinic

1. Contextualize your data; submit initial Data Management Plan
2. Organize and pre-process your material
3. Annotate and search
   - text
   - speech (audio and video)
4. Wrangle and visualize data
5. Discover suitable tools
   – meet a Data Advisor
6. Submit your final DMP
Laura Herzberg, University of Mannheim

Corpus literacy in German linguistics: the usage of corpus tools and platforms in academic classrooms

- German Linguistics department at the University of Mannheim offers corpus linguistic classes (5-6 ECTS) as an integral part of the bachelor's and master's curricula in a variety of study programs:
  - Bachelor's program in German Studies Language, Literature, Media
  - Bachelor's program in Media and Communication Studies
  - Bachelor's program in Culture and Economy
  - Interdisciplinary study programs: Media and communication studies combined in one program

- Corpora are also gaining in importance as resources in language teaching (B.A./M.A. of Education) → in a teacher training class, students learn how to use corpora with regard to their own future teaching career

- Corpus platforms and tools - CLARIN centers: Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) and Leibniz Institute for German Language (IDS) Mannheim
● introduction to the corpus platforms provided by these centres with the help of tutorials and complementary tasks; introduction to empirical linguistics; limits & chances of corpus research

● students develop linguistic research questions, e.g. in the field of computer-mediated communication
  ○ topics: comparison of German registers/vocabularies, word pairs, investigation of spoken elements in written texts, creation of lexicon entries
  steps:
    ❑ corpus data: blogs, discussion forums, website/internet corpora
    ❑ literature study
    ❑ query and sample the data
    ❑ analyse data
    ❑ discuss results
“My enthusiasm for corpus linguistics was [...] awakened - I would like to attend more of these seminars and would like to continue working in corpus linguistics.”

“Thanks to the course I am now able to perform [corpus] analyses -> it has a practical relevance for me and I see the how independent research works”

“...not only receive theoretical knowledge, but also work empirically with corpora”

- individual research projects
- corpus tools & platforms as knowledge gain & “practical” approach

“I would have found it better if we had worked with the platforms more often or more intensively.”

network / platform for virtual discussion / planning of projects

corpus project meetups – (interdisciplinary) tandem courses – students create task force wrt their field of study to investigate a topic during the course of one semester
A comparable corpus of original and translated Lithuanian ORVELIT (Originalios ir vertimų lietuvių kalbos tekstynas)

Composition: 4 sub-corpora of original and translated fiction and popular science literature

Raw (3 998 484 tokens) and morphologically annotated versions.

Access: CLARIN-LT repository

Applications:
- Corpus-Based Translation Studies (e.g., English language interference; features of translations)
- Register variation
- Comparison of original and translated language (relevance: translators, editors, teaching/learning and research material)
Contrastive Stylistics

- MA level course in the programme of Applied English Linguistics at Vytautas Magnus University.
- Curriculum extract:

<table>
<thead>
<tr>
<th>Study programme outcomes</th>
<th>Course outcomes</th>
<th>Content</th>
</tr>
</thead>
</table>
| Apply current concepts from translation theory and practice in analysing and translating English and Lithuanian texts of different functional styles and genres. | Identify register and genre features in original and translated English and native language texts. | ● Detailed description of register, genre, and style: fiction.  
● Project: a comparison of original and translated texts of a chosen genre; |
| Develop the skills needed to work with specialized texts (business, legal, academic, etc.), to choose appropriate translation strategies for dealing with such texts. | Demonstrate the knowledge of finding and using electronic language resources and tools of their analysis. | ● CLARIN ERIC and Virtual Language Observatory.  
● Language resources in CLARIN-LT repository. |
## CLARIN in class

<table>
<thead>
<tr>
<th>ORVELIT</th>
<th>CLARIN</th>
<th>TASKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Parallel and comparable corpora for translation and Lithuanian language research</td>
<td>● Virtual language observatory &lt;br&gt; ● CLARIN resource families &lt;br&gt; ● CLARIN-LT repository &lt;br&gt; ● CLARIN-UK: LTWAC; LT-FORUM; WIKI-LT</td>
<td>● Search/ access relevant resources  &lt;br&gt; ● Provide feedback</td>
</tr>
<tr>
<td>● Sharing experience in corpus creation</td>
<td>● CLARIN-D: Data Management Plan</td>
<td>● Prepare a project proposal for a DIY corpus</td>
</tr>
<tr>
<td>● Researching translated and original language: unannotated data</td>
<td>● CLARIN-LT: ORVELIT &lt;br&gt; ● CLARIN-UK: #LancsBox</td>
<td>● Download the corpus  &lt;br&gt; ● Learn to use the basic functions of corpus analysis tools &lt;br&gt; ● Provide feedback</td>
</tr>
<tr>
<td>● On the way: working with morphologically annotated data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In sum

- Step-by-step guidance on the basis of personal corpus creation and research experience helps students to:
  - search for open access language resources and their analysis tools on their own;
  - plan individual research projects;
  - gain knowledge of corpus analysis tools;
  - raise questions and conduct small-scale research;
  - critically report their findings in relation to previous research.
Katarzyna Klessa, Adam Mickiewicz University in Poznan

About the use of CLARIN tools in the courses taught to students of empirical linguistics and language documentation

ELLDo

elldo.amu.edu.pl

(MA in English)

Lingwistyka Komputerowa

Studia licencjackie i stacjonarne

Common Language Resources and Technology Infrastructure

Computer Linguistics

computerlinguistics.amu.edu.pl

(katarzyna.klessa.pl

(BA - mostly - in Polish)
Corpus linguistics

Course for 1st/2nd grade of MA studies

6 ECTS

Initial part of the course covers fundamentals of CL. In the “CLARIN-related” part students:

- explore CLARIN corpus resources & tools and solve simple tasks
- discuss how the tools & resources help solve research problems, e.g. formulated in MA diploma projects
Documentary linguistics

Course for 1st grade of MA studies

5 ECTS

Students:

- learn about data & metadata standards
- search online repositories, archives
- practice annotation of text & speech corpora
Experimental phonetics

Course for 2nd grade of BA studies

2 ECTS

Students:

● get familiar with fundamentals of grapheme-to-phoneme conversion (GTP)
● learn to understand the speech signal display
● learn about automatic segmentation / alignment
● collaboration with CLARIN -> tool development
Who are we?

The Applied Linguistics Laboratory (LABLING) at the Konstantin Preslavski University of Shumen is technological partner with the ClaDa-BG National consortium. The LABLING team`s research focuses on creating computer corpora of children`s speech and collections of associative data.

Results from two years of work on the ClaDa-BG project

• Bulgarian LabLing Corpus – corpus of Bulgarian children`s speech, published on CHILDES (https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html)
• LABLASS - web-based system for systematizing and organizing word associations into dictionaries
**Bulgarian LabLing Corpus**

The LabLing corpus includes two segments: the longitudinal corpus and the narrative corpus. *The longitudinal corpus* contains the transcribed data of 4 Bulgarian girls. *The narrative corpus* contains 91 transcripts of preschool children`s narratives extracted from 50 monolingual native speakers of Bulgarian.

**LABLASS**

Currently, the pilot version of the system has been built and is being constantly updated. The abilities of the web-based LABLASS system developed within the ClaDa-BG project are not limited to including available lexicographic resources but are instead much broader, which results in creating new dictionaries, the visualization and comparison of data from different sources. 12 students who have been trained to work with the system enter the associative data for the creation of several dictionaries:

- Dictionary of Bulgarian word associations from the early 20s of the 21st century
- Dictionary of word associations of monolingual (Bulgarian) and bilingual (Bulgarian and Turkish) persons
- New dictionary of child word associations
Practical data application

• in the curricula of linguistic disciplines - the curricula of the disciplines Psycholinguistics, General Linguistics and Child Linguistics have been extended;

• in lectures as material for demonstrations;

• in Psycholinguistics practical modules as methodological models for creating original corpora by the students themselves;

• in students` course assignments and theses


• in teaching resources

Popova, Velka. 2020. Psycholinguistics as Experimental Linguistics. Shumen

Theme 2 - Resources for the Humanities
CLARIN in the Classroom: Academic skills

Wiktoria Mieleszczenko-Kowszewicz,
Faculty of Psychology, SWPS University of Social Sciences and Humanities,
Warsaw, Poland
Introduction

Goals

- Students get acquainted with the concept of competent judges, examples of research where qualitative data are used.
- Show students the process of searching words from their own category in text corpora.
- Show students the coding process of qualitative data in practice.
## Resources & tools

<table>
<thead>
<tr>
<th>Stage</th>
<th>Clarin tool</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation before analysis</td>
<td>N/A</td>
<td>Create/download own corpora of text</td>
</tr>
<tr>
<td>First stage</td>
<td>LEM (option: sentiment analysis)</td>
<td>Frequency analysis of words</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>Decide if word belongs to the category</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>Word selection</td>
</tr>
<tr>
<td>Second stage</td>
<td>LEXP</td>
<td>Choose proper meaning of chosen words</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>Preparation of Excel file with chosen meanings</td>
</tr>
<tr>
<td></td>
<td>LEM (option: own category)</td>
<td>Analysis text corpora</td>
</tr>
</tbody>
</table>
Lessons scenarios

Lesson 1

- Create/download own text corpus
- Create a coding manual

Lesson 2

- First stage of words’ evaluation

Lesson 3:

- Second stage of evaluation (meanings of words chosen in the first stage)
The making of the siParl tutorial

Kristina Pahor de Maiti and Darja Fišer
Faculty of Arts, University of Ljubljana, Slovenia; Jožef Stefan Institute, Ljubljana, Slovenia
Introduction

Starting point

- Many parliamentary corpora in CLARIN
- Available via online concordancers
- Relevant for a broad range of disciplines
- Relevant for trans-national research

Goal

- Showcase the potential of this special data type
- Transfer CorpLing techniques to other SSH disciplines
Obstacles and Solutions

Resource & tools

- Scenario 1: Use several parliamentary corpora for comparative analysis
  - Very uneven time span, metadata & concorancers
- Scenario 2: Use the Hansard corpus for internationally understandable examples
  - Concordancer doesn’t offer functionalities needed for the selected research problem

Use case

- Target audience: SSH fields
  - Embedding into their theoretical and methodological framework
- Avoid culture/language-specific phenomena
- Avoid research questions which require detailed linguistic knowledge
Obstacles and Solutions

Delivery format

• Self-standing tutorial (theory supporting the practical tasks)
• Exploit the online environment (hyperlinks to external resources & to the concordancer, screencasts)

Testing and feedback

• Expert review + students testing sessions
• Your feedback welcome as well: https://bit.ly/30thMBN
Conclusion

Lessons learned

• Creating tutorials requires a broad spectrum of skills and is very time consuming
• New versions of tools/resources require regular updates of the tutorial
• Lack of standardisation in corpus creation and search environments hinder reuse and comparative research

Recommendations

• Promote common encoding standards for corpora and make them available through a single concordancer
• Document how corpora were created, structured and annotated because this knowledge is essential for proper use of the corpus
• Foster collaboration between lecturers to jointly contribute to the CLARIN training suite
CLARIN-IT from Pisa to Venice, Macerata and Siracuse: classes and seminars for students and scholars

Federico Boschetti and Monica Monachini, CNR-ILC Pisa and CLARIN-IT
Digital Philology and Computational Linguistics
Overview

Universities

- University of Pisa (MA, Summer School)
- University of Macerata (PhD)
- University Ca’ Foscari of Venice (MA)
- Venice International University (MA)

High Schools

- Liceo Classico “T. Gargallo”, Siracusa
- Liceo Classico “G. Galilei”, Pisa

F. Boschetti and M. Monachini, CNR-ILC & CLARIN-IT
Strengths

Universities

Students become aware of
- Open Science
- FAIR data
- the role of Research Infrastructures for (Digital) Humanities and (Computational) Linguistics

High Schools

Very young students (16-18 y.o.) desire
- to feel part of a community (making resources useful for the others)
- to explore available resources relevant for their future studies

F. Boschetti and M. Monachini, CNR-ILC & CLARIN-IT
Weaknesses and plans to overcome them

Weakness

- Students are only consumers of Linguistic Resources
- Few universities, few students
- Resources targeted for the research, not for the education

Plans

- Engage them more and more in the productive process of the Linguistic Resources
- Actions to promote CLARIN in the classrooms (awareness campaigns)
- Create (and/or adapt) resources for the students

F. Boschetti and M. Monachini, CNR-ILC & CLARIN-IT
Theme 3 - Integrating Computation
CLARIN in the Classroom -
Teaching *Computational Linguistics* to Master students within a *Digital Humanities degree* program at Pisa University

Simonetta Montemagni – Giulia Venturi
ILC-CNR - Pisa University – CLARIN-IT
Course goal

- twofold goal, covering both the practical utility of NLP in real world applications with a specific view to the area of SSH and its promise for improving the understanding of human language and/or for exploring humanistic texts

- in both cases, the adaptation of existing tools and resources to the specific language variety which needs to be automatically processed is the typical situation to be tackled
  - e.g. historical varieties of language, social media or domain-specific language, or different textual genres / registers

- the domain adaptation topic investigated through a project aimed at developing a linguistically annotated test corpus belonging to a non-standard variety of language use to assess the performance of existing NLP tools
  - for the project, tools and resources distributed via CLARIN are used
Used CLARIN resources and tools

Treebank

UDPipe linguistic annotation pipeline
Domain adaptation project using CLARIN tools and resources

Multi-topic Twitter corpus

Multi-topic annotated corpus
Domain adaptation project using CLARIN tools and resources

- Multi-topic annotated corpus
  - Manual revision of the automatic linguistic annotation by two students working separately
  - Interannotator agreement analysis
  - Merging and harmonization of the annotated corpus
  - Parsing evaluation against different UDPipe models trained on different varieties of language use (news vs social media language)
How CLARIN could support this type of project?

Creation of an integrated domain adaptation platform for different languages also including annotation editing, interannotator agreement, parsing evaluation and visualization modules

- Manual revision of the automatic linguistic annotation by two students working separately
- Interannotator agreement analysis
- Merging and harmonization of the annotated corpus

TARGET annotated GOLD corpus

Parsing evaluation against different UDPipe models trained on different varieties of language use

TARGET annotated corpus

Great teaching potential of the project: students are confronted with the real problems connected with the automatic analysis of specific varieties of language use and are requested to find the most appropriate (i.e. both UD compliant and linguistically grounded) annotation
CLARIN in the Classroom: Case of Latvia

Inguna Skadiņa, Ilze Auziņa and Baiba Saulīte
Institute of Mathematics and Computer Science, University of Latvia

Activities of CLARIN Latvia are supported by the European Structural Funds project No. 1.1.1.5/18/I/016.
Background and Current Activities

- Teaching of the Computational Linguistics course started in 2003 at Liepāja University:
  - course for the master students in linguistics
  - course for the master students in computer science
  - later Computational Linguistics was included in a course for doctoral students
    *Novel approaches to Linguistics* at Liepāja University
- Latvia joined CLARIN ERIC in 2016
- CLARIN-LV repository has been registered in March, 2020
- In 2017 new Computational Linguistics course for Master Students in English philology was started
- In Autumn, 2020 Computational Linguistics course started for Master Students in Baltic Philology
In lecture we introduce students with CLARIN research infrastructure, Virtual Language Observatory and resource families.

- We also highlight some LRTs that are more relevant to the research interests of students for particular year of studies.
- In the seminar students present their findings – tool or resource they found in CLARIN VLO and that seemed interesting for her/him.
Needs and problems

- We are familiar with language resources and tools for Latvian
- Tools and resources that might be of interest for students of English philology are not so well known for us
- We would be very interested in collective work of CLARIN partners to create an aligned list of language resources and tools for all languages (e.g., morphologically annotated corpora, treebanks, POS taggers, etc.) that can be used in teaching

https://www.clarin-d.net/en/word-level-based-comparative-textanalysis
Integrating Computation into the Humanities: Using Clarin Data in the Digital Humanities Hackathon in Helsinki

Mikko Tolonen, University of Helsinki
DIGITAL HUMANITIES AT THE UNIVERSITY OF HELSINKI

• Use of data science within the realm of SSH research (humanities and social science)
• Collaborative effort combining the expertise of SSH researchers and data scientists
• Also, to examine digitalisation as a cultural and social phenomenon
Core Concept of the DHH Hackathon

- In miniature size, reproduce an actual multidisciplinary digital humanities research project
- Learn the quirks of such a process, operating between traditions on complex data
- Do away with boundaries between students, teachers, researchers, between research and teaching

“toughest but most rewarding week of yours studies”
Exciting for all of us

Goal: *research-level challenges* for both humanists and computer scientist, opportunities for doing work that neither of them can do alone

Both humanists and computer scientists are first-class citizens here!

- **Bad:** computer scientists helping humanists to do their research
- **Good:** *computer scientists and humanists together doing multidisciplinary research that is interesting for both of them*

Computer scientists aren’t here for IT support, they are also doing research!
DHH19 Hackathon, Clarin focus on Parliamentary data

DHH19 Participant backgrounds:
- Computer Science: 9
- Social Sciences: 9
- Humanities: 20

People with at least moderate CS knowledge: 20
The Past and The Future in European Parliamentary Debates

How does the European Parliament talk about ‘the past’ and ‘the future’?

Political discourse and policies are heavily influenced by our understanding of the past and the future. The language used in political debates can reveal the morals, frames and ideologies of those who speak in these debates.

THE PAST AND THE FUTURE

The questions ‘What is Europe?’ and ‘What and How should Europe be?’ have been debated by invoking images of the future and memories of the past. An analysis of the EP plenary speeches by Construction Grammar Conceptual Network Analysis revealed inextricable links between speeches referring to the future, the past and the present.

A close analysis of a subcorpus containing speeches referring to the past with structural topic modelling revealed recurrent debates about the topic of a (common) European history as well as the topic of the institutional, cultural and territorial developments of the EU.
IN-DEPTH: TOTALITARIANISM

There is no significant difference between East and West in connecting Stalinism & National Socialism. The volume of discourse on the topic follows anniversary years and is perhaps also connected to the accession of former Soviet states in 2004.

IN-DEPTH: CLIMATE CHANGE

Climate change is a topic primarily related to the future. We can note the relative amount of speeches on climate change evolving over time, which is likely to correspond to the urgency of the topic on a global scale.
The virtual challenge...

Real life interaction in multidisciplinary mode is the heart of DHH hackathon. How to transfer it to online space is a question we have not yet tackled.

For more information on DHH, see:

Collaborations
Lonneke van der Plas, University of Malta (UM)

UPSKILLS, an Erasmus+ project that will foster research-based teaching

Funded by an Erasmus+ Strategic partnership

Consortium of 8 partners:
- University of Malta
- University of Belgrade
- University of Bologna
- University of Graz
- University of Rijeka
- CLARIN ERIC
- University of Zurich
- University of Geneva
- and several associate partners
Main aim:
Tackle skills gaps and mismatches in students of language-related disciplines

Rationale:
Linguists are needed in research and industry jobs
But need for transferable forward-looking skills, such as critical thinking and problem solving, knowledge of research design and data analysis, project management, and digital skills

How:
Innovative pedagogies such as online educational games
modular and blended learning
real-world applications (work-based learning)
integrating existing research and research infrastructures into teaching
• Intellectual outputs we will create:
  ○ Needs analysis
  ○ Guidelines on research-based teaching
  ○ Learning Content
  ○ Educational games

• Just started, will run for 3 years

• Several multiplier events planned for each IO
  ○ Needs analysis 4/’21
  ○ Guidelines on research-based teaching 10/’21
  ○ Learning Content 7/’22
  ○ Educational games 6/’23

• Very happy to welcome more people on the advisory board
  > have a say in what is needed most

• Will stay in touch!!!
Discussion: Towards a CLARIN Training suite...

- How easy to find, use were the resources?
- Was it more difficult to use resources from national consortia other than yours?
- What type of resources are most needed for teaching your subject?

- What type of support should be offered to improve your teaching material, to make it available to a broader community?
- How can we reach out to a larger number of teachers and lecturers?
- What types of events should we organise (e.g. training for tools, training for online teaching, virtual discussion groups)?
Closing remarks

Do not forget the poster style discussion session!

...and join the training mailing list

https://lists.clarin.eu/cgi-bin/mailman/listinfo/training