CLARIN in Training and Education

Koenraad De Smedt
CLARINO
University of Bergen, Norway
desmedt@uib.no

Iulianna van der Lek
CLARIN ERIC
Utrecht University, The Netherlands
i.vanderlek@uu.nl

Henk van den Heuvel
CLST / CLS
Radboud University
Nijmegen, The Netherlands
henk.vandenheuvel@ru.nl

Antonio Balvet
UMR STL 8163
Dept. of Language Sciences
University of Lille, France
antonio.balvet@univ-lille.fr

Maarten Janssen and Silvie Cinková
UFAL, Faculty of Mathematics and Physics
Charles University
Prague, Czechia
(janssen|cinkova)@ufal.mff.cuni.cz

Amelia Sanz
Dept. of Romance Studies
Complutense University
Madrid, Spain
amsanz@filol.ucm.es

Stavros Assimakopoulos
Institute of Linguistics
and Language Technology
University of Malta
stavros.assimakopoulos@um.edu.mt

Louis ten Bosch
CLST / CLS
Radboud University
Nijmegen, The Netherlands
louis.tenbosch@ru.nl

Abstract

To help realise its potential as the research infrastructure for language as social and cultural data, CLARIN is supporting the training of students and scholars in using its language data, tools and services. Lecturers and teachers in the CLARIN network have integrated CLARIN language resources into higher education programmes and other training activities. This paper showcases some recent courses and training initiatives, along with inventories and new learning materials, partly developed in EU-funded projects, which are accessible through the CLARIN Learning Hub. Each section briefly describes the motivation behind the initiative, the authors' experience, related efforts in the field, and future perspectives.

1 Introduction

CLARIN, the European research infrastructure for language as social and cultural data, offers data, tools and services to support a wide research community (Fišer and Witt, 2022). As an integral part of its outreach, CLARIN aims to "contribute to the education of new generations of data professionals for whom language data will increasingly demand advanced methods and tools" (De Jong et al., 2022, p. 55). The challenges and opportunities are clear. On the demand side, students and early-stage researchers increasingly need training in using digital language data and tools, not only to do classroom assignments and thesis projects, but also in preparation for careers that require technical know-how. On the supply side, a growing number of staff members in the CLARIN community have been applying both their competency and their CLARIN resources, tools and services to training.

CLARIN has an extensive Knowledge Infrastructure[1] to maintain contact with its users and to provide user support. It also has a Learning Hub[2] giving access to open educational resources, including online training modules to learn new skills and materials to design new university courses, training and workshops. Additionally, the hub contains best practices and guidelines developed in educational projects, such as UPSKILLS,[3] or created in collaboration with other research infrastructures.

Because CLARIN has been promoted as the European infrastructure for Digital Humanities (De Smedt et al., 2018), it has also been a long-time broker of educational services in that area, in particular through the Digital Humanities Course Registry[4] (Wissik, Wessels, and Fischer, 2022), which, in cooperation with DARIAH,[5] provides information about Digital Humanities courses in Europe. CLARIN has also expressed "recognition of the importance of students, teachers, lecturers, and trainers as users of CLARIN" (De Jong et al., 2022, p. 44) through several mechanisms, such as support for the development and sharing of training materials and the *Teaching with CLARIN Award*.

Some recent initiatives in training and education by various members of the CLARIN community were introduced at the workshop *Using CLARIN in Training and Education* at the CLARIN Annual Conference 2023. These initiatives will be further explained and discussed in the following sections.

2 Privacy by Design in Linguistic Research

The workshop *Privacy by Design in Linguistic Research* was set up to accommodate the request for knowledge and hands-on exercises aimed at PhD students concerning the implications of the General Data Protection Regulation (GDPR) for linguistic research. In many countries of the European Union, data stewards are appointed at universities to assist researchers in their research data management in general and in handling *personal* research data in particular. However, this is not the case in all EU countries, nor is sufficient expertise in GDPR-related aspects of specifically *linguistic* data available at all universities where data stewards are appointed.

Therefore, a workshop for PhD students at the start of their careers was set up at the Faculty of Arts of Radboud University (The Netherlands). The learning goal is a reflection on the relevant aspects of collecting, processing and sharing personal data in the context of linguistic research. The workshop is built around three components: an introduction to the GDPR and its impact on linguistic research, a group discussion of use cases, and a role-play. Parts 1 and 2 are based on data steward experience at the Faculty of Arts at Radboud University. Part 1 offers an introduction to the GDPR and its implications for linguistic research, covering the following topics:

- What is personal data?
- GDPR and Research
- Privacy by design, 8 principles[6]
- Informed consent
- Personal data and social media

Part 2 involves a discussion in breakout groups around a use case (which can be defined or fine-tuned upon request) addressing the following questions:

- Who are the stakeholders in this use case?
- Which existing data is used and which new data is generated?
- Which personal data is involved and what is the legal basis for collecting/sharing data?
- Is there a special category of data?

[1] https://www.clarin.eu/content/knowledge-infrastructure

[2] https://www.clarin.eu/content/learning-hub

[3] https://upskillsproject.eu/

[4] https://dhcr.clarin-dariah.eu

[5] https://www.dariah.eu

[6] See video at https://www.youtube.com/watch?v=f6MUwkEJzQ4&ab_channel=RadboudUniversity

- Where will the data be stored and with whom will the data be shared during the project?
- Where will the data be stored after the project and for how long?
- Which data will be shared after the project and with whom (access level, choice of licence)?

Part 3 is a role-play on another use case, based on the materials offered by DELAD[7] and explained in the CLARIN Impact Story *Navigating GDPR with Innovative Educational Materials*.[8] The role-play is based on the Data Protection Impact Assessment (DPIA),[9] which is a multi-stakeholder approach that provides a structured way of thinking about risks and protection measures. In the workshop scenario, the participants each took the role of one of the stakeholders in a decision about data sharing: researcher, ethics board member, representative of the data subjects, security/ICT expert, legal know-how, or data manager of the archive. The risk assessment is based on protection goals.

The workshop was organised at AITLA 2023,[10] where it was inspired, motivated and tuned towards the sensitive data typically associated with atypical speech that is dealt with at the CLARIN Knowledge Centre for Atypical Communication Expertise (ACE).[11] The workshop participants consisted of PhD students at the University of Siena and three teachers. They were divided into groups of three or four persons for the use case discussions, which turned out very lively. The teachers also took an active role and showed great involvement. In the evaluation, both the PhD students and the teachers reported that they learned a lot from the workshop and from each other. Recent requests to give the workshop at other places indicate its relevance.

3 Teaching Syntax with CLARIN Corpora and Resources

The COVID-19 pandemic highlighted the necessity of efficient, self-guided learning tools, especially in e-learning environments. However, manually designing and implementing self-correcting syntax learning activities for large student groups is labour-intensive and error-prone. To tackle the challenge of producing large volumes of reliable and consistent sets of self-correcting syntax exercises and quizzes, an automated solution is proposed, based on CONLL-U formatted Universal Dependencies corpora available from the LINDAT/CLARIAH-CZ repository,[12] as well as the UDPipe dependency parsing services.[13]

Our proposal is inspired by previous projects, where the integration of IT resources and tools for "grammar tutoring" has been explored. For example, the *IT-based Collaborative Learning in Grammar project* (Borin and Saxena, 2005), was a collaboration between Uppsala, Stockholm and Gothenburg universities.[14] In this context, the Stockholm-Umeå Corpus (SUC) and the Talbanken Swedish corpora were used as a source of syntactic and morphological annotations to automatically generate interactive grammar exercises, such as multiple-choice questions, part-of-speech tagging and rule writing exercises. Other similar initiatives, such as the VISL Corpus project (Bick, 2001, 2005a,b; Uibo and Bick, 2005; Wulff, 2006), or the French LORIA-led[15] METAL and GramEx projects (Bonnin et al., 2019; Colin, 2020; Perez-Beltrachini, Gardent, and Kruszewski, 2012), have applied syntactic parsers to generate grammar exercises and gamified activities.

Based on the corpus-to-quiz processing chain outlined in Figure 1, an initial subset of syntax quizzes covering different languages has been released.[16] The generated exercises use the General Import Format Template (GIFT) format, they are designed to be easily integrated into existing Moodle courses (or other similar platforms), and adapted to different learning scenarios. In its current version, the project focuses

[7] https://delad.ruhosting.nl/wordpress/dpia-role-play-with-video/

[8] https://www.clarin.eu/impact-stories/navigating-gdpr-innovative-educational-materials

[9] DPIA is a requirement under Article 35 of the GDPR, https://gdpr.eu/data-protection-impact-assessment-template/.

[10] https://aitla2023.wordpress.com/programma/

[11] https://ace.ruhosting.nl/

[12] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2895

[13] https://lindat.mff.cuni.cz/services/udpipe/

[14] With initial funding from the Swedish Agency for Distance Education (DISTUM).

[15] Laboratoire Lorrain d'Informatique Appliquée.

[16] https://github.com/abalvet/ACE/tree/main/v0.9/moodle-quizzes

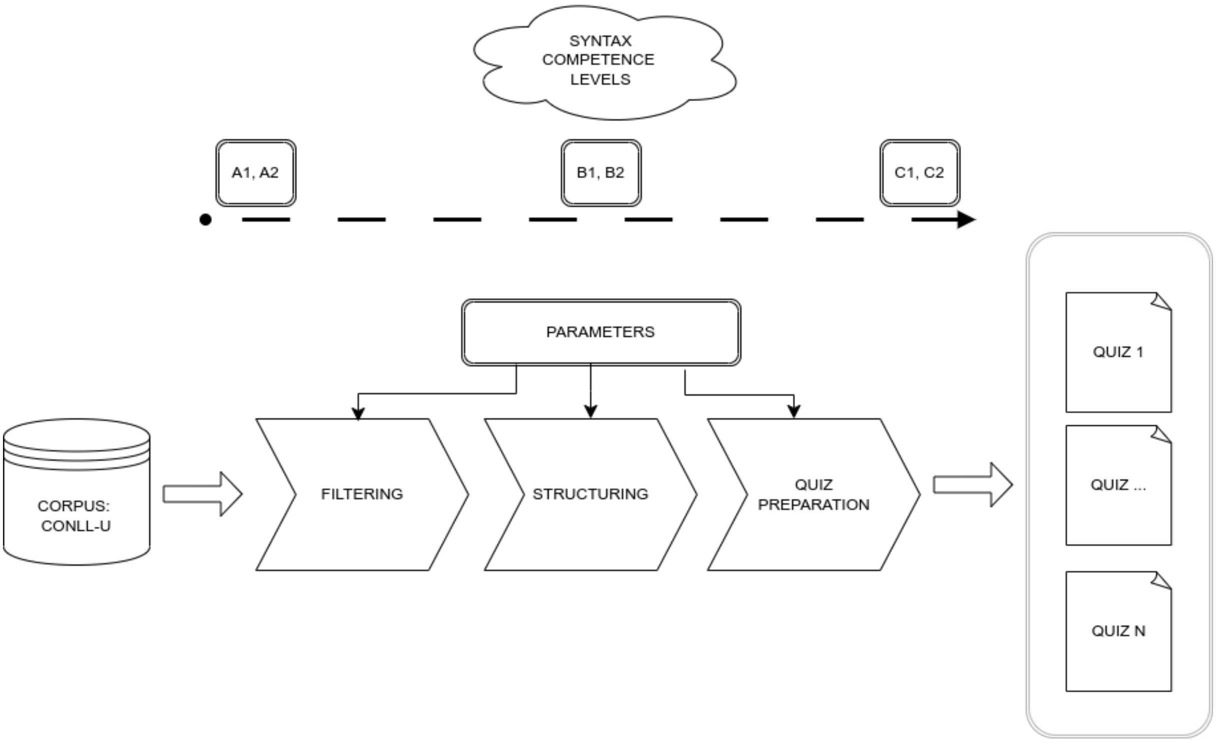


Figure 1: Corpus-to-quiz processing chain.

on French, but since it relies on CONLL-U corpora, the general procedure is adaptable to other languages, with limited overhead.

As illustrated in Figure 1, our approach leverages existing manually-verified syntactic annotations extracted from reference CONLL-U corpora, such as the French Treebank (Abeillé, Clément, and Toussenel, 2003) or Sequoia (Candito et al., 2014), to produce Moodle quizzes, aiming at university students in introductory syntax courses. Since the annotations have been extensively verified, syntactic parsing errors for French are extremely low, so that high-quality material is provided for the generated exercises. Consequently, the corpus-to-quiz processing chain significantly reduces manual editing and subjectivity in exercise creation.

The software for automating the generation of syntax quizzes is available on GitHub.[17] From a technical standpoint, we currently use Python scripts to process CONLL-U corpora to generate Moodle quiz questions. Different parameters allow instructors to filter specific features, in order to structure and automatically generate quizzes for different syntax competence levels.[18] By using the right conjunction of arguments, instructors can, for example, specify the number and type of distractors for a given syntax competence level. Other parameters determine the subset of part-of-speech (PoS) tags to target, the syntactic complexity of corpus-extracted sentences (e.g., simple versus complex sentences), and morphological or written form features for the targeted words, lemmas or PoS-tags.[19]

Looking ahead, we aim to generate new types of exercises, focusing on syntactic functions, constituency structure, and formal representations (i.e. dependency graphs and constituent trees). Plans are also underway to integrate AI tools like generative Large Language Models, as well as symbolic formalism-based ones,[20] for creating new sets of sentences to parse, or for providing learners with dynamic and personalised feedback. Since the proposed features are not natively available in standard

[17] `corpus2quiz` at `https://github.com/abalvet/ACE`

[18] Syntax Competence Levels range from 'A1' (beginner) to 'C2' (expert), in the spirit of the European Common Reference Framework linguistic competence levels.

[19] Regular expressions such as `^.*ment$` can be used in conjunction with PoS-tag filters to target sentences where a noun such as *complé**ment***, or an adverb *rapide**ment*** occur.

[20] E.g. ELVEX, `https://github.com/lionelclement/Elvex`, by L. Clément (Univ. Bordeaux).

Moodle distributions, a range of Learning Tools Interoperability (LTI) compliant web services will have to be developed. These web services will provide educators with new activities that can be seamlessly integrated into existing Moodle (or any other LTI-compliant platform) courses.

4 Learning Programming for Language-Related Studies

Handling digital text and quantitative language data for study and scholarship often presupposes knowledge and skills in programming. McGillivray et al. (2020) note that "while the humanities have developed a core set of methods and techniques for the rigorous interpretation of their sources, traditionally they lack training in the core subjects of modern data, computer and information science". Even though off-the-shelf tools and services for language processing exist, experience shows that black-box software is not always well understood, is not always capable of dealing with data in all its variety and formats, and does not support all possible angles of investigation. In practice, humanities students and scholars need to broaden their training by acquiring computer programming skills.

For working with digital language data, as for data science in general, two programming languages currently stand out: Python and R. Python has gained popularity because of its smooth learning curve and the many available packages, including modules for natural language processing (NLP) and machine learning. Although a wealth of Python courses is available, few of them specifically target students in language-related programmes such as linguistics, language studies, digital humanities and cognitive science. At the University of Bergen (Norway), Python is taught as part of an revamped introductory course on Natural Language Processing (NLP) at bachelor's level. After a few years of experience in teaching this course, it was decided to make the main course materials publicly available as an open, free-standing, web-based tutorial,[21] findable on the CLARIN Learning Hub,[22] DARIAH-CAMPUS[23] and the DH Course Registry.[24]

The core of the course consists of Jupyter notebooks that combine working Python code snippets with explanatory text and exercises. Its main pedagogical strategy is learning by example. The notebooks demonstrate basic language processing, quantitative data analysis and visualisation, all at an introductory level. The examples use language data from CLARIN and other relevant sources to the largest possible extent. Ideally, the course should be presented by a teacher and the exercises should be supervised, but the modules are also suitable for self-study. The online course does not offer solutions to all exercises, but a small quiz has been added providing solutions after submitting answers to the questions. Progressing from simple to more complex programming, the course treats the following main topics:

1. String operations, including search and substitution with regular expressions. Some attention is given to the properties and treatment of non-Latin scripts, relevant for studying various languages.
2. Tokenisation, n-grams and frequencies from text. In this context, basic functionality of the Natural Language Toolkit[25] is introduced.
3. Representation, analysis and visualisation of quantitative data in dataframes. This is relevant for processing quantitative corpus data and results from surveys and experiments, but it does not go as far as a statistics course.
4. Accessing information on the web, including reading plain text, extracting text from HTML, importing CSV tables and accessing APIs.
5. Establishing a workflow in which raw quantitative data is obtained from a data source (such as corpus frequencies), normalized and finally visualised as plots and tables, which are exported for direct inclusion in a LaTeX article. Workflow management is very useful for all students who need to write papers and theses using empirical data.

[21] Citation: *Introduction to programming for NLP with Python*. Web-based course at the University of Bergen. `https://mitt.uib.no/courses/38115`. Licensed under Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

[22] `https://www.clarin.eu/content/introduction-programming-nlp-python`

[23] `https://dariah-campus-7dh56510k-dariah.vercel.app/resource/posts/introduction-to-programming-for-nlp-with-python`

[24] `https://dhcr.clarin-dariah.eu/courses/view/753`

[25] NLTK, `https://www.nltk.org`

In presenting these themes, attention is paid to algorithmic thinking, suitable data structures and expressive programming constructs. Given the target audience and the desire to keep the course introductory and compact, many advanced programming approaches are left out. During the course, students learn how to process various language-related data, including the following resources accessible through CLARIN.

1. Texts at the Oxford Text Archive, findable via the CLARIN VLO, are used to demonstrate reading plain text webpages into Python.
2. A dataset of Norwegian compounds with *korona-* or *corona-* extracted from the Norwegian Newspaper Corpus (De Smedt, 2021) is used for reading, sorting and counting language data.
3. An overview of NorGramBank resources (Dyvik et al., 2016) containing the number of sentences and the year for each text, serves as an example for grouping and summing data.
4. An online table with Slovenian occupations in masculine and feminine forms from CLARIN.SI[26] is used in a data structure for simple gender-dependent translation.
5. Frequencies from the Corpus of London Teenage Language (COLT), accessed through Corpuscle (Meurer, 2012), are imported and processed.
6. The DH-LAB at the National Library of Norway (Birkenes, Johnsen, and Kåsen, 2023) is demonstrated.

Experience with the course shows that students get started quickly because they do not have to install any software. The combination of text and code in Jupyter notebooks makes the course materials largely self-explanatory, but their use in classroom teaching is still preferable, especially for absolute beginners. In addition, UiB has organised local group sessions led by teaching assistants who offer help with the exercises, which is well appreciated. Although the course was primarily designed for students in linguistics and language studies, it has also attracted students from information science, communication and media studies, cognitive science, digital culture, computer science, computer technology, digital security, medicine, law, administration and organisation studies, and film and TV production.

After a few iterations, the course is now fairly stable, but further improvement is always possible. For example, the practical activities could include more language resources from CLARIN, but not many fully open datasets make good examples for beginners. The notebooks are currently on the Google Colaboratory (Colab) platform, which offers hosted runtime as a free service. This makes the notebooks very easy to use, as they do not require students to install any other applications but a web browser. Colab has some limitations, which currently do not present problems, but its conditions for use may change. Alternative platforms, such as Binder, Kaggle or Deepnote, have been successfully tested but are not essentially better. Some students prefer to run the Jupyter notebooks on their own machines, using a platform such as Visual Studio Code. Ideally, the code and runtime should be hosted on an open academic Jupyterlab cloud service, but so far, none have been found that do not present administrative or technical hurdles.

Other courses with somewhat similar learning goals and target audiences exist, but none use CLARIN resources. Folgert Karsdorp and Maarten van Gompel offer an introductory open web course on *Python programming for the Humanities*.[27] That course is also presented as Jupyter notebooks, progressing from easy to more difficult, and includes exercises. The notebooks are downloadable but an executable version is also available on Binder. The course offers text preprocessing, tokenisation, frequencies, access to plain texts from the web, and extraction of text from HTML pages. However, it progresses rather quickly to more advanced concepts, including machine learning, whereas we chose to defer this topic to a separate course, thus allowing the introductory course to be more manageable for beginners. Karsdorp and Van Gompel's course presents some concepts which do not seem important for most beginners, such as execution timing, but it fails to treat non-Latin alphabets and scripts, which are clearly relevant for language studies and the humanities; also, regular expressions are not explained in any detail, despite their tremendous usefulness in text processing.

[26] http://hdl.handle.net/11356/1347

[27] https://www.karsdorp.io/python-course/

The University of Oslo offers a course on *Algorithmic Thinking for the Humanities*, the materials of which are accessible online.[28] This course is a mixed bag. After introducing basic concepts of computer programming and some very useful text processing and text statistics with Python, the UiO course proceeds to correlation with plotting, which unfortunately uses examples that are outside the language domain. However, the course does treat non-Latin alphabets, accessing texts on the web, and NLTK. Then, there is a substantial section on Game Theory, which does not seem relevant for most students in the language sciences but could be suitable for information science. The code snippets are not presented in Jupyter notebooks, which means there is a small practical hurdle for their execution.

Quite a few other courses exist that are either too general or too specialised. Clearly, the choice between different course plans depends on training goals, level and programme context. In any case, programming is a complex skill to acquire, so step-by-step instruction with good examples is essential to keep students motivated and on schedule.

5 NLP Annotation for Digital Scholars

Computational linguistics is gradually moving away from corpora in the traditional sense and Natural Language Processing (NLP) pipelines. The data used in large language models are vast; NLP would be too slow to run over such data, and is no longer needed in purely statistics-based practical applications. For that reason, NLP pipelines are no longer a key topic in computational linguistics. However, research in domains like DH or less-resourced languages (LRLs) typically relies on qualitative studies over relatively modest-size corpus data, where the annotations added by NLP make searches much more flexible.

DH often deals with texts for which off-the-shelf NLP tools have a sub-par performance because they are written in historical or local variants, or in a genre too remote from standard language, such as poetry. Moreover, NLP tools are non-existent for many LRLs. Since corpus annotation is no longer a key topic for NLP research, it is often up to the researchers to work towards the appropriate NLP tools. Yet researchers often lack the technical expertise to do so.

We are involved in an ongoing effort to provide *NLP annotation for scholars* as a pedagogical concept. The goal is to demonstrate to our students that creating training data and training NLP models can be a challenging task, but it doesn't require the level of technical expertise they may believe is necessary. We, furthermore, teach our students to define their research questions in common linguistic terms, formulate them in terms of Universal Dependencies, and query them with a corpus query language.

The first step is to select a portion of text in the pre-established corpus collection and manually annotate it either from scratch or after pre-processing by a (sub-optimal) tagger. For the manual annotation, the system will ask for the correct lemma, POS tag, and morphological features for each word. By default, syntactic dependencies will not be annotated in this step since scholars without linguistic background easily get discouraged at the mention of syntax. At the same time, those scholars usually have a sufficient grasp of morphological categories. The manual annotation is iteratively used to train and improve the tagger and to facilitate further annotation with improved pre-processing. Deep learning means fewer training data are needed to reach an adequate tagger accuracy.

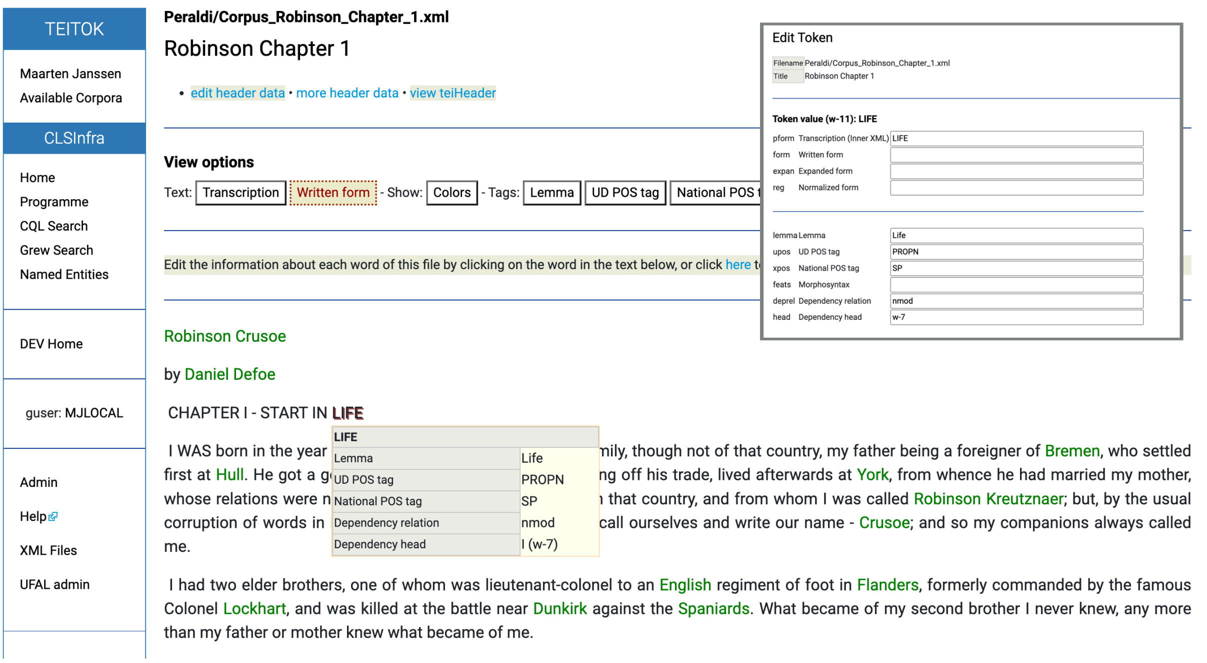
As a graphical interface for the courses, we use TEITOK (Janssen, 2016), an online environment explicitly designed to integrate NLP annotation in complex document structures and make the resulting corpus searchable across the different annotation layers. TEITOK is explicitly designed not only to make corpora searchable but also to edit them. For example, the platform provides a button to an NLP pipeline via the graphical interface, without using command line tools by hand. Furthermore, using the graphical interface, it lays out several easy ways to correct errors made by the automatic NLP pipeline.

There are many other courses on corpus annotation, but our use of a GUI in the learning process means that students can focus on the task of enriching the annotations without being distracted by technical details. Although most students have sufficient knowledge about morpho-syntax, deciding on the correct tags for actual occurrences in a corpus always takes some getting used to and applying a tagset consistently even more so. Understanding how tags are assigned helps improve not only annotation but also querying

[28] `https://uio-ccse.github.io/algoritmisk-tenkning-humanister/intro.html`

the corpus once annotated, since students will be more aware of the kinds of decisions made in the process and the kind of errors NLP tools make.

TEITOK is not built upon plain text documents the way traditional corpus tools are, but instead uses full-fledged documents in the Text Encoding Initiative (TEI), which can contain various types of additional annotations, such as alignment to a facsimile or an audio or video track, typesetting information, footnotes, etc. Therefore, students are asked to bring any type of documents they want to use as the basis for their annotated corpus. Provided the documents are in a well-established format, they can typically be converted to TEITOK. This means that students can work directly on their research data and see the search results directly in the complete original context. Figure 2 shows a document brought by one of the students with the automatic annotation, and as an inset the edit mode to manually correct annotation errors.



Peraldi/Corpus_Robinson_Chapter_1.xml

Robinson Chapter 1

- edit header data • more header data • view teiHeader

Edit the information about each word of this file by clicking on the word in the text below, or click here t

Robinson Crusoe

by Daniel Defoe

CHAPTER I - START IN LIFE

LIFE	
Lemma	Life
UD POS tag	PROPN
National POS tag	SP
Dependency relation	nmod
Dependency head	l (w-7)

I WAS born in the year [...]mily, though not of that country, my father being a foreigner of Bremen, who settled first at Hull. He got a g[...]ng off his trade, lived afterwards at York, from whence he had married my mother, whose relations were n[...] that country, and from whom I was called Robinson Kreutznaer; but, by the usual corruption of words in [...]call ourselves and write our name - Crusoe; and so my companions always called me.

I had two elder brothers, one of whom was lieutenant-colonel to an English regiment of foot in Flanders, formerly commanded by the famous Colonel Lockhart, and was killed at the battle near Dunkirk against the Spaniards. What became of my second brother I never knew, any more than my father or mother knew what became of me.

Edit Token

Filename	Peraldi/Corpus_Robinson_Chapter_1.xml
Title	Robinson Chapter 1

Token value (w-11): LIFE

pform	Transcription (Inner XML)	LIFE
form	Written form	
expan	Expanded form	
reg	Normalized form	
lemma	Lemma	Life
upos	UD POS tag	PROPN
xpos	National POS tag	SP
feats	Morphosyntax	
deprel	Dependency relation	nmod
head	Dependency head	w-7

Figure 2: Document view and token edit in TEITOK.

From the very start, the teaching process provides an automatic tagger and lemmatiser, which will become increasingly accurate with more training data. The NLP pipeline used for this is UDPIPE,[29] a cutting-edge parser that typically scores high, although other NLP pipelines could also be used.

We are currently working on integrating the course with the newly designed UDMorph system,[30] an infrastructure for morpho-syntactically tagged corpora following the UD standards, parallel to the UD infrastructure for treebanks, but for corpora that do not have dependency relations. By contributing the data to UDMorph, the newly trained tagger will automatically become available to the research community for online use, or for download and subsequent local use. Attribution data are generated by the system so as to make sure the students get credit for their work. If the training data are submitted to UDMorph as well, the data become available to the community, where it can be used to train potentially even more accurate models.

These hands-on courses are typically well received by the students since they learn what NLP tools can do, which errors they make and how to correct them. Furthermore, students learn how to use NLP tools without having to do any programming or deal with the command line interface, which makes it much more feasible to cover significant ground during the course. Finally, the students can immediately

[29] https://lindat.mff.cuni.cz/services/udpipe/
[30] https://lindat.mff.cuni.cz/services/teitok-live/udmorph/

see the results, and experience the improvements in the tagger first-hand. Students can use our research material with added annotations to formulate and run queries on their own data.

6 DH-Course Registry: A Bridge Between Infrastructures, DH Master's Degrees and Industry

European Research Infrastructures, such as CLARIN and DARIAH, are well-placed to provide a conduit between industry and education, given their wide-ranging contacts with both communities. DARIAH and CLARIN already collaborate closely within the context of the DH Course Registry,[31] maintained by both infrastructures (Wissik, Wessels, and Fischer, 2022). The registry, a platform to collect metadata on digital humanities programmes across Europe, has been a glue between the research infrastructures and the DH programmes, leading to a new joint initiative. In the spring of 2023, we set out to explore effective strategies and best practices for facilitating the career success of graduates of DH master's programmes in the private sector.

The skills acquired within Digital Humanities (DH) postgraduate degrees are interdisciplinary and, therefore, transferable, something that has been recognised among larger multinational companies. Moreover, a strong humanities background and familiarity with DH methods can benefit the commercial sector. Yet among small and medium enterprises (SMEs), employing a graduate from a field still in its relative infancy compared with more traditional disciplines can be considered a risk. Therefore, an effort should be made to highlight DH skills, while it also becomes necessary to identify the gaps between the current provision of training among DH scholars at the master's level and the needs of companies and future employers of DH graduates.

Thanks to the metadata of the DH courses and programmes described in the database, representatives of 25 DH programmes in Europe have been proactively contacted to investigate the skills gap between the DH curriculum and the job market requirements and identify best practices in setting up effective internship models with companies and organisations from the commercial sector and/or GLAM. The joint DARIAH–CLARIN workshop (Sanz et al., 2023) at the DARIAH annual conference in June 2023 unveiled opportunities for both infrastructures to enhance their roles in several areas and cultivate synergies between infrastructures, cultural heritage, industry and academia. Specifically, there is a clear call to enhance the infrastructures' involvement at the university policy level, increasing awareness of the profound impact of AI and the emerging job profiles within the digital humanities field, e.g. engineering linguist, intelligence analyst, data analyst, data consultant, data scientist, bot designer, games designer, digital technician, heritage digitiser or project manager.

One of the main findings of the workshop, reported by Paul Spence from King's College London, was that DH graduates with a mixture of critical thinking, coding, digital design skills, research software engineering and analysis skills, and UX/UI design skills are much sought after not only in the cultural heritage sector but also the commercial one. Paul Spence also indicated that it is quite common for students to use their dissertation or school project portfolios as a platform to launch new careers. Furthermore, Maria Goicoechea, from the Complutense University of Madrid, pointed out that big technology companies are heavily reliant on linguistic profiles to support their engineering teams throughout developing web, platform, and software features. In addition to possessing a solid understanding of technical requirements beyond Python, candidates are expected to demonstrate proficiency in key digital techniques such as regular expressions, command lines, and markdown. Based on these first insights, the working group envisages further in-depth analysis, a white paper, and a workshop involving representatives from academia, research infrastructures, cultural heritage, and industry.

7 CLARIN in the UPSKILLS Project

The skills gap and employability topics were also addressed in the UPSKILLS[32], a recently completed Erasmus+ strategic partnership project (2020–2023), which aimed to identify and tackle the gaps and mismatches in skills for linguistics and language students. Employment prospects for graduates in language-

[31] https://dhcr.clarin-dariah.eu

[32] https://upskillsproject.eu

related disciplines (linguistics, foreign languages, language pedagogy, translation and interpreting) are still mainly focused on teaching positions or positions as translators. This starkly contrasts their potential employability given the omnipresence of language and communication in society and the number of companies that make language their main business. Seeing how not only smaller companies but also technology giants – such as Google, Amazon, and Facebook – continuously work with language data, it is no surprise that the demand for digital research skills in language-related domains is constantly growing. However, linguistics and language-related university curricula are rarely oriented towards such skills, which means their graduates tend to be poorly prepared for the corresponding careers.

To this end, the UPSKILLS consortium partners, including CLARIN, ran a detailed needs analysis and developed a new curriculum component alongside supporting learning content to be embedded in existing programmes. More specifically, eleven learning blocks were developed, focusing on research, data acquisition and data handling skills, which can be browsed and/or downloaded from the project website.[33] The learning content is complemented by guidelines for research-based teaching (including the use of research infrastructures, such as CLARIN, in teaching),[34] as well as a set of educational games that can be used for both instruction and testing.[35] These UPSKILLS learning blocks were mainly designed for instructors of courses in linguistics and language-related subjects; however, students can also use the materials autonomously as long as they remember that they are not typical self-study courses. In the following, we zoom in on the two learning blocks developed by CLARIN.[36]

7.1 Automatic Speech Recognition and Forced Alignment (ASR/FA)

Education in the speech sciences must bridge a gap between students with a linguistic background on the one hand and technologically oriented studies involving the speech signal on the other hand. This gap is only increasing due to AI's rapidly progressing and profound impact on Automatic Speech recognition (ASR) and other speech science and technology areas. Inspired by the research-based teaching used in the UPSKILLS project, we aimed to design an overarching course to connect linguistic backgrounds with technologically advanced research domains, taking an integrative perspective. The resulting learning block *Automatic Speech Recognition and Forced Alignment* (6 ECTS) provides an example of this integration, providing speech science/technology tailored to scholars with a non-technical background. The theoretical and practical activities are based on two well-known textbooks in the field (J. Holmes and W. Holmes, 2001; Jurafsky and Martin, 2023).

The learning content is modular, enabling other lecturers to cherry-pick and adapt it based on their needs. After introducing the underlying principles of ASR, students are taught the distinction between different types of ASR architectures, i.e. classical architectures, such as acoustic models, lexicons and language models, versus more recent approaches, such as the AI-inspired deep-learning end-to-end models. Through an active learning approach, students learn to identify basic concepts in deep learning, the challenges in the field and how to select the suitable approaches for building ASR and reasoning about ASR for specific purposes and conditions, e.g. designing welcoming robots in noisy income halls in museums, help bots in station halls, carebots in care assistance or medical environments. The learning block ends with an optional thesis (3 ECTS), for which nearly ten options for diverse topics for student projects are presented. These topics invite the student to investigate the role of Acoustic and Language Models and lexicons in ASR, the user role, and the role of ASR embedded in a more extensive human-machine dialogue system.

Although this learning block has not yet been piloted at the BA level, it has been designed based on a similar research-based course, which one of the authors has taught for years at the MA level at Radboud University in the Netherlands. Substantial attention has been paid to the balance between conceptual

[33] https://upskillsproject.eu/deliverables/io3/upskills_learning_materials/

[34] https://upskillsproject.eu/deliverables/io2/

[35] https://upskillsproject.eu/deliverables/io4/

[36] CLARIN's work in the UPSKILLS project aligns with international initiatives like the European Open Science Cloud (EOSC) and FAIRisFAIR, which promote the adoption of open science and research data management based on the FAIR guiding principles Wilkinson et al. (2016) or scientific data management across all domains, disciplines and levels. For a complete overview of CLARIN's contribution to the UPSKILLS project, please refer to the CLARIN Learning Hub at https://www.clarin.eu/content/learning-hub.

topics (e.g., Bayes) and implementation topics (e.g., Viterbi). Although a fully theoretical track is an option, successful implementation of such a course largely depends on the students' programming skills (in Python), the accessibility of data sets and the flexibility of the curriculum. For example, students should be allocated enough time to gather data from their experiments. Furthermore, all experiments must be conducted using existing and easily accessible datasets. Finally, due to the rapid progress in this research field, the learning content, including the quiz questions embedded in each ASR/FA learning block, needs to be updated regularly, at least yearly.

7.2 Introduction to Language Data: Standards and Repositories

The UPSKILLS needs analysis (Gledić et al., 2021) revealed that linguistics and language-related programmes seldom include language data standards and research data repositories specifically in their learning outcomes. This motivated the design of a learning block to introduce learners (teachers and students) to research infrastructures and language data repositories, including CLARIN, and their role in the linguistic research data life cycle and management in the context of open science and FAIR data principles. The learning block consists of six units supplemented by a glossary.

1. Introduction to the Language Resource Life Cycle and Management
2. How Research Data Repositories Help Make Language Data FAIR
3. Finding and (Re)using Language Resources in the CLARIN Repositories
4. Citing Language and Linguistic Data
5. Legal and Ethical Issues Language Data Collection, Sharing and Archiving
6. Student Project

The learning outcomes of each unit target basic research data management and FAIR skills inspired by the *FAIRsFAIR Teaching and Training Handbook for Higher Education Institutions* (Engelhardt et al., 2022) and *The Open Handbook of Linguistic Data Management* (Berez-Kroeker et al., 2022). By integrating research infrastructures, language data and tools into teaching, educators can bridge the gap between theoretical knowledge and practical aspects of linguistic research data management, equipping students with the necessary skills and competences to thrive in the evolving landscape of open science and data-driven research.

After a general introduction to language resources, their life cycle and management, learners are acquainted with the FAIR data principles and how these can be applied to corpus creation, sharing and archiving. Unit 3 consists of presentations, hands-on tutorials and practical assignments, demonstrating how the Virtual Language Observatory (VLO)[37] can be used to search, find and process digital text collections with suitable tools from the Language Resource Switchboard.[38] Furthermore, learners are introduced to the CLARIN Resource Families[39] and shown how to query large families of corpora, such as Parlamint, through the available concordancers. Unit 4 gives learners an overview of the current language data citation practices and different types of persistent identifiers research data repositories assign to deposited language resources. Finally, Unit 5 helps learners identify some common legal and ethical issues involved in language data collection, sharing and archiving, e.g. GDPR principles applied in research, copyright exceptions for text and data mining, dealing with sensitive data, and selecting appropriate licences when sharing and archiving language resources. The learning block concludes with an example of a student project (for two ECTS) that aims to teach students how to design, compile, and archive a corpus of bank bulletins using the CLARIN repositories.

Teachers can teach and adapt the whole learning block or cherry-pick only those presentations and learning activities (tutorials, handouts, exercises, quizzes, and assignments) that match the learning outcomes of a specific programme, course or student project. Some presentations and assignments can also be used as self-study materials. The content in this block is in H5P format, which can be reused in any

[37] https://vlo.clarin.eu/

[38] https://switchboard.clarin.eu/

[39] https://www.clarin.eu/resource-families

content or learning management system supporting this format, e.g. Moodle, Brightspace or Drupal. If taught as a whole, the module can amount to six ECTS or more, but this is just an estimate because only parts of this learning block have been piloted so far in a few programmes, workshops and summer schools.[40]

We conclude this section by presenting the first impressions collected via the pilots. Firstly, the lecturers appreciated the repositories learning block as it fills the knowledge gap on research infrastructures and language data, as no comprehensive courses are available. Second, the level of modularity and flexibility allowed the lecturers to download only the learning content that was in line with the overall learning outcomes of their current courses. The interactive quizzes and the glossary were also appreciated because they helped the students understand and retain the technical terms pertaining to Research Data Management (RDM). Finally, although the learning block targets students at the BA and MA levels, it also helped first-year PhD students learn new things related to RDM, FAIR data principles, and legal and ethical issues in data collection, sharing and archiving. An online guide, *Integrating research infrastructures into teaching: Recommendations and best practices* (Lek et al., 2023) complements the learning content giving a more detailed introduction to the CLARIN infrastructure and how it can be used in teaching. It also directs the instructors towards relevant learning content and activities available on the UPSKILLS Moodle platform. To increase the findability of the materials within the CLARIN community, all eleven UPSKILLS learning blocks were uploaded to the CLARIN.SI repository and, hence, they are also discoverable via the Virtual Language Observatory.[41] Last but not least, the metadata of the learning content has been added to the SSH Open Marketplace.[42] to make the content findable by trainers working in other SSH domains.

8 Discussion

For over a decade, teachers and researchers associated with CLARIN and assisted by the governance and administration in CLARIN ERIC, have been nurturing and passing on know-how to students at all levels, from undergraduate programmes to researcher training. Through close collaboration with the CLARIN trainers' network, ambassadors, and user involvement activities, as well as participation in the UPSKILLS project, CLARIN substantially increased its training and educational initiatives (including the Helsinki Digital Humanities Hackathon, ESU and MEDAL Summer Schools) and its production of training and learning materials over the past year. Additionally, CLARIN's representation in relevant training communities and task forces at the level of the European Open Science Cloud (EOSC), Research Data Alliance (RDA) and SSH Open Cloud (SSHOC) also support this endeavour.

The current paper reports on some recent initiatives presented at the CLARIN Annual Conference 2023. Through discussions with the participants of the teachers' workshop and interactions with teachers and students who took part in UPSKILLS events between 2021 and 2023, a few common questions have been identified which could be further addressed in future discussions and initiatives.

The first question concerns the inclusion of digital language data and tools in language-related disciplines and at which level they should be taught in the academic programmes. On the one hand, tools that transparently integrate language data in linguistics teaching, such as the automation of syntax quiz generation mentioned above, may be employed at any stage. On the other hand, it has been commonly agreed that targeted training in handling digital language data and tools should be introduced gradually in the programme. At the beginner's level, an overview of the use of language resources and tools may be appropriate for students in language-related disciplines. In this context, exposure to CLARIN repositories and services, including the VLO and user-friendly tools for resource exploration (such as corpus search), is natural. Still, even at an early stage, any use of digital resources must always be accompanied by training in proper data citation (Conzett and De Smedt, 2022). At more advanced levels, it is desirable

[40] So far, these pilots include the Corpus Analysis course at Leiden University, the Netherlands, and the Research Methods and Analysis Techniques in Digital Linguistics programme at the University of Ljubljana, Slovenia. In addition, parts of the block were taught at the MEDAL Summer School in Corpus Linguistics, the Lancaster Corpus Linguistics conference, and the UPSKILLS summmer school.

[41] https://hdl.handle.net/11356/1865

[42] https://marketplace.sshopencloud.eu/training-material/yweKFs

to tailor training to the students' individual needs in master's and PhD projects. Advanced students often do their own data collection, so sharing, licensing and legal issues must be foregrounded, which indicates the need for courses such as *Privacy by design*, while courses such as the aforementioned *NLP annotation for scholars* may also be very relevant at this stage.

The discussions with the UPSKILLS lecturers concluded that the implementation of research-based teaching, including the use of research infrastructures and language resources, depends on the lecturers' ability to find the balance between teaching fundamental research and more practical skills (e.g. corpus-based pedagogy and data-driven learning using corpora and tools), and on the students' background, the level of digital literacy, the study load and the flexibility of the curriculum. Additionally, the lecturers acknowledged the benefits of using existing infrastructures, such as the VLO, the Resource Families of open corpora and integrated concordancers, to create a safe environment for students to experiment and discover new corpora and tools on their own. Teaching materials and examples of learning activities and tutorials on these topics can be found in the UPSKILLS course *Introduction to Language Data: Standards and Repositories course*, including a general introduction to the FAIR data principles (Wilkinson et al., 2016) and how they can be applied in corpus development, sharing and archiving.

A second legitimate question that emerged refers to the inclusion of computer programming in the language sciences and humanities curriculum. Teaching introductory programming to undergraduates may be beneficial in promoting their algorithmic thinking and a general understanding of the possibilities and limitations of NLP, AI and quantitative approaches. It will also empower students to do their own data handling and analysis for term papers and thesis projects. Furthermore, basic programming skills are useful even for scholars and professionals who avoid writing code, but who likely need to know and use specific programs when interacting with technical people, in order to ensure mutual understanding and effective cooperation. Computer programming is, however, a complex skill that may take time to acquire,[43] so care must be taken to avoid overwhelming students, but instead focus on what is most useful in the context of their studies.

A third common theme is employability and whether the skills acquired in DH, linguistics and language programs help graduates find jobs in a fast-changing workplace landscape, characterised by the rapid spread of generative AI. A balance must be found between classical academic pedagogy grounded in discipline-specific knowledge, philosophy of science and critical thinking on the one hand, and the transfer of career-focused skills, on the other hand. Fortunately, the two are not incompatible. An American study on the application of linguistics education to the workplace found that linguistics graduates have high employability and are "able to link their domain-specific linguistic skills to the transferable skills they use in the workplace" but also encourages us to "better articulate the transferable skills that are gained within a linguistics program" (Gawne and Cabraal, 2023). This suggestion is commensurate with the findings and recommendations from the UPSKILLS Needs Analysis and CLARIN–DARIAH cooperation that the present paper has reported on.

Finally, the laudable effort by universities too open up their courses and educational resources to external participants constitutes in no way a persistent offer. In our distributed and changing landscape, it is to be expected that courses evolve or are discontinued in relation to locally or globally changing needs and opportunities, while new initiatives must find their way to potential target groups. Therefore, closer cooperation between academic institutions and CLARIN aimed at persistent offers and up-to-date information in training and education may be worth pursuing. For instance, academic institutions can use the CLARIN Learning Hub to raise awareness about new training and educational initiatives and open educational resources that others can use for upskilling or reskilling purposes. Moreover, the Learning Hub can promote available expertise within the network and exchange best practices in teaching and training using language resources and technologies available through the CLARIN infrastructure.

Since creating learning resources from scratch is both time-consuming and expensive, it is crucial to design them with reuse in mind, using best practices in SSH. One such methodology is the Skills4EOSC FAIR-by-design approach, which has been adapted to meet the needs of the CLARIN community and

[43] The art of programming, as well as text writing, is bound to be transformed by generative AI.

proposed during the Bazaar at CLARIN2023.[44] Additionally, it is important to publish these resources in open access whenever possible and include their metadata in the SSH Open Marketplace. By doing so, other trainers in the community can easily find, cite and reuse the resources for new instructional purposes. Finally, lecturers and teachers can use the DH Course Registry to increase awareness about new DH programmes and courses across Europe and beyond, identify gaps in the existing DH syllabi and design new programmes that increase the employability opportunities of DH graduates.

Acknowledgments

The work reported in this paper has been funded in part by the Research Council of Norway under grant 295700 *CLARINO+* and a consortium of institutions led by the University of Bergen, by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004984 *Computational Literary Studies Infrastructure*, by the Ministry of Education, Sports and Youth of the Czech Republic under the programme LM2023062 *LINDAT/CLARIAH-CZ* and by the European Union's Erasmus+ strategic partnership for Higher Education action under grant agreement 2020-1-MT01-KA203-074246 *UPSKILLS – UPgrading the SKIlls of Linguistics and Language Students*.

References

Abeillé, Anne, Lionel Clément, and François Toussenel (2003). "Building a treebank for French." In: *Treebanks: Building and using parsed corpora*. Ed. by Anne Abeillé. Text, Speech and Language Technology 20. Dordrecht: Springer, pp. 165–187. DOI: `10.1007/978-94-010-0201-1_10`.

Berez-Kroeker, Andrea L., Bradley McDonnell, Eve Koller, and Lauren B. Collister (2022). *The Open Handbook of Linguistic Data Management*. The MIT Press. DOI: `10.7551/mitpress/12200.001.0001`.

Bick, Eckhard (2001). "The VISL System: Research and applicative aspects of IT-based learning." In: *Proceedings of the 13th Nordic Conference of Computational Linguistics (NODALIDA 2001)*. ACL. URL: `https://aclanthology.org/W01-1702.pdf`.

– (2005a). "Grammar for fun: IT-based grammar learning with VISL." In: *CALL for the Nordic languages*. Ed. by Peter Juel Henrichsen. Vol. 30. Copenhagen Studies in Language, pp. 49–64. URL: `https://edu.visl.dk/pdf/CALL2004.pdf`.

– (2005b). "Live use of Corpus data and Corpus annotation tools in CALL: Some new developments in VISL." In: *Nordisk Sprogteknologi 2004*. Ed. by Henrik Holmboe. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004. Copenhagen: Museum Tusculanums Forlag, pp. 171–185. URL: `https://edu.visl.dk/pdf/corpus_and_CALL_form.pdf`.

Birkenes, Magnus Breder, Lars Gunnarsønn Bagøien Johnsen, and Andre Kåsen (2023). "NB DH-LAB: a Corpus Infrastructure for Social Sciences and Humanities Computing." In: *CLARIN Annual Conference Proceedings 2023*. Ed. by Krister Lindén, Jyrki Niemi, and Thalassia Kontino. CLARIN Annual Conference Proceedings. ISSN: 2773-2177. Leuven: CLARIN ERIC, pp. 30–34. URL: `https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf`.

Bonnin, Geoffray, Estelle Perry, Charlotte Baraudon, and Stéphanie Fleck (2019). "Design participatif d'un tableau de bord enseignant." In: EIAH workshops. Paris. URL: `https://hal.science/hal-02476952/document`.

Borin, Lars and Anju Saxena (2005). "Grammar, Incorporated." In: *CALL for the Nordic languages*. Ed. by Peter Juel Henrichsen. Copenhagen Studies in Language 30. Copenhagen: Samfundslitteratur, pp. 125–145. URL: `https://samfundslitteratur.dk/bog/call-nordic-languages`.

Candito, Marie, Guy Perrier, Bruno Guillaume, Corentin Ribeyre, Karën Fort, Djamé Seddah, and Éric Villemonte de La Clergerie (2014). "Deep syntax annotation of the Sequoia French treebank." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, pp. 2298–2305. URL: `http://www.lrec-conf.org/proceedings/lrec2014/pdf/494_Paper.pdf`.

[44] Making the CLARIN Training Materials FAIR-by-Design, `https://www.clarin.eu/sites/default/files/CLARIN2023_Bazaar_12.pdf`

Colin, Émilie (2020). "Traitement automatique des langues et génération automatique d'exercices de grammaire." PhD thesis. Université de Lorraine. URL: http://docnum.univ-lorraine.fr/public/DDOC_T_2020_0059_COLIN.pdf.

Conzett, Philipp and Koenraad De Smedt (2022). "Guidance for Citing Linguistic Data." In: *The Open Handbook of Linguistic Data Management*. Ed. by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. Open Handbooks In Linguistics. Chapter: 11. The MIT Press. DOI: 10.7551/mitpress/12200.003.0015.

De Jong, Franciska, Dieter Van Uytvanck, Francesca Frontini, Antal van den Bosch, Darja Fišer, and Andreas Witt (2022). "Language matters." In: *The infrastructure for language resources*. Ed. by Darja Fišer and Andreas Witt. Berlin, Boston: De Gruyter, pp. 31–58. DOI: 10.1515/9783110767377-002.

De Smedt, Koenraad (2021). "Contagious 'Corona' Compounding by Journalists in a CLARIN Newspaper Monitor Corpus." In: *Selected Papers from the CLARIN Annual Conference 2020*. Linköping Electronic Conference Proceedings 180. Ed. by Costanza Navarretta and Maria Eskevich, pp. 83–92. DOI: 10.3384/ecp18010.

De Smedt, Koenraad, Franciska De Jong, Bente Maegaard, Darja Fišer, and Dieter Van Uytvanck (2018). "Towards an Open Science Infrastructure for the Digital Humanities: The Case of CLARIN." In: *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*. CEUR Workshop Proceedings, pp. 139–159. URL: http://ceur-ws.org/Vol-2084/paper11.pdf.

Dyvik, Helge, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse, and Martha Thunes (2016). "NorGramBank: A 'Deep' Treebank for Norwegian." In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. Paris: ELRA, pp. 3555–3562. URL: http://www.lrec-conf.org/proceedings/lrec2016/pdf/943_Paper.pdf.

Engelhardt, Claudia et al. (2022). "How to be FAIR with your data – A teaching and training handbook for higher education institutions." In: URL: https://hdl.handle.net/10468/12492 (visited on 01/24/2024).

Fišer, Darja and Andreas Witt, eds. (2022). *CLARIN. The infrastructure for language resources*. Digital Linguistics 1. Berlin, Boston: De Gruyter. 810 pp. DOI: 10.1515/9783110767377.

Gawne, Lauren and Anuja Cabraal (2023). "Linguistics education and its application in the workplace: An analysis of interviews with linguistics graduates." In: *Language* 99.1, e35–e57. DOI: 10.1353/lan.2023.0003.

Gledić, Jelena, Jelena Budimirović, Maja Đukanović, Tanja Samardžić, Sandra Jukić, Adriano Ferraresi, Gaia Aragrande, Lonneke van der Plas, Iulianna van der Lek, and Nađa Soldatić (2021). *Survey of business sectors hiring linguists and language professionals*. UPSKILLS Intellectual Output. DOI: 10.5281/zenodo.5030890.

Holmes, John and Wendy Holmes (2001). *Speech Synthesis and Recognition*. 2nd ed. London: CRC Press. DOI: 10.1201/9781315272702.

Janssen, Maarten (2016). "TEITOK: Text-Faithful Annotated Corpora." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), pp. 4037–4043. URL: http://www.lrec-conf.org/proceedings/lrec2016/pdf/651_Paper.pdf.

Jurafsky, Dan and James H. Martin (2023). *Speech and Language Processing*. 3 (draft). URL: https://web.stanford.edu/~jurafsky/slp3/.

Lek, Iulianna van der, Darja Fišer, Tanja Samardzic, Marko Simonovic, Stavros Assimakopoulos, Silvia Bernardini, Maja Milicevic Petrovic, and Genoveva Puskas (2023). *Integrating research infrastructures into teaching: Recommendations and best practices*. UPSKILLS Intellectual Output. DOI: 10.5281/zenodo.8114406.

McGillivray, Barbara et al. (2020). *The challenges and prospects of the intersection of humanities and data science: A White Paper from The Alan Turing Institute*. DOI: 10.6084/m9.figshare.12732164.v5.

Meurer, Paul (2012). “Corpuscle: A new corpus management platform for annotated corpora.” In: *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*. Ed. by Gisle Andersen. Studies in Corpus Linguistics 49. Amsterdam/Philadelphia: John Benjamins, pp. 31–49. URL: https://books.google.no/books?id=RJmPfmQq_20C&pg=PA31.

Perez-Beltrachini, Laura, Claire Gardent, and German Kruszewski (2012). “Generating grammar exercises.” In: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. ACL, pp. 147–156. URL: https://aclanthology.org/W12-2017.pdf.

Sanz, Amelia, Vicky Garnett, Tom Gheldof, Edward Gray, Adeline Joffres, Iulianna van der Lek, and Anna Woldrich (2023). “Digital Humanities and Industry: Identifying Employment Niches. A first overview on challenges and potential solutions.” In: DARIAH Annual Event, Budapest, Hungary, 6-9 June 2023. DOI: 10.5281/zenodo.8071224.

Uibo, Heli and Eckhard Bick (2005). “Treebank-based research and e-learning of Estonian syntax.” In: *Proceedings of Second Baltic Conference on Human Language Technologies*. Tallinn, pp. 195–200. URL: https://edu.visl.dk/pdf/HLT05_Uibo_Bick.pdf.

Wilkinson, Mark D. et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship.” In: *Scientific Data* 3.160018. DOI: 10.1038/sdata.2016.18.

Wissik, Tanja, Leon Wessels, and Frank Fischer (2022). “The DH course registry: A piece of the puzzle in CLARIN’s technical and knowledge infrastructure.” In: *CLARIN. The infrastructure for language resources*. Ed. by Darja Fišer and Andreas Witt. Berlin, Boston: De Gruyter, pp. 389–408. DOI: 10.1515/9783110767377-015.

Wulff, Anette (2006). “VISL in Danish schools.” In: *English Teaching: Practice & Critique* 5.1, pp. 142–147. URL: https://core.ac.uk/download/pdf/50642933.pdf.