CLARIN

Selected papers from the
CLARIN Annual Conference 2023
Leuven, Belgium

Selected papers from the

CLARIN Annual Conference 2023

Leuven, Belgium, 16–18 October 2023

edited by Krister Lindén, Thalassia Kontino and Jyrki Niemi

Linköping Electronic Conference Proceedings 210
ISSN 1650-3740 (digital) • ISSN 1650-3686 (print) 2023
ISBN 978-91-8075-740-9 (PDF)

Introduction

Krister Lindén
Program Committee Chair
University of Helsinki,
Finland
krister.linden@helsinki.fi

Darja Fišer
Executive Director of CLARIN ERIC
Institute of Contemporary History,
Slovenia
darja.fiser@clarin.eu

This volume presents the highlights of the twelfth CLARIN Annual Conference in 2023. The conference took place from 16 to 18 October 2023 as a hybrid event with Leuven, Belgium, as the venue.

The CLARIN Annual Conference is the main annual event for those working on the construction and operation of CLARIN across Europe, as well as for representatives of the communities of use in the humanities and social sciences.

CLARIN, the Common Language Resources and Technology Infrastructure, is a virtual platform that is accessible to everyone interested in language. CLARIN offers access to language resources, technology, and knowledge, and enables cross-country collaboration among academia, industry, policy-makers, cultural institutions, and the general public. Researchers, students, and citizens are offered access to digital language resources and technology services to deploy, connect, analyse and sustain such resources. In line with the Open Science agenda, CLARIN enables scholars from the Social Sciences and Humanities (SSH) and beyond to engage in and contribute to cutting-edge, data-driven research based on language data in a range of formats and modalities.

The infrastructure is run by CLARIN ERIC[1], a consortium of participating countries and institutes that was established in 2012 and has grown considerably in size since. Currently there are 24 member countries, 2 observers, and more than 100 associated research institutions, which are all encouraged and supported to be represented at the annual conference. The event is central for the CLARIN community and is one of the crucial instruments for CLARIN to function as a knowledge hub. At the conference, consortia from all participating countries and the various communities of use meet in order to exchange ideas, experiences and best practices in using the CLARIN infrastructure. Moreover, CLARIN 2023 was also intended for the wider humanities and social sciences communities in order to exchange ideas and experiences within the CLARIN infrastructure. This includes the design, construction and operation of the CLARIN infrastructure, the data, tools and services that it contains or for which there is a need, its actual use by researchers, its relation to other infrastructures and projects, and the CLARIN Knowledge Infrastructure. In January 2023, a call[2] was issued for which 52 abstracts were submitted. The authors of the submissions to the main conference sessions represented 19 CLARIN ERIC countries.

All submissions were reviewed anonymously by three reviewers (PC members and reviewers invited by PC members). Out of the 52 submitted abstracts, 37 submissions were accepted for presentation at the conference (acceptance rate 0.71). The 37 accepted submissions were grouped in the following subjects:

- Corpora (4 papers)
- Infrastructure (6 papers)
- Metadata and annotation (3 papers)
- ParlaMint (3 papers)
- Tools (4 papers)
- Posters (17 papers)

[1] http://www.clarin.eu
[2] https://www.clarin.eu/content/call-abstracts-clarin-annual-conference-2023

Krister Lindén and Darja Fišer 2023. Introduction. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. i–iv. DOI: https://doi.org/10.3384/ecp210

The accepted contributions were published in the online Conference Proceedings[3].

As in 2018–2022, a PhD Students Session was organised as a combination of a one-minute pitch followed by a highly interactive poster session. In total, 15 posters were presented. New in CLARIN 2023 was an award for the Best PhD Poster Presentation, for which the National Coordinators' Forum acted as the selection committee.

The 2023 edition of the CLARIN Annual Conference was shaped as a hybrid event. Live and on several screens, more than 230 registered participants were able to follow the quality content and learn what CLARIN is about. The conference programme contained both traditional conference elements and novel items better suited to the virtual set-up:

- **Keynote 'Lost in Meaning – Found in Translation: Natural Language Understanding with Multilingual Data'** by Jörg Tiedemann introduced the OPUS ecosystem of multilingual data, and presented experiments, studies and developments that the ecosystem has enabled both within and outside of the FoTran project.
- **Keynote 'Ethical Issues of Generative AI'** by Laurence Devillers outlined characteristics of generative AI systems and the main ethical issues related to the systems, and presented recommendations to tackle the issues, crafted in CNPEN (French National Pilot Committee for Digital Ethics).
- Regular sessions of accepted conference papers consisted of presentations followed by Q&A.
- During the **PhD Students Session**, PhD students presented posters with their work in progress: studies supported by or contributing to the CLARIN infrastructure. The aim of the session was to put the spotlights on the next generation of researchers and to enable them to receive feedback on their work from CLARIN experts.
- The **Best Student Poster Presentation Award** was presented during the closing ceremony of the conference and given to Nina C. Rastinger from the Austrian Centre for Digital Humanities for her poster *Re-Reading Lists in Historical Newspapers: Digital Insights into an Overlooked Text Type*.
- The workshop **Using CLARIN in Training and Education**[4] showcased the latest training and educational initiatives, as well as learning materials created within the CLARIN network.
- As usual, the **CLARIN Bazaar**[5] provided an informal setting for conversations with CLARIN people and a space to showcase ongoing work and exchange ideas. The presenters were grouped together by topic to encourage interaction:
  - CLARIN K-centres
  - Training and Education
  - Software Demos and Tools
  - Projects, Collaborations and CLARIN Consortia

After the event, CLARIN published a rich set of relevant materials:

- The complete conference programme and most of the slides presented: https://www.clarin.eu/content/programme-clarin-annual-conference-2023
- Recordings of keynotes and presentations, available on the CLARIN YouTube channel: https://www.youtube.com/playlist?list=PLlKmS5dTMgw19rdP_23ZFtmLXCApVT4H0
- Conference summary: https://www.clarin.eu/content/clarin-annual-conference-2023-summary

[3] https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf
[4] https://www.clarin.eu/event/2023/using-clarin-training-and-education-clarin2023
[5] https://www.clarin.eu/content/clarin-bazaar-2023

The authors of the accepted papers, student submissions, participants of the workshop 'Using CLARIN in Training and Education' as well as CLARIN-funded projects were invited to submit papers for the post-conference proceedings. Out of the conference contributions, 17 substantially extended contributions were accepted as Selected papers from the CLARIN Annual Conference 2023.

We would like to thank all PC members and reviewers for their efforts in evaluating and re-evaluating the submissions, Thalassia Kontino from the CLARIN Office for her indispensable support in the process of preparing these proceedings, and our colleagues at the Linköping University Electronic Press, who ensured that the digital publication of this volume came about smoothly. In order to support the programme chair and the programme committee in the organisation of reviewing and programme planning, a programme subcommittee was established. To ensure continuity, the programme chair from the preceding year's conference was a member of the committee. The members of the 2023 programme subcommittee were Krister Lindén (chair), Tomaž Erjavec, Monica Monachini, Maciej Piasecki and Vincent Vandeghinste.

Members of the Programme Committee for the CLARIN Annual Conference 2023:

- **Krister Lindén, University of Helsinki, Finland (Chair)**
- Starkaður Barkarson, Árni Magnússon Institute for Icelandic Studies, Iceland
- Lars Borin, University of Gothenburg, Sweden
- António Branco, Universidade de Lisboa, Portugal
- Koenraad De Smedt, University of Bergen, Norway
- Tomaž Erjavec, Jožef Stefan Institute, Slovenia
- Cristina Grisot, Switzerland
- Eva Hajičová, Charles University Prague, Czechia
- Marinos Ioannides, Cyprus University of Technology, Cyprus
- Monica Monachini, Institute of Computational Linguistics "A. Zampolli", Italy
- Costanza Navarretta, University of Copenhagen, Danemark
- Maciej Piasecki, Wrocław University of Science and Technology, Poland
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center, Greece
- German Rigau, Basque Center for Language Technology, Spain
- Gijsbert Rutten, Leiden University, The Netherlands
- Kiril Simov, IICT, Bulgarian Academy of Sciences, Bulgaria
- Inguna Skadiņa, University of Latvia, Latvia
- Marko Tadić, University of Zagreb, Croatia
- Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania
- Vincent Vandeghinste, Instituut voor de Nederlandse Taal (Dutch Language Institute), the Netherlands & KU Leuven, Belgium

- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences, Hungary
- Joshua Wilbur, University of Tartu, Estonia
- Tanja Wissik, Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences, Austria
- Andreas Witt, University of Mannheim, Germany
- Friedel Wolff, South African Centre for Digital Language Resources, North-West University, South Africa
- Martin Wynne, University of Oxford, United Kingdom

Additional reviewers for this volume:

- Riccardo Del Gratta, Institute of Computational Linguistics "A. Zampolli", Italy
- Benito Trollip, South African Centre for Digital Language Resources, North-West University, South Africa

Contents

Teaching Syntax with CLARIN Corpora and Resources

Antonio Balvet
Department of Language Sciences
Univ. Lille, CNRS, UMR 8163 - STL - Savoirs Textes Langage
F-59000 Lille, France
antonio.balvet@univ-lille.fr

Abstract

The recent COVID-19 pandemic has brought online learning to the forefront for learners and teachers. As a consequence, the demand for self-paced and adaptive learning resources has reached unprecedented levels. Fortunately, universities had been using e-learning platforms such as Moodle, or other SCORM-compliant LMS, which has helped make the transition from on-site to on-line learning. However, teachers still have had to design and implement assessment activities in the form of self-correcting activities (true/false, multiple answer questions, mark the words, fill in the blanks questions, etc.). This step has proved to be a major hurdle in the on-site to on-line learning transition, since designing and, most of all, manually editing formative and evaluative assessment activities is a very labour-intensive task. In this article, we present a framework that takes advantage of the corpora and resources available from the LINDAT / CLARIAH-CZ Data & Tools platform in order to generate quizzes and other activities related to syntax. After some background on using NLP for teaching grammar, we present our corpus-to-quiz processing chain, and we outline preliminary results on deploying automatically generated French syntax quizzes in the classroom.

1 Introduction

The domain of education has undergone a profound transformation in recent years, a change dramatically accelerated by the COVID-19 pandemic and the associated lockdown periods. The shift from traditional classroom settings to on-line platforms has not only redefined the way we teach and learn, but also underlined the urgent necessity for deploying innovative educational tools and learning scenarios. In this context, the demand for self-paced, responsive, and personalised learning environments has become more pronounced than ever before.

Even before the pandemic, e-learning platforms such as Moodle were integral to the pedagogical strategies of many universities. Moodle, in particular, allows instructors to include third-party learning applications, thanks to the compliance with inter-operability protocols such as the Sharable Content Object Reference Model (SCORM) or Learning Tools Interactivity (LTI) protocols. These protocols power widgets and other services, such as H5P.org's widgets, or collaborative tools for text or image creation. As a consequence, Moodle is currently being used all around the world, with over 43,000 running instances in Europe alone.[1] Moodle, and other Learning Management Systems (LMS), therefore provided a welcome foundation for the sudden and necessary transition from on-site to on-line learning. However, this transition also exposed significant challenges, particularly in the realm of language education. For instructors, one of the most daunting tasks has been designing and implementing effective assessment activities under strict deadlines. Traditional forms of assessment, such as printed quizzes and tests, have had to be transformed into self-correcting activities such as true/false, multiple choice questions, and other interactive activities. While advantageous in their ability to provide instant feedback and adaptability, such self-correcting activities require a substantial amount of time and effort, especially when catering to large groups of students, who are sometimes facing bandwidth –or even hardware– access restrictions at home. This is particularly true in the context of syntax classes, a fundamental, yet complex, aspect

[1] https://stats.moodle.org/

Antonio Balvet 2023. Teaching Syntax with CLARIN Corpora and Resources. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 1–12. DOI: https://doi.org/10.3384/ecp210

of linguistics. Syntax poses a unique challenge in an online learning environment, since many students often have unpleasant memories of traditional grammar classes, which focus more on rote memorisation of exceptions rather than understanding underlying principles and parsing methods. In this context, the ability to automatically generate quizzes that are both challenging and adaptive is crucial in a modern educational landscape to enable students to progress at their own pace and according to their individual learning paths. This personalised approach is not just a response to the logistical challenges posed by online education; it is a pedagogical strategy aimed at deepening students' understanding of syntax, enhancing their engagement, and ultimately fostering a more positive and effective learning experience.

In this publication, we outline our proposal to transform the way syntax is taught and evaluated by leveraging the considerable collective research efforts that have gone into Natural Language Processing (NLP) and the extensive manually annotated corpora available from the LINDAT/CLARIAH-CZ Data & Tools platform. One very tangible aspect of our proposal is the design and implementation of a processing chain that converts Universal Dependencies corpora into ready-to-integrate Moodle quizzes. Our processing chain automates the generation of syntax quizzes and other formative activities, letting educators focus on the design of learning scenarios and on providing their students with the necessary feedback, and not on the minutiae of setting up individual quiz questions, thus making the process of deploying formative and evaluative self-correcting activities less prone to error and subjectivity. It also provides a sound foundation for establishing carefully curated collections of standardised syntax tests that educators will be able to share and deploy at their respective teaching sites. Finally, it also offers students a more engaging and stimulating learning environment, thanks to immediate scoring and progression feedback. Although the current version of our corpus-to-quiz processing chain focuses on French syntax quizzes, generated from the French Treebank (Abeillé et al., 2003) and Sequoia (Candito et al., 2014) corpora, indexed on the CLARIN infrastructure and available from universaldependencies.org, the principle presented here is applicable to any Universal Dependencies CONLL-U[2] formatted corpus, with limited overhead. After making the case for syntax "consciousness-raising" and the need for consistent Syntax Competence Levels (SCL) assessment, we provide an outline of projects that have relied on NLP solutions and reference corpora for teaching grammar. We then present our corpus-to-quiz processing chain, and how the generated formative and evaluative activities can be integrated into Moodle courses. Lastly, we report preliminary results of deploying such automatically generated quizzes for both distance and on-site learning, and we offer first insights of deploying standardised automatically generated formative and evaluative activities in the context of our introductory syntax course.

2 Syntax Consciousness-Raising and e-learning

The recent COVID-19 pandemic has highlighted the need for self-paced and adaptive learning resources. Although universities around the world had been using e-learning platforms prior to this event, teachers were still faced with a very labour-intensive task, since designing and implementing self-correcting assessment activities for potentially large groups of learners, in a distance learning context, proved very time-consuming. Furthermore, designing and implementing such assessment activities by hand is error-prone and subjective by nature. Our proposal aims at optimising the design and implementation of such assessment activities by taking advantage of publicly available syntax-annotated corpora. In doing so, we propose a concrete approach to promote syntax 'Consciousness-Raising' among students, thus falling in line with projects such as VISL[3] and the METAL/Gramex presented below.

2.1 From grammar to syntax Consciousness-Raising

The concept of "Consciousness-Raising" (C-R) in grammar teaching, as defined by Rutherford (1987), Schmidt (1990), and Ellis (2016), marked a significant shift in second language acquisition (SLA) pedagogy, especially for English as a Foreign Language (EFL). The status of explicit grammar teaching

[2]CONLL is a tabular format popularised by the Computational Natural Language Learning conferences. CONLL-U corpora are those corpora which follow Universal Dependencies annotation guidelines.

[3]Visual Interactive Syntax Learning project.

has been the subject of intense debates in the last decades, to the extent that many countries have abandoned explicit grammar teaching in SLA and in first language acquisition. However, C-R in grammar can be seen as a transition from a focus on pure grammatical forms to a more nuanced understanding of form-meaning relations: it is often defined as an inductive approach that improves learners' awareness by encouraging them to notice, analyse, and internalise the grammatical aspects of the target language within context, thus fostering a deeper, more meaningful acquisition of language structures. We propose to transpose the notion from the domain of **grammar** to that of **syntax** teaching, in the context of higher education.[4]

We define syntax as the set of rules governing the production and decoding of well-formed sentences in any natural language, associated with an explicit, as well as practical, knowledge of discovery procedures (syntactic tests) and parsing strategies that allow students to look beyond individual surface words and identify the underlying patterns –both at the structural and functional dependency levels– at play in any particular sentence. To this end, we rely on corpus-derived automatically generated exercises, which are described in more detail below. In order to properly achieve syntax C-R, we first need to design and implement sound and consistent learning scenarios designed, precisely, to make students notice, analyse, and internalise syntactic structures, rules and procedures. The proposed learning scenarios, and corresponding sets of self-correcting formative and evaluative syntax activities are designed with a well-defined Syntax Competency Levels (SCL) progression in mind, based on different linguistic features to determine 6 basic SCL, from "beginner" to "intermediate" and "advanced" (see section 2.2).[5] Supplementing traditional "chalk-and-talk" teaching methods with extensive sets of automatically generated quizzes represents a transformative approach to teaching syntax. Our proposal is predicated on the belief that a deep and intuitive understanding of syntax–not just as a set of abstract rules, but as a living structure of communication–is essential for educators who aim to impart linguistic knowledge effectively. For teachers, particularly those instructing in their native language, such a framework offers a dual benefit: it not only reinforces their syntactic proficiency, but also equips them with a methodology to later teach syntax at primary and secondary school levels. By engaging in targeted and adaptive exercises, future teachers can better identify and address their own knowledge gaps, ensuring that they can provide a more comprehensive and nuanced education to their students. For learners of a second language (L2), especially higher education L2 learners, we see targeted exercises focused on syntax competence as an invaluable means of internalising the language they seek to master. Unlike traditional methods that may emphasise the rote memorisation of grammar rules, syntax-focused responsive and adaptive exercises facilitate deeper engagement with the language, encouraging learners to recognise and apply grammatical structures in context.

2.2 Syntax Competence Levels

A critical aspect of our approach is the concept of "Syntax Competence Levels" (SCL). This concept, inspired by the well-established language proficiency levels of the Common European Framework of Reference (CEFR), aims to categorise and define various stages of syntactic understanding and abilities in learners, ranging from A1 (beginner) to C2 (advanced). The British Council's "Grammatical Proficiency Levels" offer a parallel, yet broader, framework, encompassing a wide range of grammatical knowledge including tense, mood, and the usage of prepositions in phrasal verbs. In contrast, "Syntax Competence Levels" concentrate on tasks such as part-of-speech identification, the identification of syntactic structures, and the recognition of functional relationships, preferably in sentences taken from authentic documents rather than forged ones. From a practical standpoint, at the A1 level, the focus is on basic part-of-speech identification and understanding simple sentence structures. This foundational level is crucial for beginners, as it lays the groundwork for more advanced syntactic concepts. As learners progress to higher levels, the complexity of the structures and the depth of syntactic analysis increase.

[4] We address groups of students aiming, in their vast majority, at becoming primary or secondary teachers. These students are native French speakers who often lack a proper understanding of the structural aspects of French syntax.

[5] The definition of SCL is still a work in progress. So far, we have defined 6 SCL, ranging from A1 to C2, based on lexical parameters (e.g. frequency, familiarity), structural (e.g. simple vs complex sentences) and functional features, as well as overall sentence readability and topic.

At the intermediate levels (B1 and B2), learners engage with more complex sentences, delving into intricate syntactic structures and beginning to explore functional relationships between sentence elements (e.g. verb-dependent subordinate clauses, recursive structures). The advanced levels (C1 and C2) challenge learners to master sophisticated aspects of syntax. This includes a deep understanding of nuanced syntactic structures, the ability to analyse and interpret complex functional relationships, and the skill to apply this knowledge in varied linguistic contexts. Such a progression allows for a more targeted approach in teaching and assessing syntax, a fundamental component of linguistic competence that is often challenging for learners. We see this specialisation as particularly beneficial in an academic setting, where a deep understanding of syntax is crucial for students pursuing linguistics or related fields. By implementing these Syntax Competence Levels in our self-correcting activities, we ensure that the generated formative and evaluative activities are not only tailored to the individual learner's current level of understanding, but also provide a clear pathway for progression.

3 Corpora and NLP tools for Computer-Assisted Language Learning

In the domain of Computer-Assisted Language Learning (CALL), several projects have explored the use of reference corpora and NLP tools, to automatically derive vocabulary questions and grammar exercises. These include Fill In the Blank (FIB), Shuffle exercises,[6] true/false and multiple answer question quizzes. Such seminal projects as Bitzer et al. (1961) illustrate how computer technology has, almost from the very start, been seen as a tool to assist in –and optimise– the process of second language acquisition. More recent approaches (Aldabe et al., 2006; Borin & Saxena, 2005; Heck & Meurers, 2022; Lee & Seneff, 2007; Mitkov et al., 2006; Perez-Beltrachini et al., 2012; Smith et al., 2010) are particularly relevant to the present discussion. Following Perez-Beltrachini et al. (2012), we distinguish approaches that rely on reference corpora to extract relevant material from those that rely on some form of word or text generation under specific linguistic constraints.

Our corpus-to-quiz processing chain belongs to the former type of approach. It aims both at reducing manual edition to a minimum and at overcoming the subjectivity (and errors) associated with manually created exercises. Here, we report experiments conducted on French, although the approach can be extended to other languages, since CONLL-U corpora are readily available from universaldependencies.org for over 100 different languages. Our proposal falls in line with a larger ongoing effort to integrate technology –and more recently, Artificial Intelligence– into pedagogy. The integration of technology in syntax C-R has been marked by significant milestones: years before Moodle and other LMS were a widespread commodity, substantial work had been done to leverage Natural Language Processing (NLP) tools for educational purposes. In this context, key projects such as VISL and the LORIA-led Gramex/METAL[7] project are of particular relevance, each contributing uniquely to the advancement of "smart" syntax-aware learning activities.

3.1 The Visual Interactive Syntax Learning project

The VISL[8] project, initiated in the early 2000s, stands as a pioneering effort in this domain. Spearheaded by a consortium led by the University of Southern Denmark, VISL used a Constraint Grammar (CG) parser architecture, designed to accommodate various languages (Bick, 2001, 2004, 2015; Karlsson, 1990; Karlsson et al., 2011; Uibo & Bick, 2005; Wijlff, 2006). Based on this CG parser, large electronic corpora were produced, allowing researchers from the VISL consortium to implement a platform very similar to the well-known "Sketch Engine" presented in Kilgarriff et al. (2008). In addition to syntax-aware concordancers, the VISL consortium also designed several gamified activities based on CG-parsed corpora, for different languages.[9] This project was instrumental in establishing large, syntactically parsed corpora, which formed the basis for a range of syntax-learning tools, among which syntax-aware con-

[6] Fill In the Blank exercises require the user to produce the correct form in a given context, whereas Shuffle exercises require them to produce a valid sentence based on a set of predefined words.

[7] Models and learning analytics for language learning.

[8] Visual Interactive Syntax Learning.

[9] A syntactic labyrinth, as well as a syntactic Tetris and other syntactic games were implemented as Java applets, which means they are unfortunately not functional anymore.

cordancers and an assortment of interactive exercises –some of them gamified– derived from CG-parsed corpora. The innovative approach of the VISL project to grammar learning, emphasising visual and interactive elements, marked a significant departure from traditional syntax teaching methods.

3.2 METAL and Gramex: teaching French grammar with NLP tools

The METAL and Gramex projects, led by LORIA,[10] illustrate another approach to the adoption of NLP tools to teach French grammar. At this point, it is worth underlining that the very first version of Gramex was dedicated to generating "grammar exercises used for language learning i.e., grammar exercises whose syntax and lexicon are strongly controlled" (Perez-Beltrachini et al., 2012). Therefore, it relied on a symbolic, formalism-driven text generation platform to produce sentences based on a set of preestablished patterns to generate FIB and Shuffle grammar exercises. The latest evolution of the Gramex/METAL project saw a change of perspective on the issue of grammar and syntax C-R: the project now relies on high-precision, robust parsers for the automatic parsing of preexisting documents[11] to generate grammar exercises. Notwithstanding these technical changes, the emphasis is still on grammar C-R, with a main target composed of primary and secondary school pupils. This effort contributed significantly to the development of effective automated teaching tools for young learners.[12] In recent years, the Gramex/METAL project expanded its scope by forming collaborations with Metz and Nancy-based primary schools[13] to design comprehensive grammar teaching material associated with targeted exercises. In addition, since the project uses authentic documents (texts and on-line resources), it allows educators to generate exercise questions directly from specific content or URLs, to provide a more engaging and stimulating learning experience. Gramex also addresses collaborative learning by designing exercises that can be carried out by student groups, to foster a more interactive and supportive learning environment. This integrated approach to grammar instruction aims to make grammar learning more accessible, interactive, and effective for primary and secondary school pupils.

4 Corpus-to-quiz: leveraging reference corpora for syntax C-R

The work carried out by VISL and METAL/Gramex represents a significant conceptual foundation on which our project is built. These projects have shown promising perspectives on the use of annotated corpora and NLP for syntax C-R, but unfortunately, the developed applications and activities are no longer functional in the case of VISL,[14] or not yet available in the public domain at the time of writing, in the case of Gramex/METAL. Our approach follows the steps of VISL in that it relies on sentences extracted from existing reference corpora, one important difference being that the corpora used for French exhibit high-quality manually revised syntactic annotations.[15] Our Python-based corpus-to-quiz processing chain, together with a first version of ready-to-integrate part-of-speech and functional relationship identification quizzes, is available from the ACE-Annotated Corpora for online Exercises github repository. Both VISL and Gramex/METAL rely on high-precision parsers, while our approach capitalises on the availability of richly manually annotated corpora from the universaldependencies.org platform. In contrast to Gramex/METAL, our approach aims to address syntax C-R at the university level, targeting a demographic that has largely been overlooked in previous projects: that of aspiring primary and secondary school teachers, as well as French as a Foreign Language instructors. As stated earlier, although French is the main target in its current version, our framework is designed to be adaptable to different languages and educational levels. This adaptability is achieved by leveraging the CONLL-U formatted corpora for Universal Dependencies, making it possible to extend our methodology to a wide range of languages and syntactic structures. Finally, our approach focusses on the exercise generation

[10] Laboratoire Lorrain de Recherche en Informatique et ses Applications.

[11] Authentic documents taken from reference texts, or forged sentences that illustrate a specific grammar topic.

[12] See Perez-Beltrachini et al. (2012) and Colin (2020).

[13] See Gramex.

[14] All developed applications and gamified activities were developed as Java applets, which have been officially deprecated in 2017.

[15] More precisely, the French Treebank has undergone comprehensive manual verification, while a hybrid manual/automatic process has been implemented for Sequoia.

aspect, while all authentication procedures and learning analytics logging are currently being handled by Moodle, thus keeping software infrastructure and personal data management issues to a minimum. Nevertheless, in the coming years, we have plans to develop third-party LTI-compliant web services to specifically overcome limitations in the learning profiling capabilities of most LMS, including Moodle, as well as user interactivity: in its current version, Moodle only natively supports true/false or multiple-answer quizzes, which is a serious limitation for the purpose of our approach to syntax C-R.

4.1 The corpus-to-quiz processing chain

As stated above, our "corpus-to-quiz" processing chain is designed to transform annotated linguistic data into engaging formative and evaluative activities. Our approach relies on CONLL-U formatted corpora, specifically the French Treebank (FTB) and the Sequoia corpora. The French Treebank initiated in the late 1990's, adopting the Penn Treebank methodology.[16] As a consequence, the French Treebank (ca. 700,000 tokens) was essentially a manual annotation project, based on sentences sampled from *Le Monde*, a reference French newspaper. The initial annotations were based on an XML schema in order to capture constituency structure together with functional relationships. It is worth mentioning here that the FTB initiated in a context were dependency annotations were not as widespread as they are now. Moreover, due to copyright restrictions, the corpus is distributed under a specific license.[17] In contrast to the French Treebank, the Sequoia corpus was compiled from different sources: newspaper articles from *L'Est Républicain*, French Europarl samples, French Wikipédia articles, and medication leaflets issued by the European Medicines Agency. Sequoia is, therefore, more varied, and smaller than the FTB, since it totals a little over 70,500 tokens. Both corpora were later processed to accommodate the CONLL-U format and the Universal Dependencies syntactic annotation guidelines.

Figure 1 gives an overview of the main modules of our corpus-to-quiz processing chain.

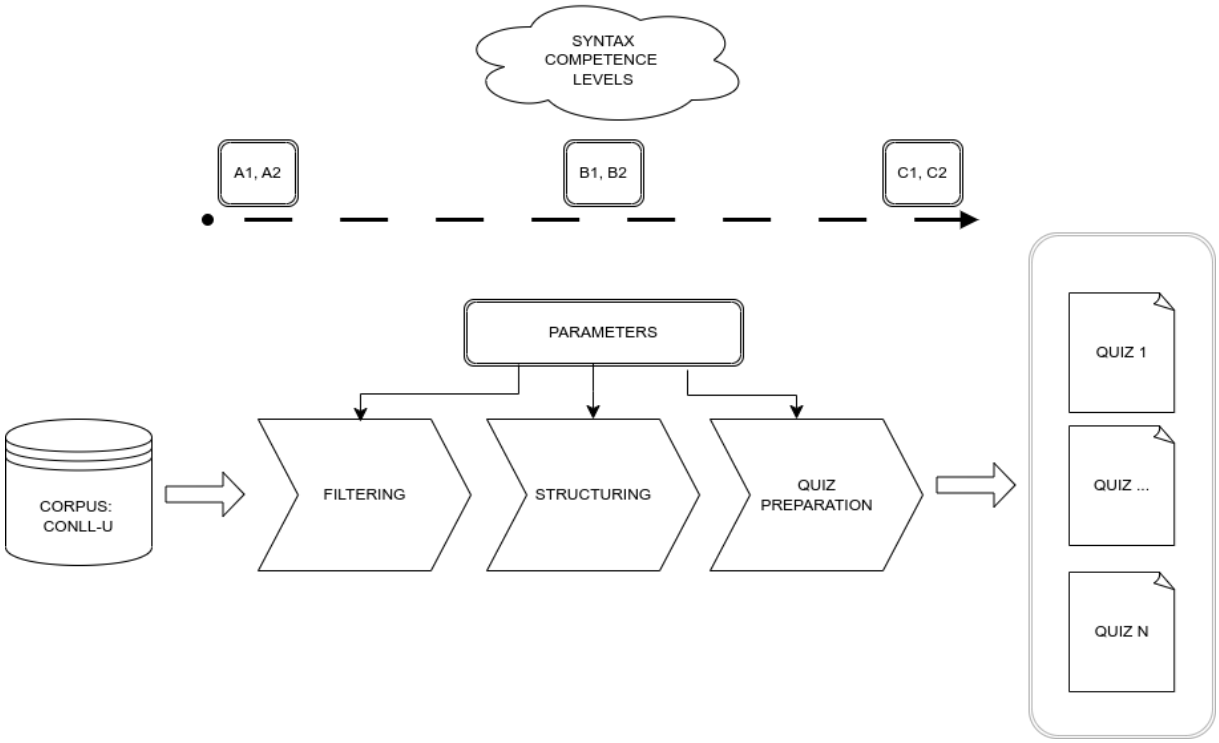


Figure 1: The Corpus-to-Quiz processing chain

As illustrated in Figure 1, the processing chain begins with the selection ('filtering' module) of sentences realising predefined types of syntactic structures from our reference corpora. These sentences are

[16] See Marcus et al. (1993).

[17] As a consequence, the FTB is not directly available from universaldependencies.org, but from the dedicated FTB website, under two distinct licenses (research/commercial).

chosen based on their CONLL-U formatted Universal Dependencies annotations. The underlying Python scripts are the powerhouse of this chain, meticulously processing each sentence to extract its syntactic elements, such as parts of speech, dependency relations, and functional roles. The extracted sentences are then processed to structure and output ready-to-use sets of questions ('structuring' and 'quiz preparation' modules), forming a consistent SCL progression step.

At each step of our processing chain, SCL form the overarching controlling element in that each target syntactic (and morphological) feature can be set at startup, with dedicated execution parameters (see section 4.2). Upon processing, the extracted sentences are then transformed into a structured quiz format, using the General Import Format Template (GIFT), a Moodle-specific text markup formatting. Figure 2 shows an example of an individual question that belongs to a quiz on parts-of-speech.

```
214 // question: 8245282  name: Parties du Discours
215 ::Parties du Discours::[markdown]Donner la partie du discours du mot **que** dans la phrase\:\nElle confirme **que**
    l'instauration d'un mécanisme de régulation de l'activité de la médecine libérale n'est pas seulement une Arlésienne.{
216         ~V Subj
217         ~V Part Passé
218         ~PRO Obj
219         ~PRO Int
220         ~V Inf
221         =Conj de Sub
222 }
```

Figure 2: A GIFT-formatted quiz question

As shown in Figure 2, GIFT allows for some level of output formatting (HTML or Markdown) and structuring: the individual question shown is associated with a particular section header ('::Parties du Discours::'), which will ensure that the question will be included in the desired subsection of a Moodle question bank upon importing the GIFT file.[18] Figure 3 shows how this GIFT-formatted question is rendered by Moodle, in the context of a formative or evaluative syntax quiz on parts-of-speech.

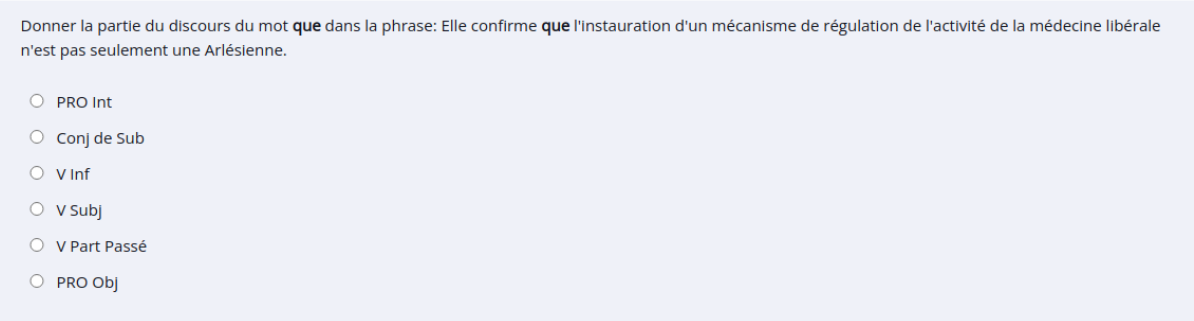


Figure 3: A GIFT-formatted question rendered in Moodle

As can be seen in Figure 3, the set of possible answers is parsed to render the question as a multiple choice quiz with only one correct answer (i.e. 'Conj de Sub' or subordinating conjunction).[19] In this particular case, the set of possible answers is randomly shuffled each time a student accesses the quiz.[20]

The versatility of the GIFT format allows for a wide range of question types — from True/False, to Multiple-Choice and Fill-In-the-Blank questions — making it ideal for testing different aspects of syntactic knowledge with a minimum of formatting and editing. As can be seen, GIFT is also a good trade-off between simplicity and control: it is far more readable than Moodle XML (or other heavily structured formats) while giving quiz managers and final users (i.e. colleagues importing preexisting quiz sets into their own Moodle courses) the necessary control to manage how the questions will be parsed and rendered by Moodle. GIFT-formatted quizzes can easily be imported into Moodle question

[18] Tags can also be associated with individual questions, thus providing a free-text indexing system (e.g. 'simple sentence', 'ditransitive verb', 'readibility:0.8', etc.) of crucial importance when dealing with sets of thousands of individual questions.

[19] An HTML form with radio buttons (only one selection allowed), instead of checkboxes (multiple selections allowed).

[20] Test parameters such as randomised shuffling, total allotted time, number of trials, minimum score to pass, completion status, etc. need to be set manually for each test.

banks, facilitating a streamlined transition from quiz design and generation to deployment in online and in-person classroom settings.[21]

4.2 Controlling the corpus-to-quiz output with SCL-informed parameters

As mentioned above, our scripts use execution parameters provided at startup to control every step of the quiz generation process. These parameters can be used to determine:

- the number of distractors to be included in the generated question
- the particular subset of part-of-speech tags, or of functional relationships
- the way part-of-speech tags and functional relationship labels will be presented to the user
- constraints on the token or on the lemmatized form
- constraints on specific morphological features (tense, mood, number, etc.).

As an illustration, these parameters can be combined to generate a set of individual questions, based on a list of relevant part-of-speech tags such as: prepositions, determiners, adverbs, pronouns (personal and relative) and (coordinating and subordinating) conjunctions, since distinguishing the different grammatical words is often considered a challenging task by A1-A2 SCL students. To make the quiz more engaging, we can not only set the quiz generation parameters to include some nouns, verbs, or adjectives, but also target specific words (or lemmas) belonging to those categories. In the context of syntactic C-R, such features are essential since native French speakers routinely confuse coordinating conjunction "ou" (or) and relative pronoun "où" (where/when), coordinating conjunction "et" (and) and verb "est" (is), or nouns bearing the suffix "-ment" with manner adverbs (e.g. noun "département" and adverb "doucement").[22] In the domain of syntactic functions, the same principles apply: questions can be generated focussed on specific functional relationships and specific parts of speech. For example, a formative set of quizzes has been set up as preparatory exercises for the end-of-semester final exam. These quizzes were targeted at infinitival subordinate clauses (e.g. "Travailler à l'étranger inquiète mes parents", working abroad worries my parents), as well as completive subordinate clauses (e.g. "Ils savent qu'un temps d'adaptation sera nécessaire", they know they will need some time to fit in) that realise either a subject or a direct object function.[23] In the case of these end-of-semester preparatory exercises, participating students had to complete 100 questions in 20 minutes. Each set of 100 questions was randomly sampled from over 2,000 individual questions so that students could extensively test their ability to apply the proposed parsing strategies to new examples taken from a consistent subset of syntactic phenomena. In addition, the quiz was set so that each individual question had a variable number of distractors (between 4 and 8), while a subset of questions targeted at different structures was added to maintain high levels of engagement and attention.[24]

At this point, it is worth underlining that native Universal Dependencies annotations do not fully align with the simplified view of parts-of-speech and functional relationships we use in introductory syntax courses. Therefore, a mapping table has been integrated into our quiz generation scripts: a tag such as "ADP" (adposition) was translated into "Préposition" (preposition), while the function label "nsubj" (nominal subject) was rendered as "Sujet" (subject). These cases are trivial one-to-one correspondences, but since UD annotation conventions follow the 'universal' annotation guidelines, meant for comparative syntax studies and not specifically for French syntax, in many cases the transposition requires

[21] Moodle quizzes can be printed on paper for classical pen-and-paper tests, while still retaining the possibility for an (almost) immediate feedback once response sheets have been scanned and processed (typical processing times: 15 minutes for over 200 response sheets, manual corrections included).

[22] "department" and "slowly". All examples provided are taken from actual syntax quiz questions we routinely use to identify students eligible for syntax tutoring classes.

[23] These cases are particularly challenging for students who have not yet embraced the systematic use of syntactic tests (e.g. pronominalization, passivisation and cleft sentences) and still resort to naïve strategies based on "who did what to whom" questions, in other words, semantic rather than syntactic parsing strategies typical of A1 SCL students.

[24] In our draft SCL progression, such a setting corresponds to A1-A2, which is the SCL required to pass introductory syntax courses, with C1-C2 as the target for completing a full L3 curriculum in linguistics.

some level of shallow parsing of the involved constituents. For example, in French (and many other languages) completive subordinate clauses can realise a verbal root's direct object, e.g.: in "ils savent **qu'un temps d'adaptation sera nécessaire**" the completive subordinate clause can be seen as a direct object of "savent",[25] However, in UD annotations, this constituent will be marked as "ccomp" (clausal complement) and not "obj". Therefore, our mapping table strategy is not just a simple equivalence list, as it must take into account structural constituents as well as functional ones.[26] This level of detail in tailoring syntax quizzes is crucial for promoting a deeper understanding of syntax among students, as it allows instructors to align the generated quizzes with a predefined learning scenario and SCL progression.

5 Discussion and perspectives

5.1 First results and classroom application

We first introduced automatically generated quizzes in our introductory syntax course in 2019. Since then, more than 800 first-year university students have been exposed to these self-correcting exercises. Initially designed as a complement to traditional in-person "chalk-and-talk" teaching methods, these quizzes were originally nothing more than a quick (and admittedly dirty) way to automatically generate more syntax quizzes than could be achieved by hand.[27] However, the unforeseen shift to online learning due to the COVID-19 pandemic propelled these quizzes to the forefront of our syntax education methodology. This transition has not only been a necessity, it has also been an opportunity to rigorously test and refine our approach in a real-world educational context. The COVID-19 pandemic, therefore, represented both a challenge and an opportunity to ramp up the initial concept's development, to arrive at the present situation where we are able to offer a range of curated formative and evaluative activities, providing students with a comprehensive tool for self-assessment and practice throughout the academic year. Figure 4 shows descriptive data on a whole promotion of first year students (n = 170)[28] enrolled in our syntax introductory course (2023-2024 academic year).

The scores for different quizzes are shown: 'POS Quiz' 1 and 2 are formative activities in preparation of mid-term or end-of-semester exams, targeted at parts-of-speech (rated for A1 to B1 SCL). 'TUTOR POS Quiz' is a special offline quiz we use to assess which students seem in need of personalised tutoring, 4 weeks into the semester. As can be seen, performance is rather poor for this particular test, but this can be explained by the fact that most first year students find PoS-tag identification and basic constituent identification tasks challenging at this stage, since these are not routinely taught in secondary school French grammar courses. Finally, 'PREP Func Quiz' is a formative activity on function identification, in preparation of end-of-semester exams (rated for B1 SCL). As is evident from the scores' distributions, a great deal of variation can be observed, which is to be expected if we take into consideration that the different POS quizzes are given at different moments throughout the semester[29] and, as the name suggests, POS Quiz 2 is more stringent than POS Quiz 1. The preparatory quiz on functions is much harder than any of the POS quizzes, since it targets functions such as direct and indirect objects, including cases where subordinate and infinitival clauses realise the targeted functions. In spite of seemingly high student attrition rates,[30] the data show relatively high rates of engagement, since no less than 90 students (53%) engage in any of the proposed quizzes. At this point, it is worth mentioning that those quizzes are not mandatory, even though a small bonus was attributed to the most assiduous students.[31] As a last comment, a clear distinction can be seen between high performers, i.e. students who embrace the proposed methodology, and those who still resort to naïve strategies and thus perform (much) lower.

The results of introducing these new learning activities have been encouraging, so far; students rou-

[25] From a French syntax viewpoint, such a constituent passes the personal pronoun test: "ils **le** savent", which is evocative of a direct object, a valid argument with regard to this verbal root's argument structure.

[26] The underlying strategy is to introduce UD's original annotations and syntactic representation logic at later stages of SCL acquisition.

[27] A first test was made using Perl scripts, which required considerable post-editing of the generated quizzes.

[28] The difference between the total number of enlisted students and the numbers in Figure 4 is due to student attrition.

[29] Every 3 to 4 weeks.

[30] Actually, these figures are 'normal': an average of 30% student attrition is commonly observed every year.

[31] Moodle analytics, based on completion statuses, were essential to assess engagement in this regard.

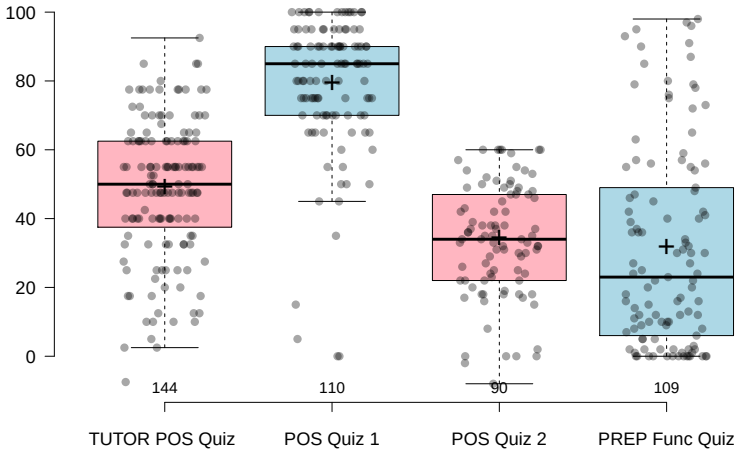


Figure 4: Syntax quizzes scores boxplot (sept. 2023-dec. 2023)

tinely engage with the proposed quizzes, finding them a reassuring tool for preparing for mid-term and end-of-term exams. One of the most important features, in the context of our syntax introductory course, is the instant feedback provided by such self-correcting learning activities. This feature is instrumental in improving student motivation and engagement, a notable improvement over traditional syntax exercises.[32] While there is still considerable room for improvement, especially in catering for less advanced students in their SCL progression, the immediate feedback mechanism not only helps students understand their mistakes in real time, but also allows them to track their progress, fostering a sense of achievement and encouraging continuous learning.

From the instructors' perspective, integrating these quizzes into the curriculum has been equally beneficial. Since our syntax quizzes can be easily imported into existing Moodle courses, they are now being used by other colleagues in charge of other syntax courses, as well as advanced students in charge of one-to-one tutoring classes. Moodle's learner analytics, although not as comprehensive as would be necessary for individual learner profiling, have nevertheless been invaluable in identifying the effectiveness of individual questions and overall quiz structures, allowing us to pinpoint "hard" and "easy" questions and adjust the quiz content accordingly, leading to a more targeted and effective teaching approach. The collected analytics have also provided insights into student learning patterns, aiding in the refinement of the Syntax Competence Levels and the overall curriculum. The quizzes have become more than just an assessment tool; they are now an integral part of our teaching methodology, paving the way for a truly dynamic, interactive, and personalised learning environment.

5.2 Perspectives

In this article, the emphasis has been on generating syntax quizzes from reference corpora. So far, Moodle does not support more interactive learning activities, such as drawing syntactic diagrams, or alternatives to multiple-choice questions (e.g. parsing student submitted hypotheses on the status of a given

[32] Students are receptive to the fact that the proposed quizzes are an experiment, which elicits feedback where, in previous years, none was to be had, positive or otherwise.

constituent).[33] In the context of our project, the emergence of the Learning Tools Interoperability (LTI) protocol[34] has opened new avenues for integrating third-party learning activities, surpassing the capabilities of previous standards such as SCORM (Sharable Content Object Reference Model), supported by well-known platforms such as H5P or Hot Potatoes.[35] LTI stands out as a more versatile and powerful tool, particularly in facilitating the incorporation of third-party web services into learning platforms such as Moodle. The LTI protocol will enable complex processing based on student data within a Moodle (or other LTI-compliant LMS) course. This capability is crucial for our project, since native Moodle activities are limited to variations of Multiple Choice quizzes. Furthermore, although Moodle provides learner activity and performance logging to some extent, no real learner profiling features are available. A fine-grained and comprehensive learner profiling is necessary for the next steps of our project: analysing learner responses and interactions is central to fine-tuning our SCL progression programme. From our standpoint, deploying LTI-compliant third-party smart web services is the way to go to provide a real adaptative, responsive, and personalised learning environment while making sure that the proposed content remains challenging and appropriate for each student's skill level.

Looking toward the future, our plans involve leveraging LTI to its fullest potential. By doing so, we aim not only to enhance the Moodle experience, but also to explore integration with other LMS platforms that support LTI. This broader integration aligns with our goal of making syntax education more accessible and effective across various educational contexts. Moreover, the use of LTI opens up possibilities for developing more sophisticated web services that can perform a range of functions, from learner profiling to complex syntactic analysis. These services can operate as standalone tools or in conjunction with Moodle, providing educators and learners with a versatile and powerful suite of tools for syntax education. In conclusion, the integration of our corpus-to-quiz processing chain into various platforms through the LTI protocol represents a significant step forward in our project. It not only enhances the capabilities of our system within Moodle, it also sets the stage for future expansions and innovations in the field of syntax education technology. By harnessing the power of LTI, we are poised to offer an educational experience that uses new educational tools, with a responsive and personalised learning environment in mind.

References

Abeillé, A., Clément, L., & Toussenel, F. (2003). Building a Treebank for French. *Treebanks: Building and using parsed corpora*, 165–187.

Aldabe, I., De Lacalle, M. L., Maritxalar, M., Martinez, E., & Uria, L. (2006). Arikiturri: an automatic question generator based on corpora and NLP techniques. *Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan, June 26-30, 2006. Proceedings 8*, 584–594.

Bick, E. (2001). The VISL system: research and applicative aspects of IT-based learning. *Proceedings of the 13th Nordic Conference of Computational Linguistics (NODALIDA 2001)*.

Bick, E. (2004). Grammar for fun: IT-based grammar learning with VISL. *Copenhagen studies in language*, *30*, 49.

Bick, E. (2015). DanProof: pedagogical spell and grammar checking for Danish. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 55–62.

Bitzer, D., Braunfeld, P., & Lichtenberger, W. (1961). PLATO: an automatic teaching device. *IRE Transactions on Education*, *4*(4), 157–161.

Borin, L., & Saxena, A. (2005). Grammar, Incorporated. In P. J. Henrichsen (Ed.), *CALL for the nordic languages* (pp. 125–145). Samfundslitteratur. https://samfundslitteratur.dk/bog/call-nordic-languages

[33] Keeping cheating in check is not the only concern. Students must be able to explain how and why they arrive at a given response.

[34] Released in its first version in 2008, by IMS Global, a not-for-profit learning consortium.

[35] Launched in 1995, the SCORM standard is developed by the Advanced Distributed Learning (ADL) consortium, with main contributors from the IMS (Instructional Management System) project.

Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, D., & de La Clergerie, É. V. (2014). Deep syntax annotation of the Sequoia French Treebank. *International Conference on Language Resources and Evaluation (LREC)*.

Colin, É. (2020). Traitement Automatique des Langues et génération automatique d'exercices de grammaire. *Theses, Université de Lorraine*.

Ellis, R. (2016). Grammar teaching as Consciousness Raising. *Teaching English grammar to speakers of other languages*, *2*(1), 128–150.

Heck, T., & Meurers, D. (2022). Parametrizable exercise generation from authentic texts: effectively targeting the language means on the curriculum. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 154–166.

Karlsson, F. (1990). Constraint Grammar as a framework for parsing running text. *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Karlsson, F., Voutilainen, A., Heikkilae, J., & Anttila, A. (2011). *Constraint Grammar: a language-independent system for parsing unrestricted text* (Vol. 4). Walter de Gruyter.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2008). The Sketch Engine. *Practical Lexicography: a reader*, 297–306.

Lee, J., & Seneff, S. (2007). Automatic generation of CLOZE items for prepositions. *Eighth Annual Conference of the International Speech Communication Association*.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational linguistics*, *19*(2), 313–330.

Mitkov, R., Le An, H., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural language engineering*, *12*(2), 177–194.

Perez-Beltrachini, L., Gardent, C., & Kruszewski, G. (2012). Generating grammar exercises. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 147–156.

Rutherford, W. E. (1987). The meaning of grammatical Consciousness-Raising. *World Englishes*, *6*(3), 209–216.

Schmidt, R. W. (1990). The role of consciousness in Second Language Learning1. *Applied linguistics*, *11*(2), 129–158.

Smith, S., Avinesh, P., & Kilgarriff, A. (2010). Gap-fill tests for language learners: corpus-driven item generation. *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, 1–6.

Uibo, H., & Bick, E. (2005). Treebank-based research and e-learning of Estonian syntax. *Proceedings of Second Baltic Conference on Human Language Technologies: Second Baltic Conference on Human Language Technologies*, 4–5.

Wijlff, A. (2006). VISL in Danish schools. *English Teaching: Practice & Critique (University of Waikato)*, *5*(1).

Standards Information System for CLARIN centres and beyond

Piotr Bański
Department of Digital Linguistics
IDS Mannheim, Germany
banski@ids-mannheim.de

Eliza Margaretha Illig
Department of Digital Linguistics
IDS Mannheim, Germany
margaretha@ids-mannheim.de

Abstract

The present contribution describes features of the CLARIN Standards Information System that have been designed to assist data deposition centres in CLARIN. We also show what is needed and what has been done in order to go beyond the originally designated target, so as to provide service to sibling and descendant research infrastructures, of which DARIAH and Text+ are taken as examples. This paper is aimed primarily at representatives of research infrastructure nodes, responsible for preparing and sharing data deposition information about their centres or repositories. It assumes a degree of technical knowledge or experience in using the XML format and tools, the REST API, and version control systems.

1 Introduction

Many modern research infrastructures (RIs) offer data deposition services for their users. For CLARIN B-centres, the provision of this service is a default characteristic that is subject to certification requirements and that is used as a basis of a measurement needed to calculate one of the CLARIN-ERIC Key Performance Indicators.

The range of data that constitutes language resources or accompanies them is very wide, from the prototypical electronic corpora and dictionaries through, among others, participant lists, tagsets, digital facsimiles, raw audiovisual datasets, language models of various complexity and size, and ending with datasets produced by behavioural or neurolinguistic experiments, as well as documentation of various kinds. Neither the kinds of data nor the formats used to encode it are exclusive to CLARIN. CLARIN's focus has historically overlapped with some areas served by DARIAH and, by a natural extension, with CLARIAH networks that combined DARIAH and CLARIN nodes in some of the European countries, at various points in time. In Germany, the national CLARIN-D merged with DARIAH-DE into CLARIAH-DE in 2019, and, since 2022, many former German DARIAH and all the former CLARIN-D centres (as well as some centres previously not belonging to either of the two) have formed the Text+ consortium, which is part of the German National Research Data Infrastructure, NFDI.[1]

This is illustrated in Figure 1, which does not take historical developments into account, but is rather meant to hint at the resulting inter-RI relationships. The reader should bear in mind that, while CLARIN and DARIAH are multinational networks, Text+ is restricted to Germany.

This paper showcases the Standards Information System (SIS) in the context of an extended network of inter-RI relationships. The main purpose of the SIS since around the year 2021 has been to serve as a platform for sharing and collecting information about data deposition formats supported by CLARIN centres, in lieu of centre-specific recommendations, provided individually in the form of lists or tables, differing in structure and granularity. The information is crucial to end-users who wish to deposit their data for the purpose of archiving or reuse, but it also provides an important insight into the network as a whole: the aggregated recommendations indicate trends in the usage of the particular data formats: a format may be labelled as "recommended", "acceptable", or "discouraged", and – on the safe assumption

[1] References to the home pages of the research infrastructures mentioned in this paper are gathered at the end.

Piotr Bański and Eliza Margaretha Illig 2023. Standards Information System for CLARIN centres and beyond. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 13–22. DOI: https://doi.org/10.3384/ecp210

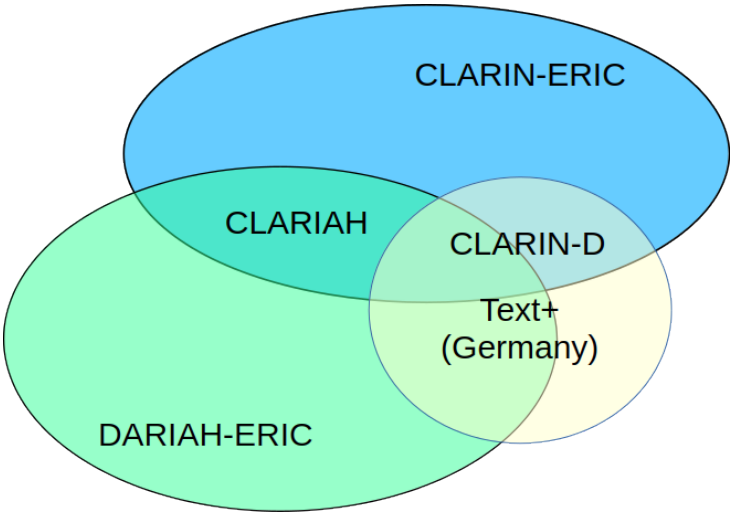


Figure 1: Relationships between selected research infrastructures representing broadly understood Human Language Technology as well as Arts and Humanities.

that CLARIN centres follow technological innovations – the changing recommendations may constitute an important input to standardisation initiatives as well as to the industry at large.

Apart from its informative goal, the SIS attempts to assist the user in thinking about their data: about the ways in which the various parts of the user's deposited data function and also about the formats that data may come in. Sections of the SIS, available from the side menu on its homepage, provide, among others, information on the functions of data in the Human Language Technology (HLT) setting as well as information on the data formats and the relationships both among formats and between formats and the standards that the formats are typically tied to.[2]

The following simplified definitions are assumed in this paper: a **standard** is a document that, following a systematic process of community consultations and revisions carried out by a standards setting organisation, sums up and recommends the best practices for dealing with certain tasks; a **(technical) specification** is similar to a standard, but its origin is less procedural and more community-oriented; a specification often enjoys the status of a *de facto* standard, before it becomes institutionally codified and disseminated. A **data format** is a serialisation of a data model defined by a standard or a specification. Note that this is a very broad statement that denotes, for example, both the XML format as a serialisation of the well-known XML standard defined by W3C, and a very narrowly defined application of XML such as a particular corpus-encoding format compliant with the ISO MAF (Morphosyntactic Annotation Format), heavily restricted by additional data models superimposed on the general XML data model. The end-user rarely has the expertise to distinguish between such cases, and it is part of the task of the SIS to suggest that, among others, "XML" alone is relatively meaningless in the context of data formats, and that it should be further qualified in order to ensure that the user's data is sustainable and interoperable – which are the usual aims of data deposition.

The paper is organized as follows. The history of the development of the SIS is briefly recounted in Section 2. The current features of the SIS and the extended features that support other RIs are elaborated on in Sections 3 and 4, respectively. Related work is presented in Section 5. Section 6 provides a summary of the main points made in the paper and indicates the paths for further development.

[2]See Figure 4 for an illustration of both the side menu and a part of a format information page.

2 Standards Information System: basic information

2.1 From CLARIN Standards Guidance to CLARIN SIS

The current CLARIN Standards Information System[3] extends the former CLARIN Standards Guidance (Stührenberg et al., 2012), contributed to the CLARIN infrastructure by CLARIN-D. Originally, the system provided information about various HLT standards and indicated relationships among them. The practical aim of the Standards Guidance was to assist users in finding standards most appropriate for their purposes. In some cases, names or abbreviations of CLARIN centres claiming to use those standards were provided, so that the user knew which centre to choose for the purpose of depositing data encoded in some specific formats (see Figure 2 for a simplified data model of the original system, indicated by the grey background). A side goal was to provide a taxonomy or even a small knowledge base of standards and technical specifications, served by eXist-DB (Siegel & Retter, 2014).[4]

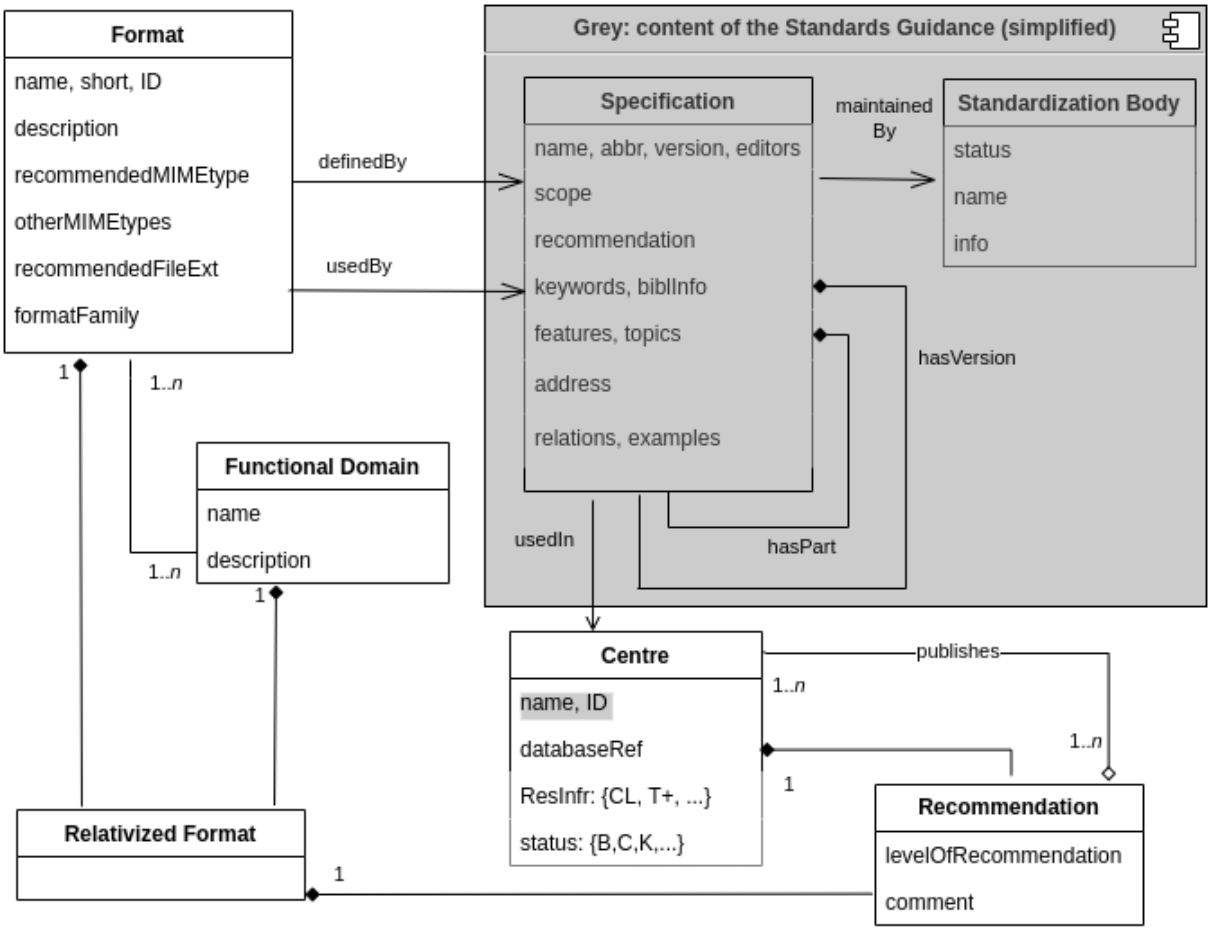


Figure 2: Simplified original data model (on grey background) with additions designed to incorporate format recommendations and research infrastructures other than CLARIN. Filled diamond arrows represent strong aggregation, hollow diamond arrows – weak aggregation, while simple relationships are represented by lines, with simple arrowheads pointing to the objects of the relationships.

Around the year 2019, the CLARIN Standards Committee undertook the task of providing a platform for CLARIN centres to share their recommendations concerning the formats for data that could be de-

[3]The SIS can be accessed at https://standards.clarin.eu/sis/, which is an alias for https://clarin.ids-mannheim.de/standards/. Its GitHub home is at https://github.com/clarin-eric/standards and the documentation is in the project wiki at https://github.com/clarin-eric/standards/wiki. The SIS is listed as a knowledge base at Fairsharing.org: https://fairsharing.org/4705.

[4]The content of the Standards Guidance is still available in the current SIS, in the "Standards and Specifications" menu item on the left. It is not actively maintained.

posited at each of them. That task eventually resulted in adapting and extending the Standards Guidance to become the tool to aggregate the information provided by centres and to visualise it in various ways. Figure 2, improving upon Bański and Hedeland, 2022, shows the extensions to the original data model needed to embrace the new functionality. The Standards Committee faced several challenges, among others concerning the classification of formats, the initial retrieval of lists of data deposition formats for some centres, or devising the least troublesome ways for the particular centres to submit their recommendations to the system. Due to the diversity of ways in which data formats are used in HLT research, it is also a challenge to visualise the recommendations and maintain them with minimal effort. The deliverable has evolved since 2021 from a complex set of Google spreadsheets that put together formats, format categories, and CLARIN centres, to the current XML format integrated in the SIS.[5]

2.2 The SIS data model

In Figure 2, the greyed area of the data model represents, with minor simplifications, the state of the CLARIN Standards Guidance up to 2021. In some cases, information about standards mentioned centre names, hence the grey background in the Centre class. In the SIS, centres are represented consistently, with an indication of the research infrastructure that they belong to (see Section 4) and of their status within that infrastructure. For CLARIN, this means their status as B- or C-centres.[6] CLARIN centres reference the CLARIN centre registry [7] as the authoritative source of information.

Format	Centre	Domain	Recommendation
PDF/A	EKUT	Image Source Language Data	recommended
PDF/A	EKUT	Textual Source Language Data	recommended
PDF/A	FIN-CLARIN	Documentation	recommended
PDF/A	FIN-CLARIN	Textual Source Language Data	acceptable
PDF/A	IDS	Documentation	recommended
PDF/A	IDS	Image Source Language Data	recommended
PDF/A	DANS	Documentation	recommended
PDF/A	DANS	Other	acceptable
PDF/A	DANS	Textual Source Language Data	recommended
PDF/A	MI	Image Source Language Data	recommended
PDF/A	MI	Textual Source Language Data	recommended
PDF/A	ZIM	Image Source Language Data	recommended
PDF/A	ZIM	Textual Source Language Data	recommended
PDF/A	LAC	Contextual Data	recommended

Figure 3: Fragment of format recommendations by CLARIN centres concerning the PDF/A format. Centres may comment on their recommendations (the circled *i* shows the comment in a pop-up).

A crucial element of the SIS is the set of functional data domains that serve to fine-tune the purposes for which the individual data items are collected: for example, data encoded in the PDF/A format are perfect for the purpose of documentation, but definitely not ideal for the purpose of providing annotation for audiovisual sources, or collections of statistical data. This is illustrated in Figure 3, which is a screenshot of a fragment of aggregated format recommendations.

While functional domains are hard-coded in the system, the instantiation of the Format class in the data model of Figure 2 may take two forms. In most cases, there already exists a format description – a

[5] Much of the history behind the task described here is documented at https://www.clarin.eu/content/standards.

[6] K-centres are typically outside the scope of the SIS, unless they are paired with a centre that offers data depositions.

[7] The CLARIN centre registry is available at https://centres.clarin.eu/ and a list of certified B-centres can be found at https://www.clarin.eu/content/certified-b-centres. Note that data deposition services are sometimes offered by centres other than "B", and that B-centres may temporarily become C-centres pending re-certification.

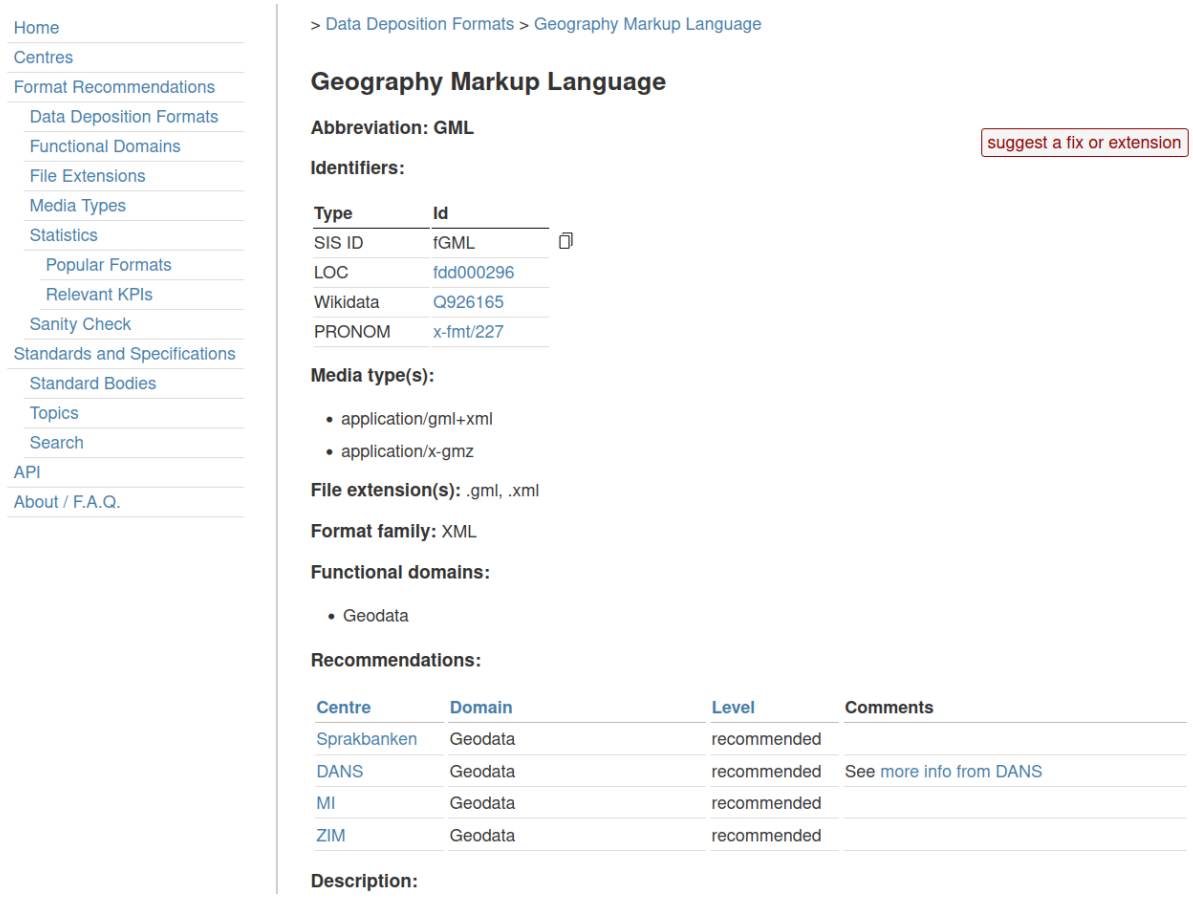


Figure 4: Format description information on GML, with cross-references to the Library of Congress, PRONOM and Wikidata at the top, and other details derived from the system. The description part is suppressed. On the left is the side menu that offers various visualisations of the underlying data.

document that describes the format and is linked from the list of recommendations (see Figure 4 for an example screenshot). In Figure 3, the fragment "PDF/A" is a link, and clicking on it displays the basic information about the PDF/A format, as well as links to related formats. It is also possible that a format does not yet have a corresponding description document in the SIS as is the case of formats mentioned in the recommendations listed in Figure 5. In such cases, the format name is not a link. Instead, it is followed by the ⊕ character and clicking on that symbol opens a pre-configured GitHub issue where the basic information on the given format can be provided, so that a physical format information document can be created on that basis. This is a way to ensure that the inventory of formats handled by the SIS can be extended according to the new or modified recommendations formulated by centres, and that the recommendations are not limited to the existing format descriptions.

A recommendation is a qualified pairing of a centre with what the model calls a "relativised format", i.e., a data format viewed from the perspective of the function that the data in that format is expected to fulfil: in the example of the PDF/A format adduced above, the domain for which this format is universally recommended is "Documentation", followed closely by "Textual Source Language Data" – although in the latter case, Figure 3 shows that not all centres are uniform in advocating that format as ideal for "Written unstructured/plain text or originally structured text (e.g. HTML) without linguistic or other mark-up added for research purposes", which is how the SIS defines the "Textual Source Language Data" do-

Figure 5: Fragment of recommendations that do not point to an existing format description document. Clicking on the ⊕ character (note the pop-up) opens a pre-configured GitHub issue.

main[8]. A complete SIS recommendation qualifies a relativised format with a degree of recommendation that the given centre determines, by choosing one of the three recommendation labels: "recommended", "acceptable", and "discouraged". An example of XML encoding of a relativised format is shown in Figure 6, where the submission of data in the domain of "Audiovisual Source Language Data" in the format identified by "fMP3" is discouraged. Additionally, the centre (in the case, IDS Mannheim) provides a comment on the reason for the negative recommendation.

```
<format id="fMP3">
  <domain>Audiovisual Source Language Data</domain>
  <level>discouraged</level>
  <comment>lossy formats should be avoided if possible</comment>
</format>
```

Figure 6: Instantiation of a relativised format with a comment (part of a centre's list of recommendations)

There is a many-to-many association between formats and functional domains: data encoded in a specific format can usually take on many functions, and conversely: a single functional domain is served by many formats. This relationship is never encoded directly – it is derived from recommendations provided by centres. If no centre were to submit a recommendation similar in structure to that in Figure 6, no association between the MP3 format and the "Audiovisual Source Language Data" domain would be derived in the SIS.

3 Standards Information System: data submission and exploitation

The current offer of the SIS towards centres can be summed up in the following three points:

1. increasingly user-oriented way of submitting information,
2. increasingly attractive way to benefit from data aggregation,
3. a way to reuse the data submitted by the centres.

Below, we elaborate on each of these points.

3.1 Data submission

For the purpose of the first data submission, all that is expected from a given centre is a single document that contains a list of formats provided together with a statement that expresses the centre's willingness to accept the particular format in some functional domain. Recall that the possible levels of recommendation are "recommended", "acceptable", and "discouraged", where the last one indicates that the centre may either have insufficient capacity to prepare such data for long-time preservation, or that the process may take a long time. Conversely, the value "recommended" indicates a promise that the deposition process

[8]The list of domains supported by the system is accessible at https://clarin.ids-mannheim.de/standards/views/list-domains.xq and the list of supported formats is at https://clarin.ids-mannheim.de/standards/views/list-formats.xq.

should be relatively painless to both parties. Submitting small-sized initial lists meant to be iteratively extended with further domains or further recommendations is also an option. The SIS offers templates that can be used for that purpose.

Note that, for many centres, the members of the Standards Committee have entered the initial recommendations on the basis of documents published by those centres. That step required a lot of interpretation on the part of the submitter, in order to adjust various kinds of the original recommendations or their varying granularity to the format used by the SIS. Such recommendations are considered "seeds" and should be reviewed, and – probably in many cases – corrected and extended by the given centre. The users are warned in such cases that they are looking at recommendations that have not been curated yet. That warning is eliminated after the centre submits curated information and appoints a contact person.

The preferred way for data submission is by pull requests (PRs) directed at the SIS source code deposited on GitHub. CLARIN developers are familiar with GitHub, so submitting a PR presents no obstacle. For technically less advanced users, the SIS offers an alternative way through editing the recommendation documents, which may be exported from the section of the SIS devoted to the given centre (even if the set of recommendations is empty). These exported files contain placeholders and templates, added in order to make the data input easier. They are additionally constrained by document grammars (W3C XML Schema and ISO Schematron), which signal errors and provide closed lists of options to choose from, where feasible. Finally, many places in the SIS offer an option to switch to editing a templated GitHub feature request, in a single click. This final way is naturally best used only for minor fixes. The wiki system that accompanies the SIS source, linked from the SIS instance, provides additional instructions and illustrations.

3.2 Data aggregation and visualisation

Aggregating structured data from several sources presents an opportunity to visualise the data in various ways and to provide statistics. For this purpose, the SIS offers, among others, word-clouds based on the format keywords, tabular displays of various sorts, extracted lists of file extensions and media types for use in processing pipelines, as well as higher-level statistics concerning, for example, the most "popular" file formats relative to the intended function of the submitted data. For CLARIN, the data aggregated in the SIS make it possible to dynamically compute the Key Performance Indicator (KPI) "collections of standards and mappings", measured by calculating the percentage of centres offering data deposition services and having published their format recommendations (see de Jong et al. (2020) and Bański and Hedeland (2022)) for discussion and further references).

3.3 Information recycling: the SIS API

Finally, the SIS offers a way for the centres to reuse the data that was submitted, via a REST API. This way, the SIS may be used as the sole tool for the maintenance of centre recommendations (and, in the case of CLARIN, to satisfy one of the B-centre certification requirements; see Bański and Hedeland, 2022 for discussion). There is no need to manage two separate instances of data: one for the SIS, and one for the centre itself to display. The API offers a way to receive the information that the centre has provided, to be transformed and styled in the way that the centre wishes.[9]

Additionally, other potentially useful information, e.g. format descriptions, can be obtained via the API and reused. Information obtained from the SIS is available under the CC0 "No Rights Reserved" waiver, with a non-binding request for the SIS to be recognised as the source.[10]

4 Extending the SIS beyond CLARIN

The SIS is in the process of constant development and receives upgrades of functionality on a nearly weekly basis. The most recent work has been influenced by meetings with the Text+ Standardisation

[9] See the example result of an API query for the data of IDS Mannheim at https://clarin.ids-mannheim.de/standards/rest/data/recommendations/IDS-recommendation.xml. The API also supports searching and exporting recommendations with some filtering criteria, such as centre, domain and recommendation level.

[10] See https://creativecommons.org/public-domain/cc0/ for the explanation of how CC0 works.

Group of the Collections cluster, and resulted in partial internationalisation of the underlying functionality: it is now possible to use language tags for centre descriptions and comments on recommendations, and to retrieve that information via the SIS API.

As for the needs of the sibling infrastructure DARIAH, including the cases where the national CLARIN and DARIAH nodes operate as CLARIAH, the SIS offers a functional domain inventory that goes beyond pure language-oriented applications[11]. Depending on the decision by the DARIAH governance (or by the individual repositories) to use the SIS, it remains to be seen whether the repertoire currently offered is going to require further adjustments and fine-tuning given the needs of DARIAH centres.

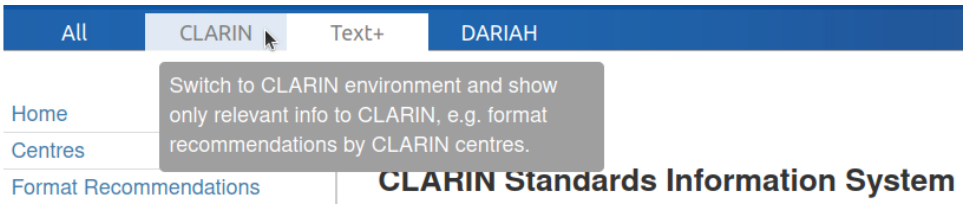


Figure 7: Switching between research infrastructures. The currently active RI "Text+" is shown on the white background. Hovering over one of the tabs displays a tool-tip to guide the user.

The SIS also provides a functionality that enables users to easily switch between RI environments and to filter the content according to the selected RI. Figure 7 shows how the switch works. "Text+" on the white background indicates that the Text+ environment is currently active. In this case, only Text+ centres and their format recommendations are listed, while the information concerning other centres is hidden. Moreover, language preferences are also taken into account in RI environments. For Text+, which prefers the German language, descriptions and comments are shown in German, as long as centres have provided them. Otherwise, the system falls back to English.

Extending the SIS beyond CLARIN opens new challenges and exposes some limitations of the system. First, some CLARIN centres may appear under different names in research infrastructures other than CLARIN. Currently, the system only allows a single name for a single centre. Whether this is acceptable or whether the centre list needs to be split depending on the RI remains to be seen.

Second, since format recommendations are grouped by the given centre, they are considered to be the same for the same centre across the RIs. When a single centre is a node in multiple RI networks, the SIS assumes that its format recommendations are the same in all these RIs. That means that it would not be possible, for example for the IDS, to recommend the CHAT format in CLARIN but discourage it in Text+. Whether this restriction is going to be problematic remains to be seen when more centres have provided their data.

5 Related work

Similarly to the SIS, re3data.org (Pampel et al., 2013)[12] standing for Registry of Research Data Repositories, provides information about global repositories for deposition of, and access to, research data across various academic fields, and assists researchers in finding a repository suitable to their data and its requirements. The content types of the research data in re3data.org, e.g. audiovisual data and raw data, are more general than those offered by the SIS functional domains, but nevertheless comparable. The SIS is more specific than re3data.org and, naturally, more oriented towards broadly defined HLT research centres – for example, it offers comprehensive details regarding the acceptability of data formats by the listed repositories.

PRONOM (The National Archives, 2002), the Digital Formats website (Library of Congress, 2023) and Wikidata (Wikimedia Foundation and contributors, 2023) present detailed information about file formats including relations to other formats and tools to support the long-term accessibility and preservation

[11] See https://clarin.ids-mannheim.de/standards/views/list-domains.xq
[12] https://www.re3data.org/

of digital materials. Information provided by these initiatives complements the basic information on the particular formats provided by the SIS, and most of the format information documents in the SIS provide cross-references to these three sources (see Figure 4).

In order to promote collaborative knowledge gathering, re3data.org, PRONOM, and Wikidata allow users to submit information through online forms. The SIS targets a much more restricted audience and uses the means made available by the GitHub environment, from pull requests to unstructured issue reports (see Section 3.1), depending on the user's choice and level of technological proficiency.

Similarly to re3data.org and Wikidata, the SIS offers a REST API, as mentioned in Section 3.3. The SIS API is geared more towards the retrieval of entire sets of recommendations, for reuse by centres.

6 Summary and outlook

The Standards Information System is a dynamic platform that adjusts to the expanding demands of data deposition centres. It used to be a relatively static information booth, which around the year 2020 began to evolve into a partially interactive system. The year 2023 is another road marker on its path, as the system opens towards research infrastructures other than CLARIN-ERIC, in the hope to become a platform for the aggregation, visualisation, and measurement of data deposited in research initiatives oriented towards the Humanities. In March 2024, the CLARIN Technical Centres Committee decided to encourage centres to submit recommendations to the SIS and actively monitor the overall progress.

At the time of writing, there are 36 CLARIN depositing centres recorded in the SIS. For 22 of them, at least rudimentary format recommendations have been recorded, and one is in the process of adding the data, after which the dynamically calculated KPI "percentage of centres offering repository services that have published an overview of formats that can be processed in their repository" should be at 63.8%.[13]

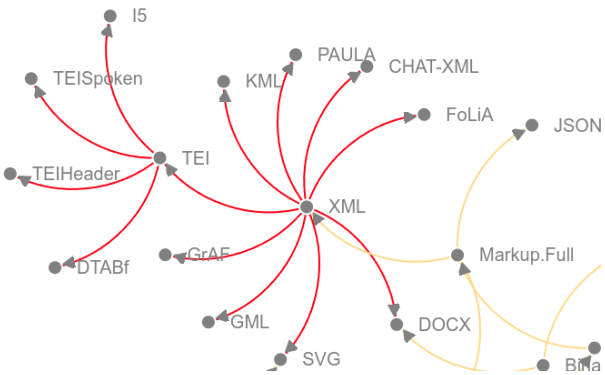


Figure 8: Formal relationships between formats encoded in "format families", a pilot project in the SIS. The figure is a snapshot of a much wider graph. All the leaves and intermediate nodes labelled with format names are clickable and open format description pages similar to that in Figure 4.

Much of the functionality that is needed to serve more than one research infrastructure is already in place. The list of functional domains is already able to accommodate demands that go beyond strictly language-resource-oriented use cases, and can be adjusted to other data functions (with the category "Other" serving as interim storage space). The list of formats is open-ended by design and can be extended both via pull requests and GitHub issues. A preliminary study of formal inter-format relatedness is at the beta stage (see Figure 8) and provides an alternative way to navigate across formats. The system is ready to be used both for CLARIN centres and beyond CLARIN.

Acknowledgments

The SIS has been developed in the context of the work done by the CLARIN Standards and Interoperability Committee (formerly, the CLARIN Standards Committee) and owes much to its former and present

[13] Note that B-centres can sometimes temporarily become C-centres during re-certification. That does not change their classification as data deposition centres, and that is what the KPI calculates.

members, as is only partially evidenced in the CLARIN Bazaar presentations offered in the previous years – a lot of ideas have been discussed, criticised and advanced during the (mostly virtual) committee meetings. We would like to acknowledge the three anonymous CLARIN conference reviewers and thank them for kind words and critical remarks. We are also grateful to the reviewers for the proceedings volume – thanks to their helpful criticism, the text has become much more readable.

Consortia and infrastructures mentioned in the paper

- CLARIAH-DE: https://www.clariah.de/en/
- CLARIN: https://www.clarin.eu/
- CLARIN-D: https://clarin-d.net/en/
- DARIAH: https://www.dariah.eu/
- DARIAH-DE: https://de.dariah.eu/
- NFDI: https://www.nfdi.de/
- Text+: https://text-plus.org/en/

References

Bański, P., & Hedeland, H. (2022). Standards in CLARIN. In D. Fišer & A. Witt (Eds.), *CLARIN: The Infrastructure for Language Resources* (pp. 307–340). De Gruyter. https://doi.org/doi:10.1515/9783110767377-012

de Jong, F., Maegaard, B., Fišer, D., van Uytvanck, D., & Witt, A. (2020, May). Interoperability in an infrastructure enabling multidisciplinary research: The case of CLARIN. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 3406–3413). European Language Resources Association. https://aclanthology.org/2020.lrec-1.417

Library of Congress. (2023). Sustainability of Digital Formats. Retrieved March 28, 2024, from https://www.loc.gov/preservation/digital/formats/index.html

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.-J., Gundlach, J., Schirmbacher, P., & Dierolf, U. (2013). Making Research Data Repositories Visible: The re3data.org Registry. *PLOS ONE*, *8*(11), 1–10. https://doi.org/10.1371/journal.pone.0078080

Siegel, E., & Retter, A. (2014). *eXist*. O'Reilly Media, Inc.

Stührenberg, M., Werthmann, A., & Witt, A. (2012). Guidance through the standards jungle for linguistic resources. In *Proceedings of the LREC 2012 workshop on collaborative resource development and delivery* (pp. 9–13).

The National Archives. (2002). *The technical registry PRONOM*. Retrieved March 28, 2024, from https://www.nationalarchives.gov.uk/PRONOM/Default.aspx

Wikimedia Foundation and contributors. (2023). Wikidata. Retrieved March 28, 2024, from https://www.wikidata.org/wiki/Wikidata:Main_Page

XSL-HoReCo and GoSt-ParC-Sign: Two New Signed Language - Written Language Parallel Corpora

Mirella De Sisto
Tilburg University, the Netherlands
M.DeSisto@tilburguniversity.edu

Vincent Vandeghinste
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
and KU Leuven, Belgium
vincent.vandeghinste@ivdnt.org

Caro Brosens
Vlaamse GebarentaalCentrum
Antwerp, Belgium
caro.brosens@vgtc.be

Myriam Vermeerbergen
KU Leuven, Belgium
myriam.vermeerbergen@kuleuven.be

Dimitar Shterionov
Tilburg University, the Netherlands
D.Shterionov@tilburguniversity.edu

Abstract

Developments in language technology targeting signed languages are lagging behind in comparison to the advances related to what is available for so-called spoken languages.[1] This is partly due to the scarcity of good quality signed language data, including good quality parallel corpora of signed and spoken languages. This paper introduces two parallel corpora which aim at reducing the gap between signed and spoken-only language technology: The XSL Hotel Review Corpus (XSL-HoReCo) and the Gold Standard Parallel Corpus of Signed and Spoken Language (GoSt-ParC-Sign). Both corpora are available through the CLARIN infrastructure.

1 Introduction

In Europe about half a million people have a sign language as their main or preferred means of communication (Pasikowska-Schnass, 2018). Nevertheless, when talking about language technology, sign language technology is extremely lagging behind in comparison to the tools available for spoken languages (Vandeghinste et al., 2023). One of the reasons is the scarcity of data.[2] This is partially due to the fact that sign languages do not have a widely-used written form, hence collecting written sign language data is not an option (in contrast to what is the case for many spoken languages).

Data collection and data storage also face a number of challenges, such as GDPR[3] restrictions, difficulties in recruiting participants, etc. A lot of short videos are scattered around different platforms and websites, which makes it difficult and time consuming to track them down and get the informed consents of the signers (Vandeghinste et al., forthcoming).

The majority of sign language data comes in the form of videos. To date there is no automatic tool able to annotate or translate sign language videos (Morgan et al., 2022; Vandeghinste et al., 2023), which means that these processes rely on very time-consuming manual work; consequently, the amount of available annotations or translations is scarce.

In addition to that, the quality of the data which are available is often rather problematic (Vandeghinste et al., forthcoming). Most of the sign language datasets readable for machine learning consist of television broadcasts with a spoken language as a source, such as Camgöz et al. (2021) and Koller et al. (2015) which is then interpreted into a sign language by a hearing interpreter. In those cases sign language is the target language of interpreting, which often occurs simultaneously and in real time, and might be influenced by the source language as well as affected by the interpreting process. Most hearing interpreters

[1] We characterize *spoken* language as contrasting with *signed* language, rather than in relation to speech data.

[2] For a detailed overview of data-related challenges, see De Sisto et al. (2022) and Vandeghinste et al. (forthcoming)

[3] The General Data Protection Regulation is a component of the EU privacy law and human rights law.

Mirella De Sisto, Vincent Vandeghinste, Caro Brosens, Myriam Vermeerbergen and Dimitar Shterionov 2023. XSL-HoReCo and GoSt-ParC-Sign: Two New Signed Language - Written Language Parallel Corpora. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 23–33. DOI: https://doi.org/10.3384/ecp210

do not use a sign language as their main or preferred means of communication (the exception being interpreters who are CODA's – Children of Deaf Adults – and some other specific cases); consequently, they are considered L2 signers.

Additionally, the length of e.g. news broadcasts and the range of specific topics with associated specific lexicon, as well as the speed at which information is disseminated, and the number of names that need to be fingerspelled, all heavily factor into the quality of the result. Interpreters are usually required to take a break every 15-20 minutes when they interpret simultaneously, to keep the quality up and avoid cognitive overload, while news broadcasts are often longer and the pace of information is very high. Different from face to face situations, the interpreter cannot ask the newsreader to repeat themselves or go slower, so in order to keep up with the pace, the interpreters might lean more towards the source text than is ideal.

Within the two projects we present in this paper, we take this into account. Along with the open distribution of these data sets (making them available for the wider research community), the quality of the data (professional translations, involvement of native signers for translation and validation, etc.), and the different (identifiable) domains, they have been collected in a way that suits their use in Machine Learning (ML) applications, and thus have the potential to stimulate the advancements in the field of signed language technology through both high-quality data for training models as well as a gold standard for testing.

After presenting related work in Section 2, we present two recent projects that each aim to address the lack of good quality data by providing parallel data of signed and spoken language data.

- the **XSL Hotel Review Corpus** (XSL-HoReCo) consists of a parallel dataset of hotel reviews in written English (the source language), videos in Sign Language of the Netherlands (Nederlandse Gebarentaal, NGT), Flemish Sign Language (Vlaamse Gebarentaal, VGT), Spanish Sign Language (Lengua de Signos Española, LSE), written Dutch, Spanish and Irish. This dataset is described in Section 3.
- the **Gold Standard Parallel Corpus of Signed and Spoken Language** (GoSt-ParC-Sign), a gold standard dataset of semi-spontaneous Flemish Sign Language (Vlaamse Gebarentaal) (VGT) videos translated into written Dutch. This dataset is described in Section 4.

Section 5 draws conclusions.

2 Related work

Although sign languages are low resource languages, there have been some data collection efforts in the past. Kopf et al. (2021) contains a comprehensive list of available corpora for sign languages, but is limited to those cases where sign language is the source. The associated Sign Language Compendium (Kopf et al., 2022)[4] requires as a criterion for inclusion that a corpus must contain (semi-)spontaneous signing, provide transcriptions or translations for at least some of its content and contain at least 10 hours of sign language recordings.

Various sign language datasets have been collected over the years, e.g. CorpusNGT (Crasborn et al., 2008) or DGSKorpus (Prillwitz et al., 2008). However, such datasets are not particularly suited for machine learning or deep learning applications, and require substantial processing prior to building language technology for signed languages (De Sisto et al., 2022; Vandeghinste et al., forthcoming).

For the signed languages addressed in XSL-HoReCo and GoSt-ParC-Sign the following data are available. For NGT existing datasets with authentic signers are the Corpus NGT (Crasborn et al., 2020) and part of the ECHO corpus (Nonhebel et al., 2004). For VGT this is limited to the Corpus VGT (Van Herreweghe et al., 2013). For LSE there is the Corpus de la Lengua de Signos Española (CORLSE)[5] and the small corpus iSignos.[6]

As already mentioned, some sign language datasets that are regularly used for sign language recognition or translation contain *non-authentic* sign language. In these cases we cannot assume that the signers

[4]https://www.sign-lang.uni-hamburg.de/lr/compendium/index.html
[5]https://corpuslse.es/
[6]https://http://isignos.uvigo.es/

belong to the respective sign language community of the language they sign, as most often they are *hearing* sign language interpreters. There is still debate in the sign language technology research community whether the price for using lower quality data can be compensated by the amount of such data, which is much more abundantly available.

Such data for VGT is available in the Content4All corpus (Camgöz et al., 2021). To alleviate thism, in the BeCoS data (Vandeghinste et al., 2022) the interpreters are deaf signers re-interpreting hearing signers (which are not on the video), so the resulting sign language, although still being the target language, can be consdered authentic. More data has been collected, such as more television broadcasts with sign language interpretation in VGT and videos of the plenary sessions of the Belgian Federal Parliament, with live interpretation into VGT and French Belgian Sign Language (Langue des signes de Belgique francophone; LSFB), but has not yet been processed nor released, partly due to legal constraints (for the broadcasts).

The availability of LSE data is more scattered and not easily gathered. For instance, the corpus created by Porta (2014) contains Spanish texts from different domains which were translated by an interpreter into LSE. However, the video data are not publicly available. Another example of data with limited availability in which LSE is the target language is the material produced by the Fundación CNSE (State Conferedation of Deaf),[7] such as an online driving license manual platform.[8] The signed videos are accessible online but the source texts are only visible in the images displaying street signs, hence, not easy to compile nor ML-usable. In most cases, not many metadata are provided concerning the source of the video material. Therefore, it is not immediately possible to evaluate the quality and the authenticity of the signing.

3 The XSL Hotel Review Corpus

The XSL Hotel Review Corpus is a multilingual parallel corpus of Sign Language of the Netherlands (Nederlandse Gebarentaal - NGT), Flemish Sign Language (Vlaamse Gebarentaal - VGT), Spanish Sign Language (Lengua de Signos Española, LSE),written English, Dutch, Spanish and Irish.

The focus on a restricted domain ensures recurrence of similar constructions and terms and facilitates the mapping of words or messages to different realisations of signs or signed utterances.

The choice for the domain of hospitality was motivated by the results of co-creation events of the SignON project:[9] during these events, deaf individuals identified the set of circumstances connected to hotels, restaurants, etc., as an appropriate environment in which sign language technology tools would be useful and acceptable. This relates to the concern of some members of the deaf community about the use of these technologies in sensitive or critical situations in which the presence of a human interpreter is preferred.

3.1 Written text

The English source text was taken from the Hotel Reviews dataset publicly available on Kaggle.[10] The original dataset contains a list of 1,000 hotels and their reviews provided by Datafiniti's Business Database. The dataset includes hotel location, name, rating, review data, title, username, and more.

For XSL HoReCo, we only used a selection of the actual hotel reviews. 300 reviews were selected according to the following criteria:

- The review is in English;
- The text is grammatically complete and correct;
- The text does not contain uncommon abbreviations (e.g. *mntns* for 'mountains').

[7] https://www.fundacioncnse.org
[8] https://www.fundacioncnse.org/dgt/
[9] https://signon-project.eu/
[10] https://www.kaggle.com/datasets/datafiniti/hotel-reviews

- In the case in which the review contains incomplete sentences, the removal of these does not affect the meaning of the whole text (an example is provided in table 1).

Original text	Text without incomplete sentence
The Southside Motel and Marina is a diamond in the rough. My room had a comfortable king size bed, nice size fridge, microwave and coffee pot. The room was clean and the staff went out of their way to make sure I always had clean towels, the room and was clean and that I had coffee supplies. The motel owners... More	The Southside Motel and Marina is a diamond in the rough. My room had a comfortable king size bed, nice size fridge, microwave and coffee pot. The room was clean and the staff went out of their way to make sure I always had clean towels, the room and was clean and that I had coffee supplies.

Table 1: Example of reviews with incomplete sentences whose removal does not affect the meaning of the text as a whole.

Within the XSL-HoReCo project, the selected reviews were translated into different languages. The translations from English into Dutch and into Spanish were performed by professional translation companies which used automatic translation (generated by DeepL) followed by in-depth human post-editing. The Irish translation was performed manually by a professional translation company.

XSL-HoReCo consists of 297 hotel reviews, corresponding to 21,464 words in the English source, 22,274 words in Dutch, and 26,469 words in Irish. Only 283 reviews were translated into Spanish,[11] which consists of 20,470 words. A distribution of the length of these reviews is presented in Figure 1 and shows that most reviews have a length between 100 and 350 characters.

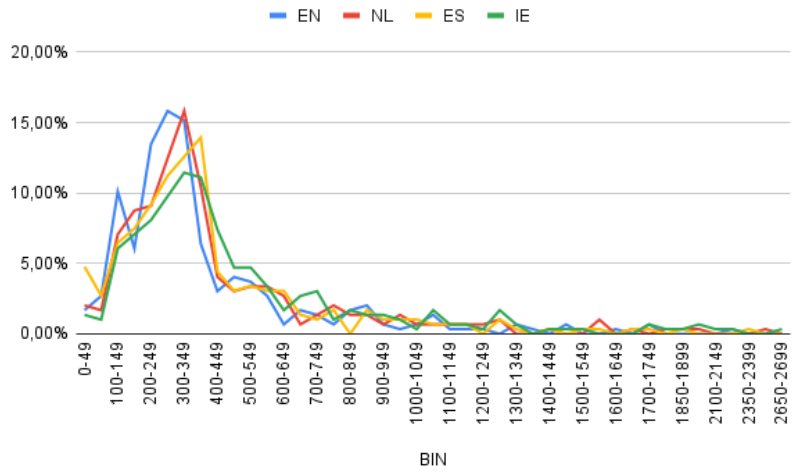


Figure 1: Distribution of the length (in characters) of the different reviews (in bins of 50 characters).

3.2 Translation into signed languages

The translations into the three signed languages (i.e. NGT, VGT and LSE) were produced according to the same guidelines (see 'Translation specifications' below), concerning the translators, the types of videos produced, and the availability of the data. All translations were made by deaf translators. This reduced as much as possible the interference of the source language. Note that NGT and VGT were translated from the manually post-edited Dutch, while LSE was translated straight from English.

[11] Due to limited budget.

The Dutch text was translated into NGT and VGT by six deaf professional translators each. For the translations into NGT, reviews were shared among translators as shown in table 2. Four translators for the NGT-HoReCo were women. In total, 167 reviews were produced by female signers.

Signer / Translator	**No. of videos**
P1	50
P2	21
P3	28
P4	49
P5	101
P6	48
Total	297

Table 2: Distribution of videos across NGT-HoReCo signers

For VGT, five were recent graduates from KU Leuven's training program for deaf translators and interpreters. Translations were divided among translators by assigning to each of them a (close to) equal number of words. Four interpreters were female and two were male.

Translations into LSE were produced by a single translator, due to the very limited availability of deaf professional LSE translators.

Figure 2 shows the distribution of the lengths of the videos for NGT, VGT and LSE, which are mostly between 10 and 60 seconds. Figure 3 shows an example of the videos and texts of the XSL-HoReCo.

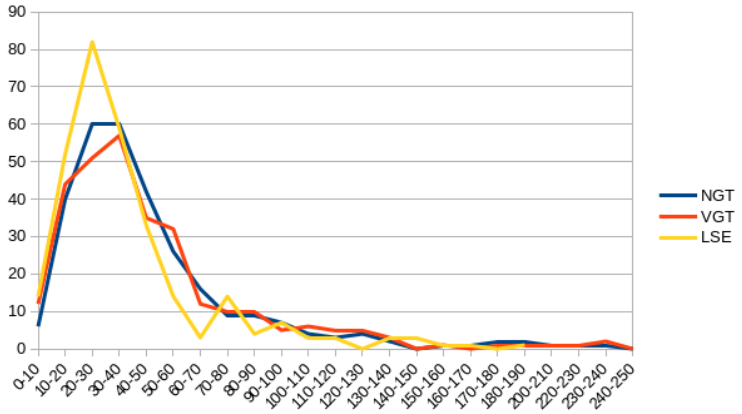


Figure 2: Distribution of lengths of videos (in seconds) in NGT, VGT and LSE

XSL-HoReCo consists of 03:51:45 hours of NGT videos, 03:59:04 hours of VGT videos and 03:09:49 hours of LSE videos, for a total of 11:00:38 hours of recording.

Translation specifications Translators were asked to make the recordings in an everyday-life, quiet environment, with a high quality camera. Each video contains one signer translating one review. Each review has been translated once. A future possible expansion of the corpus could include more translations of the same review to better account for inter-signer variation. Nevertheless, given that the corpus focuses on a single domain, i.e. hospitality, a certain recurrence of topics and signs in different possible combinations is already attested; therefore, even if to a limited extent, it allows to account for inter and intra-signer variation.

No time constraint was set for the preparation of a translation before video recording. This was done to ensure the quality of the translation and avoid a "simultaneous-interpretation effect": during simultaneous interpretation, interpreters are under time-pressure and often need to prioritize efficiency on preserving the complete content of the original message. By having the possibility of preparing the translation

beforehand, XSL-HoReCo translators could make sure that the content of the reviews would be preserved as much as possible during the translation process.

English source text	Automatically translated Dutch text	Postedited Dutch text	Automatically translated Spanish text	Postedited Spanish text	Manually translated Irish text
There's a cool area with shopping and excellent restaurants about half a mile up the road. It's a quick walk but the complimentary hotel shuttle can zip you over too.	Er is een leuke buurt met winkels en uitstekende restaurants ongeveer een halve mijl verderop. Het is een korte wandeling, maar de gratis hotelshuttle brengt u er ook heen.	Op ongeveer 800 meter afstand is een cool gebied met winkels en uitstekende restaurants. Het is een korte wandeling, maar je bent er ook zo met de gratis hotelshuttle.	Hay una zona interesante con tiendas y excelentes restaurantes a aproximadamente media milla de la carretera. Es una caminata rápida, pero el servicio de transporte de cortesía del hotel también puede acercarlo.	Hay una zona guay con tiendas y excelentes restaurantes a aproximadamente media milla de la carretera. Es una caminata rápida, pero el servicio de transporte gratuito del hotel también puede acercarte.	Tá áit iontach ann ina bhfuil siopaí agus bialanna den scoth thart ar leathmhíle suas an bóthar. Is siúlóid ghairid é ach is féidir leis dul ansin ar an tointeáil óstáin saor in aisce freisin.

Figure 3: Example from NGT, LSE and VGT-HoReCo

3.3 Availability

To ensure the availability of the data for future research, all translators signed an informed consent form agreeing with the data being publicly available under the CC-BY NC licence. The accessibility and format of the XSL-HoReCo project makes it easily expandable with more parallel languages and/or additional annotations.

The NGT side of the HoReCo is available through the European Language Grid at https://live.european-language-grid.eu/catalogue/corpus/21535 and through CLARIN at http://hdl.handle.net/10032/tm-a2-x7.

The VGT side of the HoReCo is available through the European Language Grid at https://live.european-languagegrid.eu/catalogue/corpus/23007, and through CLARIN at http://hdl.handle.net/10032/tm-a2-y3.

The LSE side of the HoReCo is available through the European Language Grid at https://live.european-language-grid.eu/catalogue/corpus/23263 and is available through CLARIN at http://hdl.handle.net/10032/tm-a2-x6.

4 GoSt-ParC-Sign

The Gold Standard Parallel Corpus of signed and spoken language focuses on spontaneous and semi-spontaneous VGT and its translation into written Dutch.

The GoSt-ParC-Sign project was developed in three phases: data gathering, manual translation, and quality control. All phases were coordinated and overseen by the Vlaamse GebarentaalCentrum (Flemish Sign Language Centre).

4.1 Phase 1: Data collection

During the first phase, roughly ten hours of publicly available semi-spontaneous VGT videos were initially identified. All VGT material contained in this corpus was produced by deaf authentic signers for a signing audience.[12] Therefore, the quality of the signing is as close as it could possibly be to real life signing. Written consent was gathered from the authors and signers of these videos to ensure that we would be allowed to redistribute the material.

The final content of the corpus is presented in Table 3 and amounts to just about 10 hours of footage.

Corpus Part	**Duration**
Spontaneous conversation from the VGT corpus	3:11:05
Talkshow ”Dagelijks Doof”	2:24:24
Vlog regarding typical language use in VGT	1:15:24
Game show ”wie wordt miljonair”	1:07:00
Various research rapports professionally translated into VGT	1:46:06
Opinion pieces in VGT	0:13:18
Total	9:57:07

Table 3: Description of the sources of the differnt parts in the GostParc-Sign corpus

The footage contains 43 different signers of different ages and different regions, as presented in Table 4. Age groups are presented in Table 5. This information is relevant for future work on this corpus because much variation and differences are attested across different regions and age groups. Since the corpus only contains already existing data, we had no real control over the distribution of these sociolinguistic factors.

Region	**Men**	**Women**
West-Flanders	4	5
East-Flanders	8	7
Flemish-Brabant	1	0
Antwerp	8	4
Limburg	1	3
Total	22	21

Table 4: GostParc signers: origin and gender

The data in the two languages are aligned at the sentence (or message) level, since there is no one-to-one correspondence between VGT signs and Dutch words.

4.2 Phase 2. Manual translation

The second phase focused on the manual translation task from VGT into Dutch text. Translations were performed by mixed teams of deaf and hearing translators, in total four deaf and six hearing translators were involved. Having a mixed team had a double purpose:

i. ensuring that the original meaning of the signed message was preserved through the deaf translator;

[12] For sign languages, it is problematic to talk about native language, since most deaf children are born in hearing families and get exposed to a sign language only later on in life. Consequently, we prefer the term ‘authentic’, indicating that the individual uses a sign language as their main and preferred language.

Age group	Men	Women
12-18	3	7
19-25	3	4
26-35	7	5
36-50	2	4
51-70	6	2

Table 5: Age groups (at the time of recording)

ii. providing a good quality Dutch text, through the native Dutch translator.

Translations were organised in ELAN (Sloetjes & Wittenburg, 2008), which allows to synchronise multiple annotation tiers with the video timeline. A 'Translation' tier was created for each of the participants in the video to contain the written Dutch translation in each ELAN Annotation Format (EAF) file of each video (an example of the format is provided in Figure 4). The image shows one tier for each of the four participants in the talkshow, this way even overlapping utterances could be correctly captured (as seen in the image). Having files in EAF can serve for linguistic research; in addition, this format can be easily adjusted into an ML-suited format with the framework proposed in De Sisto et al. (2022).

Our initial estimation was that 133 hours of translation work would lead to approximately 9–10 hours of videos being translated. This estimation was based on a consultation with professional signed language to spoken language translators, according to which we concluded that 15 minutes of translation work would correspond roughly to one minute of video translation. Unfortunately, during the translation phase we realised that, given the breadth of the topics covered in the video and the spontaneity of the signing, more time was needed for translating them. In total between 180 and 220 hours of translations lead to 8 hours of videos being translated. Consequently, in order to reach the target of having 10 hours of videos, we opted for including in the corpus additional footage publicly available with subtitles.

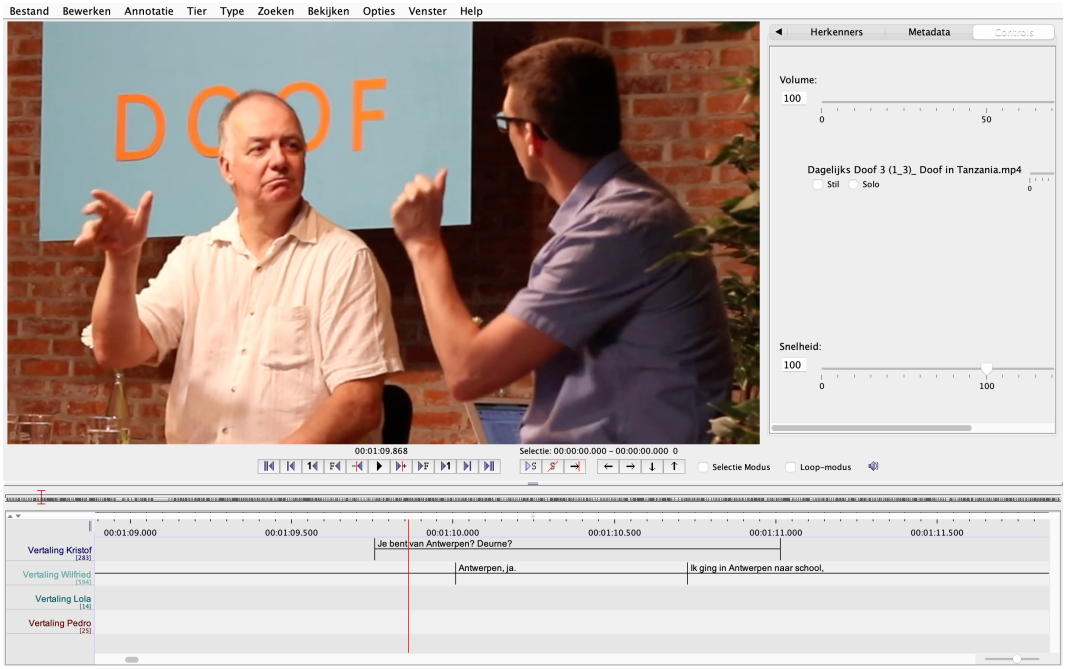


Figure 4: Example of GoSt-ParC-Sign's data format

4.3 Phase 3: Quality control

In the third phase the quality control of the translations was performed by a professional editor who was not part of the initial translation team, to ensure that the produced translations were correctly reporting the original message of the VGT videos and that the Dutch texts were of high quality.

This corpus is made available under CC BY license, at the Instituut voor de Nederlandse Taal (INT) at http://hdl.handle.net/10032/tm-a2-x9 and will soon be made available on the European Language Grid.

5 Conclusion

In this paper we have introduced two signed language data collection projects which aim at supporting advances in more inclusive language technology which also targets signed languages. The XSL-HoReCo project led to the creation of a multilingual parallel corpus of NGT, VGT, LSE, English (source text), written Spanish, Irish and Dutch. The very recently concluded GoSt-ParC-Sign project produced a parallel corpus of authentic VGT videos and a translation into written Dutch. The creation of similar parallel data is fundamental for supporting research and developments into fields such as signed language translation, recognition and processing.

In addition, another important outcome of these data collection project is the lesson learned throughout the process and from the challenges encountered, which can be useful for future high quality signed language data collection projects:

- Guidelines for recording a video need to be extremely clear and specific; potential vagueness might lead to differences in the quality, style and type of the recording.
- It is quite difficult, if not impossible, to have an exact estimation of the ratio of translation time needed per hour of signed language videos. Many factors are at play, which affect the translation process, such as topics discussed, spontaneity of the signing, monologue vs. group conversation, potential peculiarities of individual signing styles, etc. In the GoSt-ParC-Sign project, our initial estimation turned out to be dramatically lower than the actual time needed by translators.
- The translation of spontaneous signing can be particularly challenging. Just as with spontaneous speech, unplanned signing might contain unnecessary repetitions, unclear articulations, false starts; in some circumstances, identifying what is being signed can be challenging even for an expert authentic user of a signed language, independently from whether the user is deaf, hard of hearing or hearing.
- Signers remain at all time the owners of the data they produce.

Acknowledgements

The initial NGT-HoReCo project (containing NGT and English and Dutch) has been funded by the SRIA contribution Projects of the Eurpean Language Equality 2 project.[13] Extensions to VGT, LSE, Irish and Spanish were funded by the SignON project which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101017255.

The GoSt-ParC-Sign project has been awarded the EAMT Sponsorship of Activities 2022 and is partially funded by the SignON project.

References

Camgöz, N. C., Saunders, B., Rochette, G., Giovanelli, M., Inches, G., Nachtrab-Ribback, R., & Bowden, R. (2021). Content4all open research sign language translation datasets. *CoRR*, *abs/2105.02351*. https://arxiv.org/abs/2105.02351

[13] https://european-language-equality.eu/

Crasborn, O., Zwitserlood, I., & Ros, J. (2008). Het Corpus NGT. Een digitaal open access corpus van filmpjes en annotaties van de Nederlandse Gebarentaal. Nijmegen: Centre for Language Studies, Radboud University. https://www.corpusngt.nl/

Crasborn, O., Zwitserlood, I., Van der Kooij, E., & Bank, R. (2020). *Annotation conventions for the Corpus NGT* (tech. rep.). Radboud University Nijmegen, Centre for Language Studies and Department of Linguistics.

De Sisto, M., Vandeghinste, V., Egea Gómez, S., De Coster, M., Shterionov, D., & Saggion, H. (2022). Challenges with sign language datasets for sign language recognition and translation. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2478–2487. https://aclanthology.org/2022.lrec-1.264

Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, *141*, 108–125.

Kopf, M., Schulder, M., & Hanke, T. (2021). *Overview of datasets for the sign languages of Europe* (Project deliverable No. D6.1). EASIER Consortium. https://doi.org/10.25592/UHHFDM.9560

Kopf, M., Schulder, M., & Hanke, T. (2022). The Sign Language Dataset Compendium: Creating an overview of digital linguistic resources. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, J. Mesch, & M. Schulder (Eds.), *13th international conference on language resources and evaluation (LREC 2022). Proceedings of the 10th workshop on the representation and processing of sign languages: Multilingual sign language resources* (pp. 102–109). European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/pub/22025.pdf

Morgan, H. E., Crasborn, O., Kopf, M., Schulder, M., & Hanke, T. (2022). Facilitating the spread of new sign language technologies across Europe. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, J. Mesch, & M. Schulder (Eds.), *13th international conference on language resources and evaluation (LREC 2022). Proceedings of the 10th workshop on the representation and processing of sign languages: Multilingual sign language resources* (pp. 144–147). European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/pub/22026.pdf

Nonhebel, A., Crasborn, O., & van der Kooij, E. (2004). *Sign language transcription conventions for the echo project. annotation convention. version 9* (tech. rep.). University of Nijmegen. http://hdl.handle.net/2066/57889

Pasikowska-Schnass, M. (2018). *Sign languages in the EU* (tech. rep.). European Parliamentary Research Service. http://www.europarl.europa.eu/RegData/etudes/ATAG/2018/625196/EPRS_ATA(2018)625196_EN.pdf

Porta, J. (2014). *Towards a rule-based Spanish to Spanish sign language translation: from written forms to phonological representations* [Doctoral dissertation, Universidad Autónoma de Madrid. Departamento de Tecnología Electrónica y de las Comunicaciones].

Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., & Schwarz, A. (2008). DGS corpus project–development of a corpus based electronic dictionary German Sign Language/German. *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, 159.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf

Van Herreweghe, M., Vermeerbergen, M., Demey, E., De Durpel, H., Nyffels, H., & Verstraete, S. (2013). *Corpus Vlaamse Gebarentaal: Annotatierichtlijnen* (tech. rep.). Universiteit Gent and KU Leuven.

Vandeghinste, V., De Sisto, M., Egea Gómez, S., & De Coster, M. (forthcoming). Challenges with sign language datasets. In A. Way, L. Leeson, & D. Shterionov (Eds.), *Sign language machine translation*. Springer.

Vandeghinste, V., De Sisto, M., Kopf, M., Schulder, M., Brosens, C., Soetemans, L., Omardeen, R., Picron, F., Van Landuyt, D., Murtagh, I., Avramidis, E., & De Coster, M. (2023). *Report on Europe's Sign Languages* (tech. rep.). European Language Equality D1.40. https://zenodo.org/records/8047005Domain

Vandeghinste, V., Van Dyck, B., De Coster, M., Goddefroy, M., & Dambre, J. (2022). Becos corpus: Belgian covid-19 sign language corpus. a corpus for training sign language recognition and translation. *Computational Linguistics in the Netherlands Journal*, *12*, 7–17. https://clinjournal.org/clinj/article/view/144

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.

Koenraad De Smedt, Iulianna Van der Lek, Henk Van den Heuvel, Antonio Balvet, Maarten Janssen, Silvie Cinková, Amelia Sanz, Stavros Assimakopoulos and Louis ten Bosch 2023. CLARIN in Training and Education. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 34–49. DOI: https://doi.org/10.3384/ecp210

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

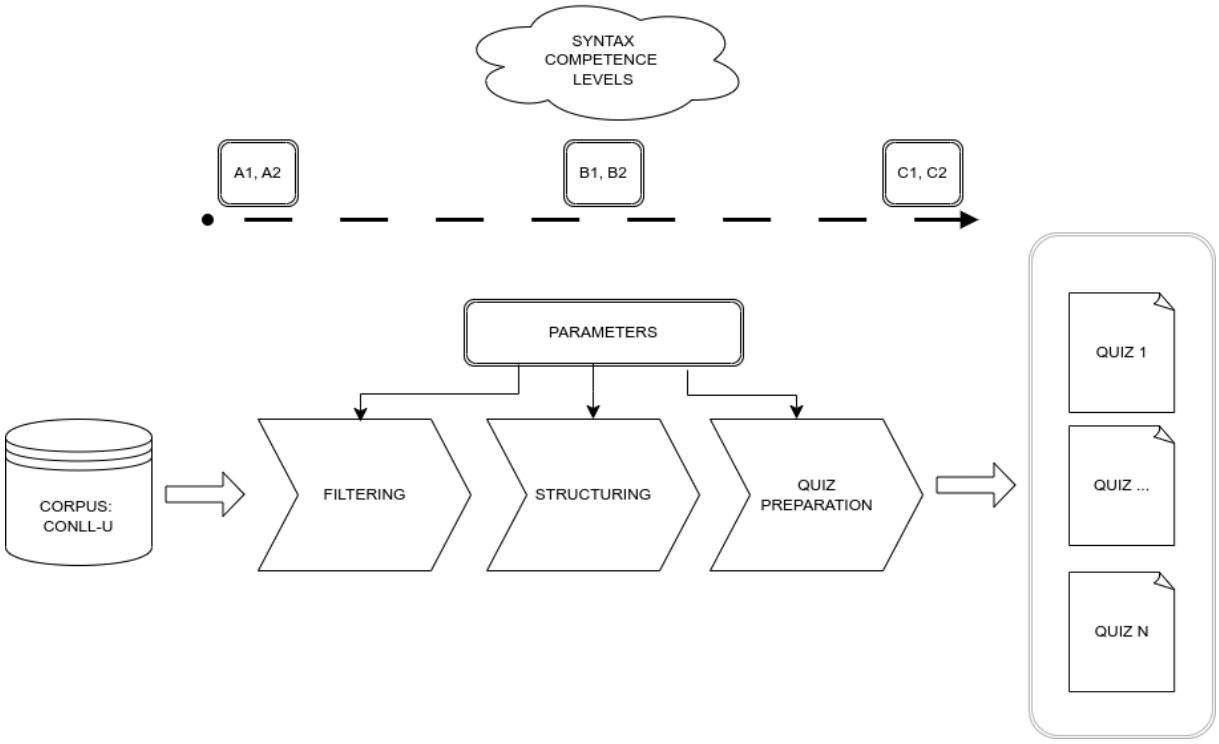


Figure 1: Corpus-to-quiz processing chain.

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 *compl**ément***, 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.

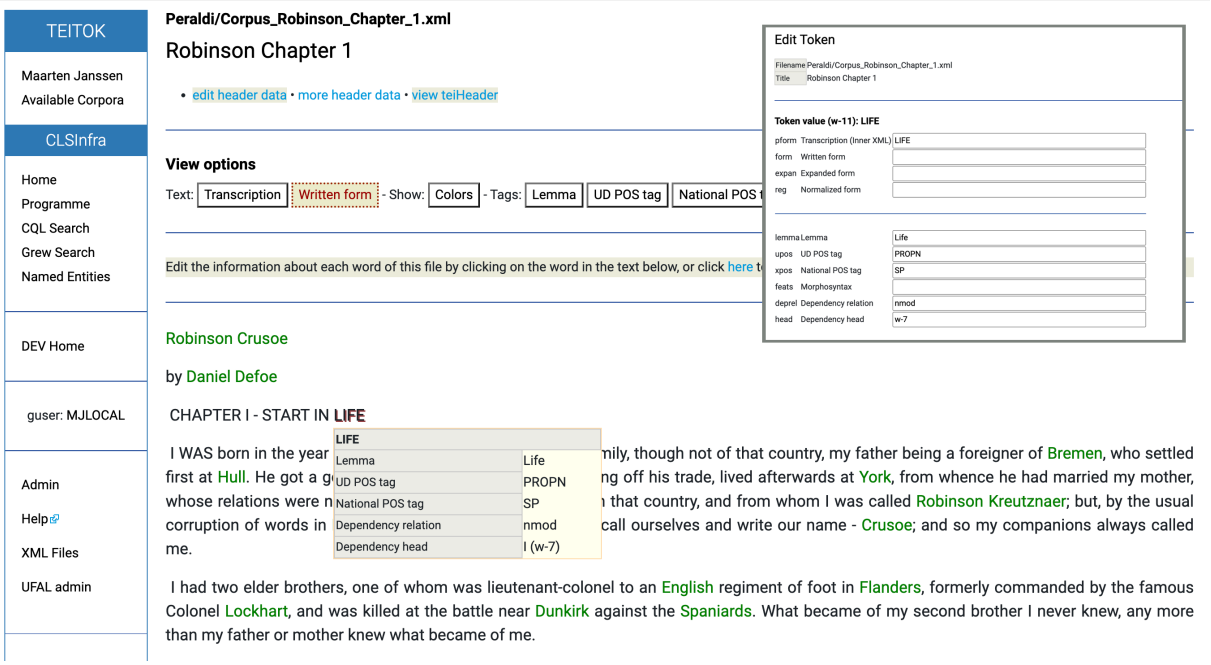


Figure 2: Document view and token edit in TEITOK.

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.

The CLARIN:EL infrastructure: Platform, Portal, K-Centre

Maria Gavriilidou
ILSP / Athena RC, Greece
maria@athenarc.gr

Stelios Piperidis
ILSP / Athena RC, Greece
spip@athenarc.gr

Dimitrios Galanis
ILSP / Athena RC, Greece
galanisd@athenarc.gr

Kanella Pouli
ILSP / Athena RC, Greece
kanella@athenarc.gr

Penny Labropoulou
ILSP / Athena RC, Greece
penny@athenarc.gr

Juli Bakagianni
ILSP / Athena RC, Greece
julibak@athenarc.gr

Iro Tsiouli
ILSP / Athena RC, Greece
tsiouli@athenarc.gr

Miltos Deligiannis
ILSP / Athena RC, Greece
mdel@athenarc.gr

Athanasia Kolovou
ILSP / Athena RC, Greece
akolovou@athenarc.gr

Dimitris Gkoumas
ILSP / Athena RC, Greece
dgkoumas@athenarc.gr

Leon Voukoutis
ILSP / Athena RC, Greece
leon.voukoutis@athenarc.gr

Katerina Gkirtzou
ILSP / Athena RC, Greece
katerina.gkirtzou@athenarc.gr

Abstract

This paper presents the CLARIN:EL infrastructure, which comprises three pillars: the language resources and technologies Platform, the Portal and the Knowledge Centre. It serves as a comprehensive and interoperable environment that supports language-related research in the fields of language technology, language studies, digital humanities, and political and social sciences. The Platform facilitates deposition, curation and sharing of digital language resources (catering for providers' needs), and access to and automatic processing of these resources (catering for consumers' needs). The Portal offers informative material about CLARIN:EL and support services to the community, including dissemination, awareness raising and training activities. The Knowledge Centre promotes digital literacy in the scientific domains served, by providing information on studies, educational and training material and publications. This paper discusses the CLARIN:EL pillars, the technical architecture, its design and implementation principles, the functionalities offered to the users, the support activities provided, usage analytics and future steps.

1 Introduction

CLARIN:EL is the Greek National Infrastructure for Language Resources & Technologies, which comprises three interconnected pillars, namely, the Platform, the Portal and the NLP:EL Knowledge Centre. CLARIN:EL serves as a comprehensive and interoperable environment that supports language-related research in various fields, such as language technology (LT), linguistics, language studies, digital humanities (DH), political and social sciences. The Platform hosts the Language Resources and Technologies and provides the user interaction mechanisms through appropriate interfaces. The Portal and the K-Centre cater for dissemination, offer informative material, and support the community as regards awareness, training, and knowledge transfer in LT and DH.

Maria Gavriilidou, Stelios Piperidis, Dimitrios Galanis, Kanella Pouli, Penny Labropoulou, Juli Bakagianni, Iro Tsiouli, Miltos Deligiannis, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis and Katerina Gkirtzou 2023. The CLARIN:EL infrastructure: Platform, Portal, K-Centre. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 50–61. DOI: https://doi.org/10.3384/ecp210

At the national level, CLARIN:EL is part of the Greek Roadmap for Research Infrastructures; currently, it forms part of the APOLLONIS infrastructure, together with DARIAH/DYAS. At the European level, it is the Greek branch of the CLARIN ERIC infrastructure (Branco et al. 2023). It supports the community through a certified CLARIN B-Centre and a K-Centre[1], it has been awarded the CoreTrustSeal[2], and it is listed in re3data (the registry of research data repositories)[3]. The CLARIN:EL network supporting the Infrastructure consists of 14 organization members (9 Universities and 5 Research Centres) practically covering the whole geographical area of the country. CLARIN:EL currently (February 2024) contains 793 resources (648 corpora, 94 lexical resources, 49 tools/services, and 2 language descriptions).

The CLARIN:EL infrastructure, with its Platform, the Portal, the NLP:EL K-Centre, the network of organizations supporting it, and, additionally, with the technical and operational interconnection with CLARIN ERIC, constitutes a valuable universe supporting language technology at the national and European levels. It serves both linguists and non-linguists, academics and non-academics, students, educators and language professionals, industry, and broad public. Through the use of concrete licensing schemes for the distribution of language resources and services, it actively safeguards Open Access and Open Science and promotes the requirements for open language data and language technology.

The following sections discuss the three pillars of CLARIN:EL with their components, the design of the infrastructure and the implementation principles; the functionalities of the infrastructure (deposition, documentation, curation of resources and services, search, retrieval and processing of resources, user management, and dashboard), the technical architecture, the support activities and materials offered to the users, as well as training and dissemination activities; finally, the paper provides infrastructure usage analytics, and concludes with future steps.

2 The CLARIN:EL Platform

The CLARIN:EL Platform consists of two interconnected subsystems: (a) the Repository, a system with all functionalities related to the provision of LRs, i.e., for depositing and documenting LRs, for curating their metadata through a specially designed metadata editor, for storing, sharing, searching, retrieving, and downloading LRs, and (b) the Workbench, a system providing integrated services that perform core Natural Language Processing (NLP) tasks, such as sentence splitting, tokenization, PoS tagging, lemmatization, parsing, chunking, named entity recognition (Prokopidis & Piperidis, 2020), as well as tasks such as text classification and verbal aggression analysis (Pontiki et al., 2020). Moreover, it offers preprocessing services that perform data format and character encoding conversion.

The Platform offers access to the resources through the Central Inventory, which provides a comprehensive catalogue of the resources and tools (corpora, lexical/conceptual resources, language descriptions and tools/services), for Greek (on its own or in combination with other languages).

The Central Inventory includes metadata records for (a) CLARIN:EL hosted data or software, (b) data or software that reside outside the CLARIN:EL platform, (c) reference data (i.e., bibliographical lists, useful catalogues, etc.). The Central Inventory can be filtered according to various features, such as resource type, language, domain, depositing organization, etc.

2.1 Deposition of Language Resources

Depositors of data or language processing services must be registered CLARIN:EL users, either affiliated to network member organizations or individuals. Resources provided by network members are associated, through the relevant metadata, to the specific organization, while those provided by individuals non-affiliated to a member organization are assigned to the Hosted Resources Repository. Resources deposited encompass written, spoken, or multimodal content. They can be sets of texts, lexical resources, language models or processing tools, and they may pertain to modern Greek language, to earlier forms of Greek, or to other languages. To be processable by the integrated services of CLARIN:EL, the corpora must be in one of the recommended text formats (plain text, XML, TMX, etc.)

[1] https://www.clarin.eu/content/clarin-centres
[2] https://www.coretrustseal.org/
[3] https://www.re3data.org/

(Piperidis et al., 2016). Data providers can get guidance and assistance via the Help pages[4] and the relevant Policy documents, Data Collection Policy[5] and Deposition Documentation[6]. CLARIN:EL offers support on various issues such as data formats, metadata, and legal aspects, through the Recommended Formats guidelines[7], online documentation for metadata and data preparation, documentation and deposition, video tutorials, and helpdesks.

CLARIN:EL favours and promotes Open Licenses; however, existing restrictions on data distribution and/or use are respected. CLARIN:EL offers a variety of standard licenses for the provider to select from, and assistance through the Legal Helpdesk. The responsibility of clearing IPR and selecting the appropriate license for the resources provided lies with the resource provider. Metadata of CLARIN:EL resources are freely available to all with a CC-BY 4.0 license.

2.2 Documentation of resources with metadata

To ensure appropriate description of deposited resources, CLARIN:EL has adopted a rich metadata schema, CLARIN-SHARE[8], which allows coherent documentation to be added to each resource. The CLARIN-SHARE metadata model builds upon the META-SHARE metadata model (Gavriilidou et al., 2012), and its application profiles, ELG-SHARE (Labropoulou et al., 2020), ELRC-SHARE (Piperidis et al., 2018), and the MS-OWL ontology (Khan et al., 2022; McCrae et al., 2015), RDF/OWL representation of the model.

The CLARIN-SHARE schema supports the objectives of the Platform. In particular, it ensures that all resources are discoverable and accessible by human users and machines (e.g., including links to URLs that offer direct access to the resource), addresses researchers' needs such as data citation (through the use of persistent and unique identifiers) and replicability of experiments (ensuring persistence and reusability of resources), and facilitates the integration of processing services with data resources (using as interoperability anchors the same attributes and values, such as data formats and annotation types, across resource types) and the documentation of processing activities and their outcomes (provenance/lineage metadata). Overall, the CLARIN-SHARE schema is an important factor contributing towards achieving FAIR data (Wilkinson et al., 2016).

The schema builds along three key concepts, each of which is associated with a distinctive set of metadata attributes:

- *resource type*, classifying resources into *corpora* (sets of text, audio, video or image files), *lexical/conceptual resources* (e.g. lexica, glossaries, ontologies, etc.), *language descriptions* (including models and computational grammars), and *tools/services*
- *media type*, which specifies the form or physical medium of the resource, i.e., *text*, *audio*, *image*, *video* and *numerical text* (referring to numerical data, such as biometrical, geospatial data, etc.). To cater for multimedia and multimodal language resources (e. g., a corpus of videos and subtitles, or a corpus of audio recordings and transcripts, a sign language corpus with videos and texts, etc.), language resources are represented as consisting of at least one media part, while multiple parts are also possible;
- *distribution*, which, following the DCAT vocabulary[9], refers to any physical form of the resource that can be distributed and deployed by end-users.

These concepts give rise to a modular structure, in which attributes are attached to the appropriate level. The level of "Language resource" includes properties common to all resource and media types, such as those used for identification purposes (title, description, etc.), recording provenance (creation, publication dates, creators, providers, etc.), contact points, etc. More technical features and classification attributes differ across resource and media types and are, thus, attached to combinations thereof; for example, a corpus may take properties specific to annotation processes, while the description of a

[4] CLARIN:EL User manual: https://clarin-platform-documentation.readthedocs.io/en/stable/
[5] Data Collection policy: https://www.clarin.gr/sites/default/files/CLARINELDataCollectionPolicy.pdf
[6] Deposition documentation: https://clarin-platform-documentation.readthedocs.io/en/stable/all/4_Data/DataPreparation.html?highlight=deposit
[7] Also in the Documentation: https://clarin-platform-documentation.readthedocs.io/en/stable/all/4_Data/FileFormats.html
[8] https://clarin-platform-documentation.readthedocs.io/en/stable/all/5_Metadata/Full.html
[9] https://www.w3.org/TR/vocab-dcat-2/

computational lexicon encodes whether it includes lemmas, examples, grammatical information, translation equivalents, etc. Technical features, such as format, size, information on licensing and mode of access are properties of the distribution.

The schema includes properties for the description of the full life cycle of language resources, from conception and creation to integration in applications and usage. All this information leads to a complex and demanding schema; to ensure flexibility and uptake by providers, only a carefully selected subset of these attributes are prescribed as *mandatory* and, thus, required to be filled in for metadata records to be approved for import in the Platform. The remaining attributes are *recommended* or *optional* i.e. providers are encouraged to fill them in and enhance the discoverability and usability of their resources.

To foster the visibility and reusability of data, CLARIN:EL exposes metadata for harvesting, thus extending their discovery. The CLARIN-SHARE metadata schema has been mapped into the broadly used metadata schemas Dublin Core and OLAC, so that the metadata records of the resources are harvested by repositories and infrastructures that support the OAI/PMH harvesting protocol [10] (e.g., CLARIN Virtual Language Observatory/VLO[11]). In addition, CLARIN-SHARE has been implemented in the form of four profiles (one for each resource category) following the principles of the Component Metadata Infrastructure and integrated in the Component Metadata Registry[12], thus enhancing its reusability and interoperability in the CLARIN framework.

Resource providers in CLARIN:EL have two options for creating metadata for their resources: to create and upload XML files that adhere to the CLARIN-SHARE metadata schema or to create metadata records using the platform's metadata editor. Users can choose to validate XML files prior to uploading them. If any inconsistencies or missing metadata are detected, error messages will be displayed, prompting the user to address the issues. Once corrections have been implemented, the XML files can then be successfully uploaded. The infrastructure supports the uploading of files either individually or in batches. The submitted XML files are once more automatically checked for completeness and well-formedness.

If users opt to create metadata for their resource through the metadata editor, they will be presented with four specific forms tailored to the four types of resources. After selecting the type of resource they want to deposit, users are redirected to the relevant form (Figure 1), where they are required to fill in at least the mandatory metadata elements.

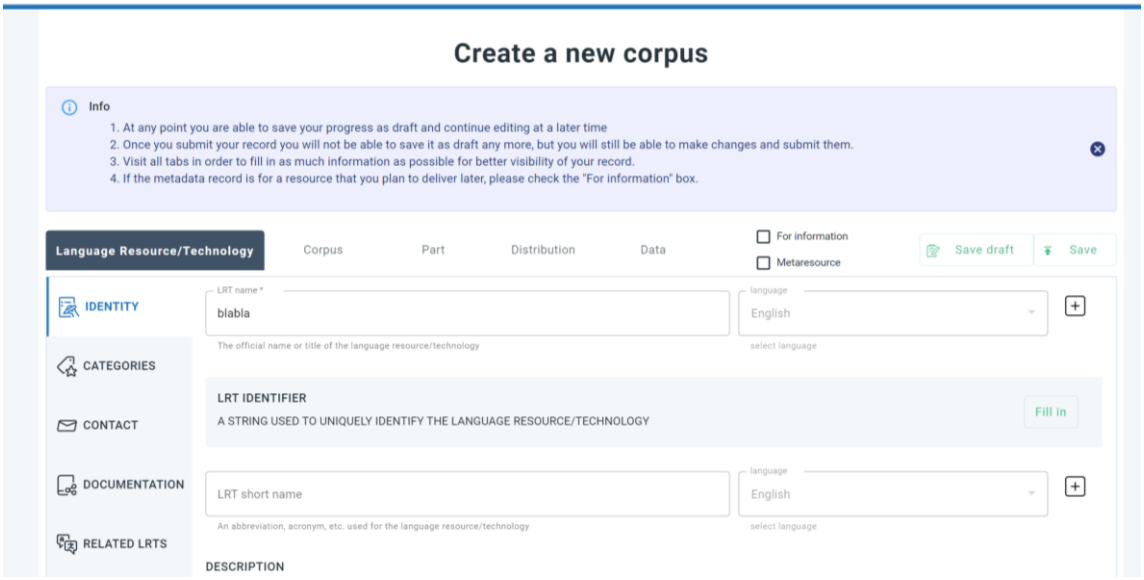


Figure 1: Creation of the resource record for a corpus

[10] https://www.openarchives.org/pmh/
[11] https://vlo.clarin.eu
[12] https://catalog.clarin.eu/ds/ComponentRegistry/#/

The metadata record cannot be saved until all required fields are filled in as indicated. Thus, the metadata editor guides the providers to the complete description and uploading of their resources, through iterative checks that make sure that all obligatory elements are filled in; it safeguards well-formedness and facilitates metadata interoperability by using controlled vocabularies (where applicable), and assists them with examples and tips.

2.3 Curation of metadata

The completion of the description by the depositor and the automatic checking by the system are followed by two rounds of manual assessment. The first round involves metadata and legal validation performed by human validators, followed by the final approval by the supervisor of each organization member, which triggers the resource's publication in the repository. Frequent quality checks, aiming at the completeness and correctness of metadata records and related datasets, are conducted centrally by the dedicated CLARIN:EL technical and metadata team. The dataset(s) associated with a metadata record are also automatically checked to assess their conformity with technical specifications as regards format and processability (interoperability with processing services). Providers are notified in cases of erroneous or sub-optimal codification of metadata, which they are requested to rectify.

Even when a resource needs to be removed from the central inventory, its findability is ensured; in these instances, users will be directed to a tombstone page, where a message indicates that the resource is (temporarily or permanently, as appropriate) unavailable.

2.4 Search and retrieval of Language Resources

CLARIN:EL presents resources in a Central Inventory (Figure 2), which users can search using keywords and facets, or browse and select a resource to view its full description; if interested, they can download it, or use the CLARIN:EL services to process it. The inventory lists metadata records containing the descriptions of the LRs (datasets or tools), which are normally (but not necessarily) accompanied by datasets or software code. Resources with no data fall into two categories: (a) metadata records in anticipation of the data that is not yet ready to be published, and (b) meta-resources, i.e., ancillary resources (e.g., bibliographical lists, literature reviews, etc.). The functionalities of browsing, viewing, and exporting metadata records, as well as downloading open-access resources are available to all users (registered or not), while user authentication and authorization are required for the use of the platform's processing services or accessing restricted resources. The downloadability of a resource depends on the license defined by the provider. Legal and technical restrictions on resources are specified by the provider via the relevant metadata elements, based on which CLARIN:EL implements the resource's access policy. For the content files of a resource to be accessible, two criteria must be met: an open access license, and storage of the content files at an access point within CLARIN:EL or externally.

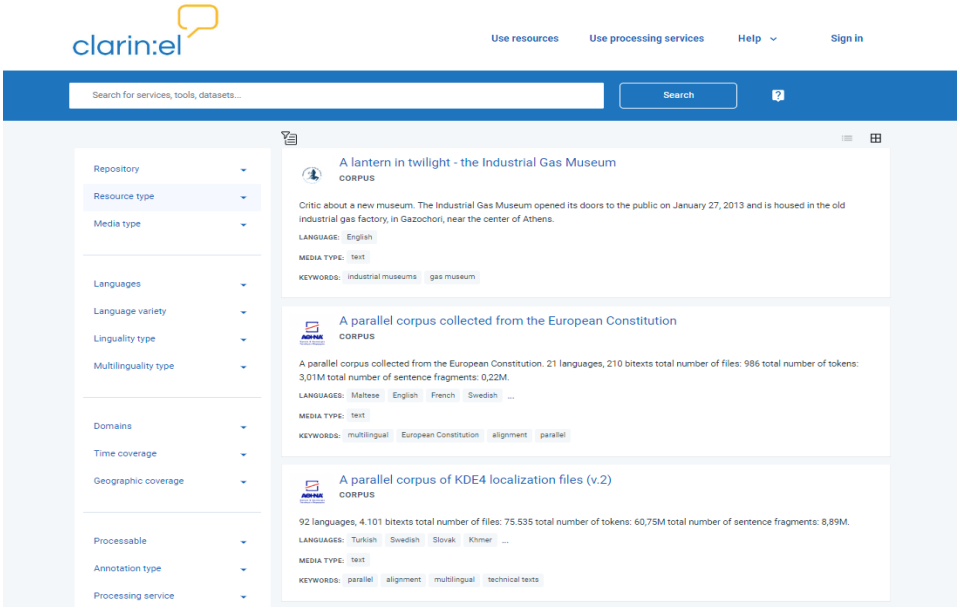


Figure 2: The CLARIN:EL Central Inventory

2.5 Processing of Language Resources

CLARIN:EL offers two types of tools/services for processing data: (a) single-task tools (e.g., lemmatizers, tokenizers, etc.) available as web services or as downloadable tools, accessible either from within CLARIN:EL or through an external link, and (b) the CLARIN:EL Workbench[13], which includes NLP web services integrated in the CLARIN:EL infrastructure. Each single-task web service can also be part of a workflow, i.e., of a pipeline of tools that operate at multiple levels of analysis (e.g., a workflow starting from tokenization and sentence splitting, continuing with PoS tagging, lemmatization and concluding with named entity recognition). The Workbench is designed to support non-expert users in their data processing tasks, by providing ready-to-use pipelines (workflows) of interoperable tools at a single click, thereby relieving them from the burden of selecting and assembling tools in a workflow from scratch (Figure 3).

Figure 3: Selecting a processing service

Users can process datasets hosted at CLARIN:EL or upload and process their own datasets (with a size limit of 2MB, currently). In the former case, the outcome of the processing is stored in the infrastructure as a new resource, with its metadata automatically generated by combining the metadata of the dataset with those of the processing service used. The outcome of the processing is available both in the data format generated by the workflow (such as XML or XMI), and in Comma Separated Values (CSV) format. The latter is provided for reasons of user-friendliness and interoperability, given that such files can be fed to other NLP services or to visualization tools residing within CLARIN:EL and/or externally.

There are two prerequisites for the integration of a processing tool (e.g., a PoS tagger) to the CLARIN:EL infrastructure: (a) the tool has to be wrapped and offered as a web-service (e.g. via a RESTful API), and b) the software application that offers this web-service along with its dependencies (e.g., libraries, settings, operating system etc.) has to be packaged in a Docker image. The technical team of CLARIN:EL has the responsibility to deploy the Docker images of the provided tools/services at the infrastructure's Kubernetes cluster and integrate them in a software module called Workflows Manager which acts as a processing orchestrator. The orchestrator chains the tools into workflows, provides the required readers for different types of input datasets (e.g. TXT, XCES, TMX, etc.), exports the processing results to CSV, handles failures/timeouts, etc. Obviously, a large number of processing jobs can be initiated in CLARIN:EL by different registered users simultaneously. This in many cases can lead to overload; i.e., the deployed services might not be able to serve all the required requests, for example, due to memory issues and network timeouts that might occur. To avoid overstretching the capacity of the execution system, a scheduler has been implemented that decides when a processing job will start running at the available computing resources (VMs, containers, etc.).

[13] https://inventory.clarin.gr/workflows/

2.6 User and Resource-lifecycle management

Members of the academia, the research community, or the general public, whether affiliated to a network member organization or not, have full access to the infrastructure. Registered users, authenticated via their academic, personal or CLARIN ERIC accounts, can upload their resources and/or tools, use the available resources and process them using the services offered by CLARIN:EL.

Registered users have full access to all CLARIN:EL functionalities and are considered potential resource providers, either as individuals or as members of their organization. There are two ways to become a CLARIN:EL user: (a) create a personal (non-organization) account, and (b) use an existing account managed by the identity provider (IdP) of the affiliated organization (e.g., research institute, university etc.) that belongs either to the national Authentication and Authorization Infrastructure (AAI) Federation (GRNET) or to the CLARIN Service Provider Federation. In this way, CLARIN ERIC users from any country have access to the Greek network. In both cases, users are stored in the User Management module, which is based on Keycloak, an open source identity and access management solution.

The activities available to the users depend on their roles, which are also defined and managed within Keycloak. The User Roles schema foresees the roles of Curator (assigned by default to all registered users), Supervisor for each organization-member, and Validator (assigned by the Supervisor). These roles are involved in the creation and publishing of a resource, with varying rights: Supervisors have the full list of permitted actions, Validators are responsible for the legal (license) and metadata quality check, while Curators have the basic set of actions for LRs deposition.

The set of states of a resource in the process of being prepared for publication in the Central Inventory is depicted in the Resource Lifecycle[14]; these states include the creation of a new resource by a curator (resource status: Draft), the automatic checking of its syntactic validity and conformity with the specifications (status: Ingested/Syntactically valid), the submission of the resource by the curator and the assignment of the resource to validators by the supervisor (status: Assigned for Validation); after the approval of the resource by the validators (status: Approved), the supervisor publishes the resource, making it visible on the CLARIN:EL inventory (status: Published).
Each member organization is responsible for its internal User Role Management, i.e. assigning roles (Curator, Validator, Supervisor), and for ensuring efficient creation, description, and publication of their own resources. Above this User Role Management at the level of member organizations, additional Validator and Supervisor roles exist at the central level of the CLARIN:EL Platform, with rights on all resources, facilitating quality assessment and ensuring completeness and correctness.

2.7 The User Dashboard

The CLARIN:EL User Dashboard, exclusively available to registered users, functions as a specialized section of the Platform. It is a key tool enhancing the overall user experience and optimizing the utilization of CLARIN:EL resources and services. The dashboard plays a dual role, not only facilitating the creation, management and processing of language resources, but also offering users a comprehensive overview of their activity within the CLARIN:EL ecosystem. Through this interface, users benefit from an assortment of features. Specifically:

- Customization: The Dashboard supports customization, offering different views tailored to the specific roles of registered users, allowing them to access information and functionalities relevant to their roles within the CLARIN:EL platform.
- Interactivity: The Dashboard is designed to be interactive, allowing users to actively engage with CLARIN:EL resources and services. This interactivity includes creating and uploading resources, utilizing Natural Language Processing services, or performing other tasks.
- Real-time Data Display: The Dashboard provides real-time data display, enabling users to access up-to-date information regarding their tasks, resources, and processing jobs.
- Alerts and Notifications: Users receive alerts and notifications through the Dashboard, which keep them informed about changes related to their activities on the CLARIN:EL platform.

The Dashboard provides the users with the following functionalities (Figure 4):

[14] https://clarin-platform-documentation.readthedocs.io/en/stable/all/3_Creating/publicationLifecycle.html#publicationlifecycle

- Resource Creation and Upload: Users can easily create and upload language resources directly from the dashboard.
- Processing Services: The Dashboard serves as a gateway for users to access and utilize the integrated NLP processing services.
- Activity History: Users can explore a comprehensive history of their activity within CLARIN:EL, including details on created resources, validation tasks, and processing jobs.
- Editable Profile: Users can customize their profiles and manage their personal information and preferences effortlessly.

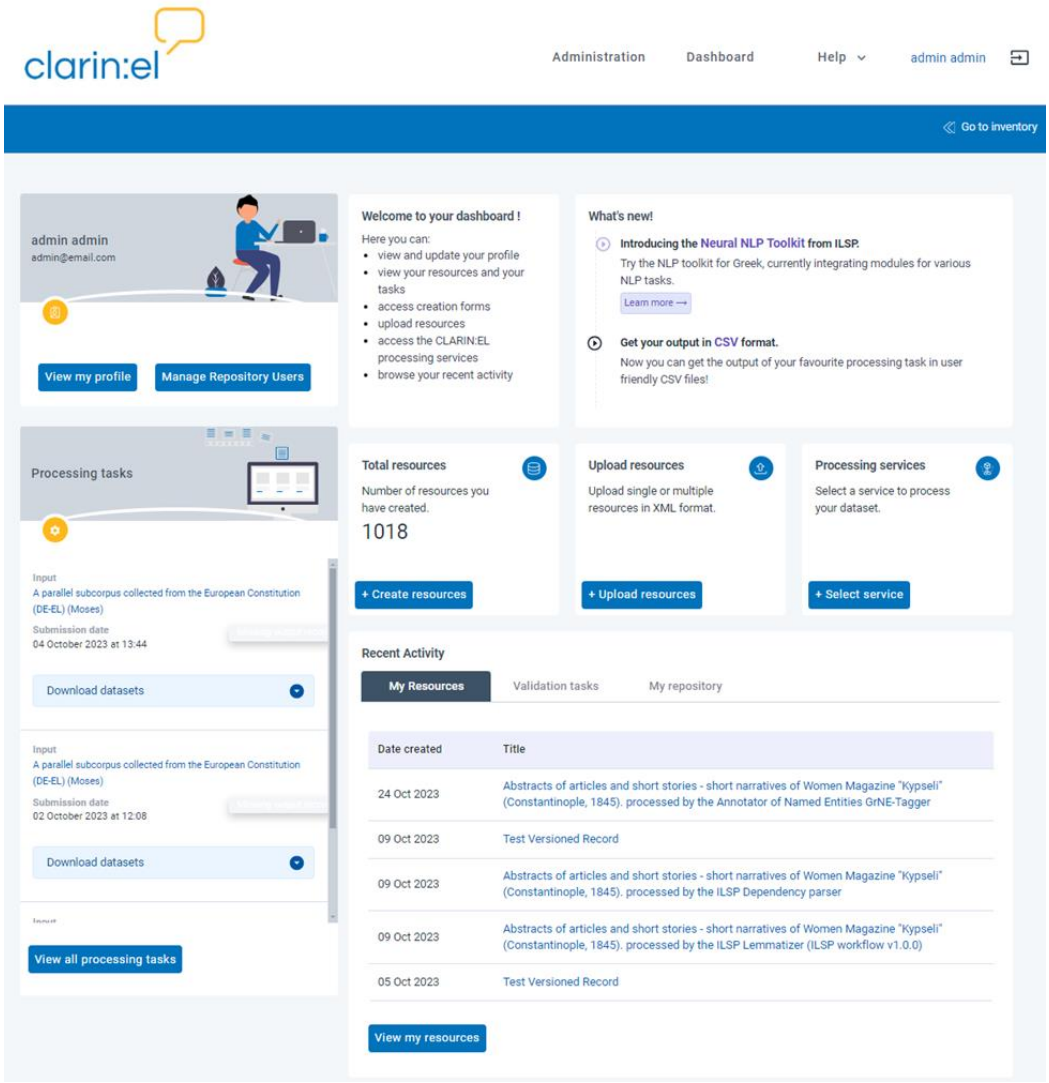


Figure 4: The User Dashboard

3 The CLARIN:EL technical architecture

The functionalities described in Section 2 are supported by the CLARIN:EL platform, designed and implemented with state-of-the art technologies. Its subsystems are built with robust, open-source, scalable technologies, and consist of several applications:

- the PostgreSQL database (DB) used for storing several types of data, such as user data, the metadata records of the LRs, etc.
- ElasticSearch for indexing
- the Repository Backend, built using the Django web framework, offers REST services for managing metadata (import, create, update, delete), authorizes access to the resources etc.

- the Repository, based on the META-SHARE software[15], with many improved architectural choices, new functionalities and features
- the User Interface that consists of web pages for searching/browsing the catalogue
- the Metadata Editor for creating/updating metadata, admin pages for validating resources etc.
- Keycloak, an identity and access management solution used for securing the applications
- the integrated NLP services
- a Workflow Manager responsible for executing NLP services and a scheduler that decides where and when a user's processing request will be executed, to avoid platform overloading
- and the User Dashboard.

All the above applications run as Docker containers at a Kubernetes (k8s) cluster, maintained and supported by the CLARIN:EL development team at ILSP/Athena RC. The LRs data are saved in a dedicated Network Attached Storage (NAS), while metadata are stored in PostgreSQL. CLARIN:EL uses Handle.net service to assign PIDs to resources, to ensure data accessibility. Procedures are in place for ensuring that hardware, software, and storage media containing archival copies of digital content are managed in accordance with security control, data protection and recovery standards. For example, an automatic procedure has been set up at CLARIN:EL's k8s cluster that regularly creates backups of the LRs and metadata records to two different servers, one on-site and one off-site. Also, a recovery procedure is in place, with a set of installation scripts and procedures that are to be used when a new instantiation of CLARIN:EL must be set up, e.g., in case of disaster. The CLARIN:EL infrastructure is protected and secured using standard practices, e.g., access to VMs and containers and k8s cluster is limited only to the administrators, a firewall is appropriately configured to restrict access to (specific) ports and VMs, SSL certificates are installed and renewed as appropriate.

4 User Support

4.1 The portal

CLARIN:EL provides several assistance mechanisms to support user needs. The Portal includes (i) information material on the infrastructure, the network and the team members working on the maintenance and operation of the Greek CLARIN, the use of the Platform and the provided services and functionalities, FAQs, as well as guidance on legal and policy issues related to data sharing, etc., (ii) dissemination material (news, events, articles and/or interviews about Language Technology), and (iii) educational and training material, namely video tutorials, scientific publications, and presentations. Publicly accessible Helpdesks enable interested parties to ask questions on technical and management issues, legal issues, or issues related to metadata creation and documentation of language data, tools and/or services. The Portal, as the main entry point of the infrastructure, besides hosting the informative material mentioned above, also provides links to direct the users to the Platform and the NLP:EL Knowledge Centre.

4.2 The NLP:EL Knowledge Centre

NLP:EL was established in March 2020 and has been since officially recognized as a CLARIN ERIC Knowledge Centre (K-Centre), while in May 2023 the Greek K-Centre has successfully received a certificate of recognition renewal. NLP:EL aims at actively supporting research and scientific advances and providing to all interested parties useful information and guidance in the fields of Natural Language Processing, Language Technology, Language Resources and Sign Language Technologies (SLT), as well as to support the digital readiness of the Greek language (Gavriilidou et al., 2023). The NLP:EL microsite (section) is organized in two main units: (a) *Knowledge*, where users can find access to an exhaustive list of tools and services for NLP and SLT, information on studies and curricula of Greek Universities related to LT and DH, educational and training material such as the *CLARIN Learning and Training Resources*[16] and the *SSH Training Discovery Toolkit*[17], a list of scientific publications (from

[15] https://github.com/metashare/META-SHARE
[16] https://www.clarin.eu/content/learning-and-training-resources
[17] https://training-toolkit.sshopencloud.eu/about

1973 to date) on NLP involving the Greek language, collected automatically from 7 databases[18], and (b) *Community*, where interested parties can be informed on NLP/LT and SLT teams active in Greece, on certified CLARIN K-Centres and on National and European LRTs Infrastructures.

4.3 Documentation

CLARIN:EL also provides detailed online documentation[19] on the Platform and all its functionalities, accessible through a "Help" button located on the home page. Through a dropdown menu, users can navigate to either the User Guide as a whole or directly to the section detailing the Recommended File Formats. This section is selectively promoted to aid users seeking information about CLARIN:EL policies regarding data and file formats during deposition. Additionally, users can utilize the search box to find particular information. The User Guide is available in both Greek and English; it is designed not for linear reading, though it can be approached that way, but rather aims to assist users in locating specific information according to their needs. It familiarizes users with the basic concepts of the infrastructure, guides them through its main functionalities (browsing, searching, viewing, downloading, and processing Language Resources), instructs them how to create and manage their resources, and explains the role and the significance of the metadata schema used for this purpose. Finally, it provides crucial information on legal issues connected to the publication, distribution, and use of language resources (licensing), as well as those connected to the use of the infrastructure itself (Privacy policy and Terms of Use). Each chapter within the User Guide is interconnected, and external links to referenced sources and documents are also included. The User Guide is available for download in PDF, HTML, and EPub formats.

4.4 Training activities

In addition to the management and the continuous updating of the material provided through the Portal and the NLP:EL Knowledge Centre, CLARIN:EL organizes a wide range of teaching/training, scientific and user support activities and events, such as webinars, workshops, hands-on sessions, summer schools, datathons, meetings with network members and CLARIN:EL users, etc. These activities are single or recurrent, aiming to educate users on LT and DH, to introduce the functionalities of the Platform and train interested parties how to use it.

4.5 The “Me, my family and other resources” initiative

The "Me, my family & other resources" initiative draws inspiration from the CLARIN ERIC Resource Families, while the title is a paraphrase of Gerald Durrell's renowned novel "My Family and Other Animals." Similarly to CLARIN Resource Families, this initiative organizes resources into thematic families, based on shared characteristics such as domain, topic, media type, time period, etc. These resource families are virtual collections of resources from various CLARIN:EL organizations, while each one maintains their individual autonomy and distinct traits. For each family, one resource is selected and highlighted as its representative; this resource is described in a dedicated portal webpage by a person involved in its creation process. Key metadata, including type, size, medium, provider, and format are also provided alongside a preview of the content of the resource. All the members of the family are also listed as hyperlinks directing to the CLARIN:EL central inventory. Each family is presented in conjunction with significant global, European, or national observances. For instance, the inaugural family, Poetry, was introduced in March 2022 to coincide with World Poetry Day on March 21st. To date, 13 resource families have been presented, namely Poetry, Medicine & Health, Museums, Tatoeba corpora, Sign Language resources, Educational Textbooks, Human Rights, Named Entity Recognition Tools, Kypseli Women's Magazines, Elections, Parliamentary Discourse, Literary Translation and Medieval and Early Modern Greek.

[18] ACL Anthology, ACM Digital Library, ArxiV, IEEE Xplore, ResearchGate, Semantic Scholar, and Springer. For ACL Anthology, papers were found on the official GitHub repository in XML format, whereas for the rest the software JabRef was used through its search interface.

[19] https://clarin-platform-documentation.readthedocs.io/en/stable/index.html

5 CLARIN:EL analytics

Matomo, formerly known as Piwik, is an open-source web analytics platform[20], which has been integrated on-premise into both the portal website (https://www.clarin.gr/) and inventory (https://inventory.clarin.gr/), providing detailed insights into visitor behaviour and site performance. With Matomo, website administrators gain valuable data to optimize content, improve user experience, and enhance overall website effectiveness. CLARIN:EL leverages Matomo Analytics to gather anonymized information about how visitors interact with the infrastructure. Using Matomo Analytics, CLARIN:EL tracks visitor interactions in real-time and over-time, offering a detailed overview of user engagement, including page views, visit duration, and bounce rates. This data helps understand how visitors navigate the website and which content attracts the most interest. Other valuable information include statistics on location and device used to access the infrastructure, as well as basic infrastructure performance statistics. Location statistics are evaluated based on the anonymized IP (e.g., 192.168.xxx.xxx). Anonymization in Matomo refers to the process of concealing or obscuring personally identifiable information (PII) from the data collected by the platform. This is crucial for privacy compliance, especially with regulations like GDPR in Europe.

In December 2023, the total number of CLARIN:EL registered users was 1,560 (+10,4% compared to December 2022). Based on Matomo Analytics, in 2023 the three pillars of the infrastructure (Portal, Platform, K-Centre) gathered a total of approximately 6,300 visits, 20,600 pageviews, 8,931 resource downloads and 3,843 processing tasks.

6 Conclusion

This paper presented the CLARIN:EL infrastructure, the design and implementation principles as reflected in its architecture, the functionalities available to the users, the support activities provided to the community, and finally, usage analytics. Future steps include the maintenance and upgrading of the infrastructure's modules, and the population of the repository with new resources (datasets and workflows). As regards dissemination and training, future objectives include the continuous support and training of CLARIN:EL users, and the roll-out of outreach activities aiming to raise awareness about LT in the research community. At the technical level, CLARIN:EL aims to ensure and increase interoperability with other infrastructures and repositories, at the national, European and international levels. At the strategy level, concrete aims constitute the enlargement of the network with new organization members and end-users, and ensuring its connectivity with EOSC, the SSHOC Marketplace, the Language Data Space and similar emerging initiatives.

Acknowledgements

This work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.), under the Emblematic Action "The emerging landscape of digital work in Humanities in the context of the European infrastructures DARIAH and CLARIN" (Project Number: 7982), https://digital-landscape.gr/.

References

Branco, A., Eskevich, M., Frontini, F., et al. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. Lang Resources & Evaluation. https://doi.org/10.1007/s10579-023-09658-z

Gavriilidou, M., et al. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 1090-1097, http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.pdf

[20] https://matomo.org/

Gavriilidou, M., Giagkou, M., Loizidou, D., & Piperidis, S. (2023). Language Report Greek. In: Rehm, G., Way, A. (eds) European Language Equality. Cognitive Technologies. Springer, Cham. https://doi.org/10.1007/978-3-031-28819-7_19

Khan, A.F., et al. (2022). When linguistics meets web technologies. Recent advances in modelling linguistic linked data. Semantic Web Preprint (2022): 1-64. https://www.semantic-web-journal.net/content/when-linguistics-meets-web-technologies-recent-advances-modelling-linguistic-linked-open

Labropoulou, P., Gkirtzou, K., Gavriilidou, M., Deligiannis, M., Galanis, D., Piperidis, S., Rehm, G., Berger, M., Mapelli, V., Rigault, M., Arranz, V., Choukri, K., Backfried, G. Gómez Pérez, J.M., & Garcia Silva, A. (2020). Making metadata fit for next generation language technology platforms: The metadata schema of the European Language Grid. arXiv preprint arXiv:2003.13236

McCrae, J.P., et al (2015). One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web. *The Semantic Web: ESWC 2015 Satellite Events*. ESWC 2015. Lecture Notes in Computer Science, vol 9341. Springer, Cham. https://doi.org/10.1007/978-3-319-25639-9_42

Piperidis, S., Galanis, D., Bakagianni, J., & Sofianopoulos, S. (2016). Combining and Extending Data Infrastructures with Linguistic Annotation Services. In: Murakami, Y., Lin, D. (eds) Worldwide Language Service Infrastructure. WLSI 2015. Lecture Notes in Computer Science(), vol 9442. Springer, Cham. https://doi.org/10.1007/978-3-319-31468-6_1

Piperidis, S., Labropoulou, P., Deligiannis, & M., Giagkou, M. (2018). Managing public sector data for multilingual applications development. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018*), Miyazaki, Japan. European Language Resources Association (ELRA). https://aclanthology.org/L18-1205.pdf

Pontiki, M., Gavriilidou, M., Gkoumas, D., & Piperidis, S. 2020. Verbal Aggression as an Indicator of Xenophobic Attitudes in Greek Twitter during and after the Financial Crisis. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, pages 19–26, Marseille, France. European Language Resources Association.

Prokopidis, P., & Piperidis, S. (2020). A Neural NLP toolkit for Greek. In *11th Hellenic Conference on Artificial Intelligence* (pp. 125-128).

Wilkinson, M., Dumontier, M., Aalbersberg, I., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18

The LiRI Corpus Platform

Johannes Graën, Jonathan Schaber, Daniel McDonald, Igor Mustač, Nikolina Rajović, Gerold Schneider, Teodora Vuković, Jeremy Zehr and Noah Bubenhofer
Linguistic Research Infrastructure
University of Zurich, Switzerland
first_name.last_name@linguistik.uzh.ch

Abstract

We present the LiRI Corpus Platform (LCP), a software system and infrastructure for querying a vast array of corpora of different kinds. It heavily relies on the PostgreSQL relational database management system, employing state-of-the-art data representation and indexing techniques, which lead to significant performance gains when querying, even for structurally complex queries involving nested logical operations and quantifiers. In this work, we describe the requirements that led to the development of this novel system, discuss methods from corpus linguistics and beyond that we considered key for such a system, and provide details on a number of technological features that we take advantage of. Our platform also comes with its own query language tailored both to the requirements in terms of information need and our philosophy of how to define corpora in an abstract way.

1 Introduction

Corpora are an important resource for empirical linguistic research, as well as text-based social studies such as digital humanities or media research (Meurers, 2005). There exist numerous tools and platforms for manipulating different corpus types (mono- vs. multilingual, text vs. multimodal, synchronic vs. diachronic, etc.), offering various means of access to corpora (web interfaces, command line interfaces, etc.), and providing different functionalities (editing, querying, annotating, etc.).[1] While, taken together, the set of tools and applications developed thus far covers a variety of corpora and corpus types, taken individually, most of them are actually designed for specific corpora or corpus types, and therefore lack the ability to generalize. Some tools provide a general interface for diverse corpora, handling a multitude of differently structured corpora and occasionally provide specific functionality needed for them (see Section 2 for a (small) overview of existing tools and applications).

Central to all corpus tools is the ability to query one or more datasets. We propose a distinction between two main querying strategies:

1. **Corpus exploration**, characterized by being an iterative, drilling-down process, where users run and refine their queries based on results, and by interfaces that prioritize speed (Hearst, 1999).

2. **Corpus analysis**, maximizing recall of results in a corpus, prioritizing accuracy and specificity, with results returned in a format suited for subsequent statistical analyses.

The latter strategy is of particular interest to approaches that require large amounts of data ("Big Data"), such as the training of large language models.

While those two strategies are more complementary than antagonistic, to our knowledge most tools tend to prioritize the latter strategy over the former (cf. Desagulier, 2019). We design our application to accommodate both approaches, supporting the construction of queries in a bottom-up fashion. When working with today's often very large corpora, one does not always need full recall: as soon as the matches delivered are a reasonable approximation to all the results, the user may choose to stop instead of waiting for complete results. We discuss, besides other things, how we employ random sampling by

[1] https://corpus-analysis.com/, a page collecting tools for corpus linguistics, lists 266 entries (as of 2024-03-05).

Johannes Graën, Jonathan Schaber, Daniel McDonald, Igor Mustač, Nikolina Rajović, Gerold Schneider, Teodora Vuković, Jeremy Zehr and Noah Bubenhofer 2023. The LiRI Corpus Platform. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 62–75. DOI: https://doi.org/10.3384/ecp210

means of logarithmic partitioning to achieve these goals (in Sections 3.2 and 4.2). The incorporation of both strategies constitutes a timely improvement over existing tools that presume their users to arrive with full-fledged, complex queries.

In this contribution, we present the LiRI Corpus Platform (LCP) a new tool that couples complex analysis of diverse corpora and interoperability with existing CLARIN resources. The software facilitates access to and re-use of research data, offering a flexible architecture, whereby a single backend can be connected to multiple frontends tailored to the requirements of specific research agendas. We will mostly focus on the data modeling and its implementation in an RDBMS, while only briefly explaining the different interfaces that we have implemented and plan to implement in the future. While interfaces are to some extent interchangeable, the dynamic data model and its implementation form a core feature and contribution of our tool.

After discussing existing corpus query systems in the next section, we will detail in Section 3 which methods from corpus linguistics were seminal to our platform, followed by an overview of technology and technological methods employed in Section 4. Section 5 lists important limitations of our platform's features, describes the current state of implementation and provides an outlook on future development.

2 Related Work

Myriad systems for storing and querying large corpora have been developed in past decades. Clematide (2015) gives an overview of corpus linguistic query language types, distinguishing between four families. Historically prominent were (I) sequence-based designs, such as CQP (Christ, 1994; Evert & Hardie, 2011) and other dialects of regular expressions, (II) structure-based designs, such as TGrep2 (Rohde, 2005) for querying syntactic trees. Many of them have organically developed within parameters set by technological restrictions of the time. For example, CQP is limited with regard to syntactic queries because its sequence-based conceptualization of text is ill-suited to express non-sequential structural information expect for containment expressed by the *within* keyword. More recently (III), the class of path-based languages which use the XPath query language have been implemented. Finally, (IV) the class of logic-based languages such as TigerSearch and ANNIS, which offers outstanding expressiveness, coupled with considerably longer retrieval times. For example, for performance reasons, the most recent major version of ANNIS (Krause & Zeldes, 2016) has abandoned the relational database PostgreSQL, and developed a custom implementation based on graphs.[2] In contrast, our proposed approach retains the expressiveness of logic-based languages while leveraging advanced data-representation and indexing techniques in order to offer faster retrieval times.

Several methods for speed-ups have been proposed. For instance, while older approaches rely on MapReduce techniques (Schneider, 2013), modern database management systems as PostgreSQL provide internal mechanisms for parallel computing and intelligent algorithms to use the most discriminate features first. Other methods are the use of sophisticated indexing and retrieval techniques, as in the the proposals of (Ghodke & Bird, 2012).

For reasons of space, our overview provides only a brief sketch of past and present resources. For example, we omit detailed description of several projects fine-tuned for particular corpora and particular tasks, for example Dependency Bank (Lehmann & Schneider, 2012), which allows fast syntactic queries on the British National Corpus (BNC).

While some published standards and guidelines have attempted to define a single digital format able to encompass all possible linguistic data and annotations (Gries & Berez, 2017; Ide & Romary, 2004, 2006), to our knowledge no actual working implementation of such proposals has been presented to date. While we do not suggest that our solution covers all possible needs, we offer an attractive solution that facilitates analysis of a very broad range of corpus types.

3 Methods

The methods described here were fundamental to the development of our corpus platform. Our target user groups come from different fields and are accustomed to different tools and methods of analyzing

[2] https://github.com/korpling/ANNIS/blob/main/CHANGELOG.md (2024-03-05).

language data. We aim at satisfying needs of most envisaged users, but we are well aware that a one-size-fits-all solution for corpus management and retrieval is unlikely to be developed in the near future.

3.1 Analysis and Collocations

In addition to showing matches in plain and keyword-in-contect (KWIC) format, LCP allows for several statistical analyses. We present two examples of statistical analyses: frequency analysis, which delivers frequency distributions, and collocations, in which a large range of standard collocation measures are offered. The query in Figure 1 illustrates those on a query for adjective premodification of a noun whose lemma is *preference*.

```
Segment s

sequence seq
    Token@s t1
        xpos2 = "ADJ"
    Token@s t2
        xpos2 = "SUBST"
        lemma = "preference"

KWIC => plain
    context
        s
    entities
        t1
        t2
```

```
AdjDist => analysis
    attributes
        t1.lemma
    functions
        frequency
    filter
        frequency > 1

LeftContext => collocation
    center
        t2
    window
        -2..-1
    attribute
        lemma
```

Figure 1: A query on the British National Corpus (BNC) searching for a sequence of adjectives and nouns with the lemma *preference*. Alongside a result set comprising both tokens within the sentence in which they appear ('KWIC'), a frequency analysis on the lemma of the adjectives ('AdjDist') and a collocational analysis ('LeftContext') are defined – for a window of the two tokens preceding the noun (positions -2 and -1).

The results pane shows three tabs, corresponding to the three result sets *KWIC*, *AdjDist* and *LeftContext* that we defined in the query above. Figure 2 reports the frequency table (left) and collocation table (right) generated by the query: the former lists absolute and relative frequencies, while the latter reports strongest associates, sorted by *simple log-likelihood* in this example.

3.2 Sampling

One strategy we deploy allowing us to deliver results very quickly is the use of random sampling combined with the incremental querying of increasingly large samples of documents, starting with the smallest sample, while constantly updating the information displayed, that is raw matches, frequencies and statistical measures.[3] Random sampling is typically the best sampling strategy if no assumptions can be made about the entire dataset (Cochran, 1977). The reliability of the partial results can be estimated by the standard error SE, which is defined as

$$\mathrm{SE} = \bar{\sigma} = \frac{\sigma}{\sqrt{n}} \tag{1}$$

where n is the number of matchses, and σ the standard deviation, i.e. the fluctuation in a sample.

While we do not know the exact standard deviation until we have found all the matches (for we cannot know the mean of the whole population) we can approximate it, as is e.g. done in the Welch's two-sample t-test, where the standard error is approximated via the samples.

[3] See also Section 4.2.

KWIC (521) | AdjDist (116) | LeftContext (426)

Filter by t1.lemma | Filter by frequency | Filter by relative freque

t1.lemma	frequency	relative frequency ↓
personal	74	8.159 %
first	62	6.836 %
strong	43	4.741 %
own	43	4.741 %
this	30	3.308 %
local	30	3.308 %
sexual	27	2.977 %
second	27	2.977 %
convertible	26	2.867 %
parental	26	2.867 %
individual	21	2.315 %
political	20	2.205 %
mating	19	2.095 %
clear	17	1.874 %

KWIC (521) | AdjDist (116) | LeftContext (426)

Filter by Text | Filter t | Filter by E | Filter by O/E | Filter by M | Filter by M | Filter by lo | Filter by | Filter by z-sc | Filter by sin

Text	O	E	O/E	MI	MI³	local-MI	t-score	z-score	simple-ll ↓
personal	75	0.109	688.971	9.428	21.886	707.123	8.648	226.987	830.498
convertible	31	0.003	10053.296	13.295	23.204	412.157	5.567	558.203	509.377
strong	48	0.125	383.64	8.584	19.754	412.013	6.91	135.347	475.422
first	64	0.759	84.343	6.398	18.398	409.484	7.905	72.6	441.183
own	50	0.462	108.337	6.759	18.047	337.969	7.006	72.92	369.448
parental	26	0.008	3132.988	11.613	21.014	301.946	5.097	285.317	366.603
redeemable	19	5.78e-4	32853.672	15.004	23.5	285.072	4.359	790.052	357.194
a	120	13.612	8.816	3.14	16.954	376.813	9.712	28.836	309.598
mating	19	0.002	7936.008	12.954	21.45	246.13	4.358	388.26	303.213
sexual	27	0.044	615.614	9.266	18.776	250.179	5.188	128.715	292.909
local	30	0.294	102.025	6.673	16.487	200.183	5.424	54.782	218.101
marked	16	0.013	1238.665	10.275	18.275	164.393	3.997	140.665	195.923
second	27	0.285	94.813	6.567	16.077	177.309	5.141	50.062	192.372
imperial	15	0.012	1268.548	10.309	18.123	154.634	3.87	137.834	184.393

Figure 2: Distribution of adjectives premodifying the noun *preference* (left) and collocations of the two preceding tokens of that noun (right).

$$t = \frac{\bar{x} - \mu}{\bar{\sigma}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{(\bar{x_1} - \mu) + (\bar{x_2} - \mu)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}} = \frac{(\bar{x_1} - \bar{x_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2}$$

Accordingly, the frequencies found in a large set of samples give us a good estimate of the population frequency, to which it asymptotically converges. Our technique of logarithmic partitioning implements this random sampling strategy for faster callback while minimally compromising on accuracy (see Section 4.2 for more details).

3.3 Syntactic Queries

Corpora that model dependency relations can be queried for syntactic relations. For example, the corpus in Figure 3 defines an entity named *DepRel* which models syntactic dependency relations. The illustrated query delivers all pairs of tokens that belong to the same sentence *Segment*, where the head is a form of the verb 'take' and the dependent is its object (the labelset uses 'dobj' for direct objects).

See also Section 4.4 for representation and implementation details regarding dependency structures.

3.4 Query Language

The CQP query language (Christ, 1994) ingeniously translated the concept of regular expressions into corpus queries, supporting any pattern of adjacent tokens with attributes that can be modeled as automata. Its widespread use attests to its versatility. Other query languages have extended CQP's potential by providing for syntactic query operators. The ANNIS query language AQL (Zeldes et al., 2009), in particular, provides a multitude of operators for both constituency and dependency syntax.[4]

Clematide (2015) classifies several query languages with regard to their purpose and functionality. In general, a trade-off between efficiency and complexity (of both the data and the query language) can be observed. However, to determine this trade-off in a sound way, or even comparing different query languages regarding those effects is very hard as pointed out by Lai and Bird (2010, p. 59):

> The problem of comparison [of query languages] is exacerbated by the fact that most implementations are tailored to the flat file representation used by a specific corpus, and cannot

[4]Fundamental syntactic queries had already been established by its predecessor TIGERSearch (König & Lezius, 2000).

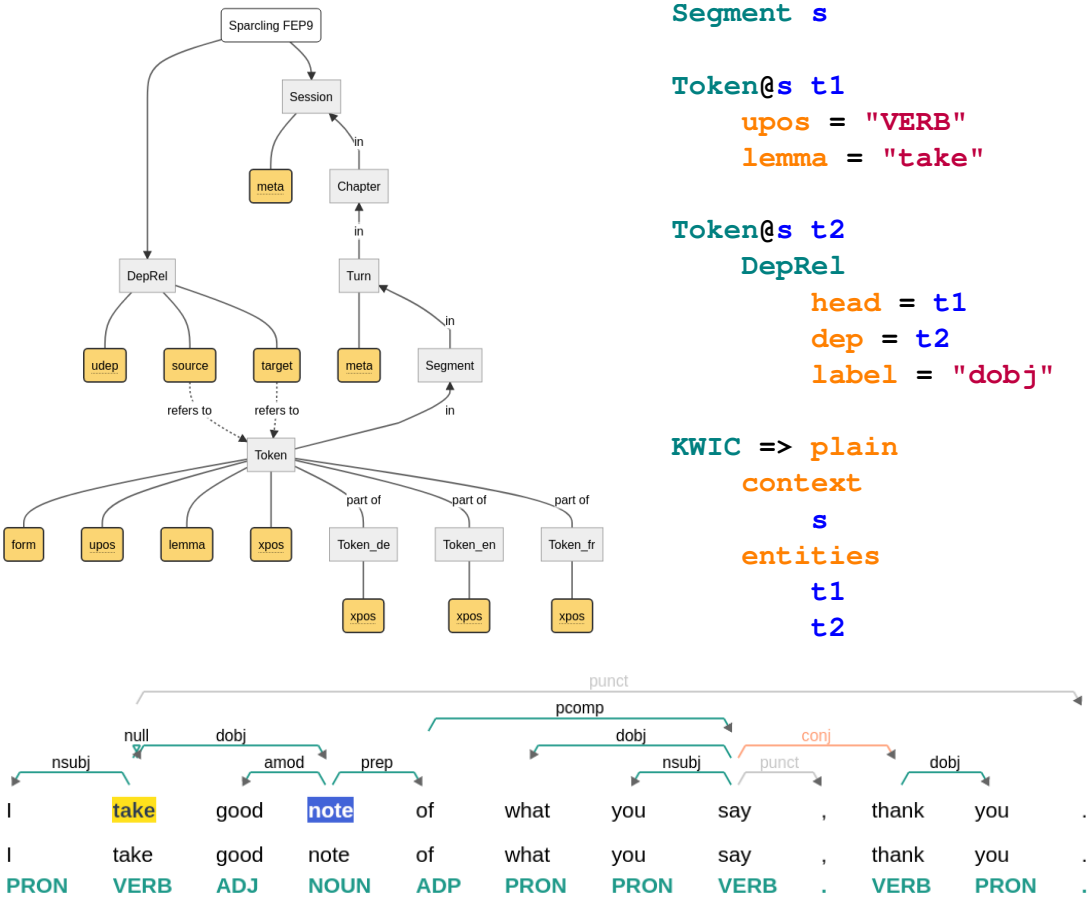


Figure 3: The automatically generated visualization of the corpus structure (left), a syntactic query for a simple dependency relation to be run against the said corpus (right) and an example sentence from the 'KWIC' result set (bottom).

> be compared directly. Comparing code is hampered by the fact that the code is not always available, and because query processing is intertwined with idiosyncratic indexing and storage. In general, these languages are not compiled into an existing general purpose language (such as SQL), which means that their relationship to such languages is not known. Moreover, standard indexing and optimisation techniques cannot be applied, and the implementations we have experimented with do not scale [...]. For these reasons, it is difficult to establish the formal expressiveness of existing linguistic tree query languages, or establish the asymptotic efficiency of their implementations.

For our query language DQD (Descriptive Query Definition), a key concept was the Entity–relationship (ER) model used for modeling data structures in software development, especially for database schemata. For corpora hosted by our infrastructure, the structure needs to be defined in terms of entities (including a list of their attributes and attribute types), and relations between them; we refer to this structure as *corpus template*. In contrast to free modeling in ER diagrams, we limit data and relation types to a predefined set, and require corpus developers to define three fundamental units:

1. the main unit of interest (e.g. tokens),
2. meaningful ordered collections of those units (e.g. sentences), and
3. the hierarchically highest unit that contains ordered elements (e.g. documents)

The query language DQD then refers to a particular constellation of entities and their relations (including non-existing parts). We can put constraints on entities (existing and non-existing ones) in form of logical formulae represented as trees. DQD uses indention to express subordination.[5]

In absence of root-level logical operators, we assume conjunction (as used for conditions on *Token 't2'* in Figure 1). Similarly, we assume existential quantification if existence is not explicitly negated (for more details on quantification, see the following section). [6] A DQD query introducing entities e_1 to e_n with their subordinated propositions p_1 to p_n is thus matching everything where $\bigwedge_i \exists e_i \wedge p_i(e_i)$ holds.

DQD uses three set-defining operators:

1. **sequence** for patterns on units with optional repetition ranges similarly to the ones introduced by the CQP query language,
2. **group** for conflating elements or sets of elements of the same type into a single set, and
3. **set** for a – potentially empty – set of units of the same type defined by a logical proposition (see *tverbs* in Figure 4)

```
Segment s

Token@s tverb
    upos = "VERB"

set tdeps
    Token@s tx
        DepRel
            head = tverb
            dep = tx
            label != "punct"
```

```
KWIC => plain
    context
        s
    entities
        tverb
        tdeps
```

Figure 4: A query searching for verbs and all their dependencies with a label different to "punct". The results set returns verbs and dependencies within the context of sentence *Segments*.

Besides the definition of such a constellation, DQD also defines one or many datasets to be returned by the query. We currently support three types, namely:

1. a **plain** view on the data within a defined **context** (e.g. sentence or document) and the entities in this context to be marked (e.g. tokens or sentences),
2. a statistical **analysis** of the results on one or more **attributes**, using one or more **functions** (e.g. frequency), applying an optional **filter**,[7] and
3. a **collocation** analysis on a defined **attribute**, with either a **center** unit (token) and a **window**, or a **space**, which is a set of tokens[8]

A list of operators are defined to operate on numbers and number ranges (including times and time ranges), strings and entities. Alignment operators can be employed for defining correspondence on parallel corpora.[9]

[5] DQD comes in a textual and a corresponding JSON format; for reasons of simplicity, we will only refer to the textual representation when discussing DQD features.

[6] By convention, we use lowercase for attributes (lemma, upos, label, ...) and camel case for entities (Token, DepRel, ...). Logical operators and quantifiers are required to be written with uppercase letters (NOT, AND, EXISTS, ...).

[7] The filtering happens after the aggregates have been calculated

[8] The latter is useful for collocation analysis in dependency relations.

[9] Along the lines of "hierarchical alignment" described in (Graën, 2018, p. 76ff).

3.5 Existential Constraints

An important aspect of query languages targeted at linguistic corpora concerns the possibility to express constraints on the *absence* of data. Consider, for example, the pseudo-query "find all sentences that have no verb", which requires the query language to have some means to express both universal quantification and negation on the level of quantification. This query can be expressed in relational calculus as $\{\langle s\rangle|\neg\exists t \in s \wedge \text{verb}(t)\}$ or, universally quantified, as $\{\langle s\rangle|\forall t \in s \rightarrow \neg\text{verb}(t)\}$.

```
Segment s

Token@s t1
    upos = "NOUN"

¬EXISTS
    Token@s t2
        upos = "VERB"
```

```
KWIC => plain
    context
        s
    entities
        t1
```

Figure 5: A query searching for sentence *Segments* with at least one noun but no verb.

While existentially quantified entities have a global scope in DQD and can be referred to from anywhere after their declaration in DQD, universally quantified entities cannot as their existence is bound to the scope of their quantification. The *Token* 't2' in Figure 5, could thus only be referred to within the scope of the '¬EXISTS' quantification. For the result set 'KWIC', 't1' is thus the only visible entity.

As universal and existential quantification can be translated to one another by negating both the quantification and the logical formula under the quantification's scope, we only support negated existential quantification in DQD – in addition to regular existential quantification.[10] Given that a query needs to declare at least one result set for it to return anything at all, and that result sets need to make use of at least one entity, all DQD queries consequently have at least one entity that is existentially quantified.[11]

Lai and Bird, 2010 demonstrate that while TGrep2 only enforces an "(implicit) outermost quantification [that] is always existential" and allows other forms of quantification on lower scoped variables while inhibiting "negation outside an entire expression", in TIGERSearch "all variables are existentially quantified" and the language further inhibits "negation to scope over this implicit existential".

4 Technology

We offer a flexible solution that we have designed for several types of corpora, ranging from Digital Editions (where a reference to an area in the original manuscript is essential), multimodal corpora (where movies and temporal annotation are required), and very large corpora (where short retrieval times are a priority, even for complex corpus templates).[12]

The corpora that we make available through the system differ in various aspects such as size, annotation layers, and complexity. While structurally complex corpora pose a challenge for the construction and processing of a dynamic query language, very large corpora, on the other hand, demand the efficient retrieval of previously unseen queries. Because of the latter, we cannot precompute partial results, but we can still precompute efficient data representations and index structures.

We accommodate structurally complex corpora, and in particular parallel corpora, by representing alignments as structural elements, similar to dependency relations.[13] We employ PostgreSQL, a modern database management system,[14] to deal with the latter challenge, unlike the developers of ANNIS

[10] Recall from Section 3.4 that all entities are existentially quantified if not declared otherwise.

[11] Even if a result set refers to a set instead of a singleton and all instances of said set are empty, it still exists.

[12] See Section 3.4.

[13] There are three main differences between these two relations: 1) alignments represent correspondence and are thus undirected, 2) alignment links are typically unlabeled, and 3) the data type of alignments are (nested) sets rather than trees (see Graën, 2018).

[14] https://www.postgresql.org/

(Krause & Zeldes, 2016), who replaced PostgreSQL as a storage and query backend with their own solution.[15]

The description of the methods we employ focuses on the central task of querying. Our methods have been designed to scale to very large corpora with complex annotation schemes. To this end, we have designed a query language named DQD (Descriptive Query Definition), which has two interconvertible representations, one in JSON and one in a textual format (see also Section 3.4). We can translate queries from other query languages like the CQP query language to DQD, but not the other way around, since there are queries which can be represented in DQD, but not in other languages (e.g. queries using syntactic relations or alignment structures).

Due to the complexity of its architecture (see the following section) and the intended use as a platform for making linguistic data available e.g. through the CLARIN-network, LCP is implemented as a service that can be accessed over the Internet; either through one of several dedicated web applications or directly via an API.[16] Both access methods require matching credentials.

In this section, we limit ourselves to central techniques that we leverage as part of our platform's results retrieval process. We describe further improvements not explained here in Section 5.

4.1 Platform Architecture

LCP is designed in a modular way; individual components can be replaced and layers be extended, allowing us to scale horizontally, address future needs when they arise, and offer tailored solutions to individual user groups. In particular, new frontend applications can easily be integrated and an increasing number of concurrent users can be dealt with by scaling up the number of workers and replicas in the database cluster.

Figure 6 shows the architecture of LCP. The entry point for natural users and automated access is the reverse proxy, which connects to Shibboleth for authentication if authentication is needed.[17] Requests are then forwarded to the individual applications.[18] Media files (audio, video and images) are sent directly to the client, provided approval by LAMa, the LiRI Account Manager, which handles users, groups and permissions.

The backend for all applications consists of an application that manages sessions and routes different types of queries to either regular workers or background workers, based on priorities. Regular workers perform incremental queries as described in Section 3.2 using Redis[19] for aggregating data retrieved from the database. Background workers, on the other hand, perform more time-consuming tasks such as exporting large datasets and converting them into different file formats.

Connections to the database run through a pgpool II instance that acts as both, connection pooler and load balancer, in front of the actual database cluster, which consists of a main server and several replicas, for reasons of performance and availability.

4.2 Logarithmic Partitioning

LCP places no inherent restrictions on corpus size.[20] Running a complex query on a multi-billion word corpus can potentially require hours if data is to be processed sequentially, prohibiting a seamless iterative querying approach, despite this being a typical workflow in corpus-driven linguistic research (Rayson et al., 2017). To quickly retrieve initial results, and thus enable researchers to refine their query and estimate

[15] "Instead of using the relational database PostgreSQL, a custom AQL implementation based on graphs called graphANNIS is used." (https://github.com/korpling/ANNIS/blob/main/CHANGELOG.md (2024-03-05).

[16] We provide a general overview of corpora and tools available at https://lcp.linguistik.uzh.ch.

[17] Automated access uses project-based access keys instead.

[18] We currently provide an application for querying text corpora named *catchphrase*. Another application for speech corpora is scheduled to be released in March 2024, and another one for multimodal corpora in May 2024.

[19] https://redis.io/

[20] This is one of the advantages of LCP over other systems like e.g. the Corpus Workbench, whose initial architecture restricts it to a maximum corpus size: "[. . .] the internal format of the indexed and compressed corpora imposes [a] 2.1 billion token limit" (Evert & Hardie, 2011, p. 13). This shortcoming is to be addressed by a new data representation format named Ziggurat (Evert & Hardie, 2015), which has not been released at this point. The authors have also witnessed the performance of ANNIS (version 3.x) dropping considerably for medium-size datasets (less than 10m tokens).

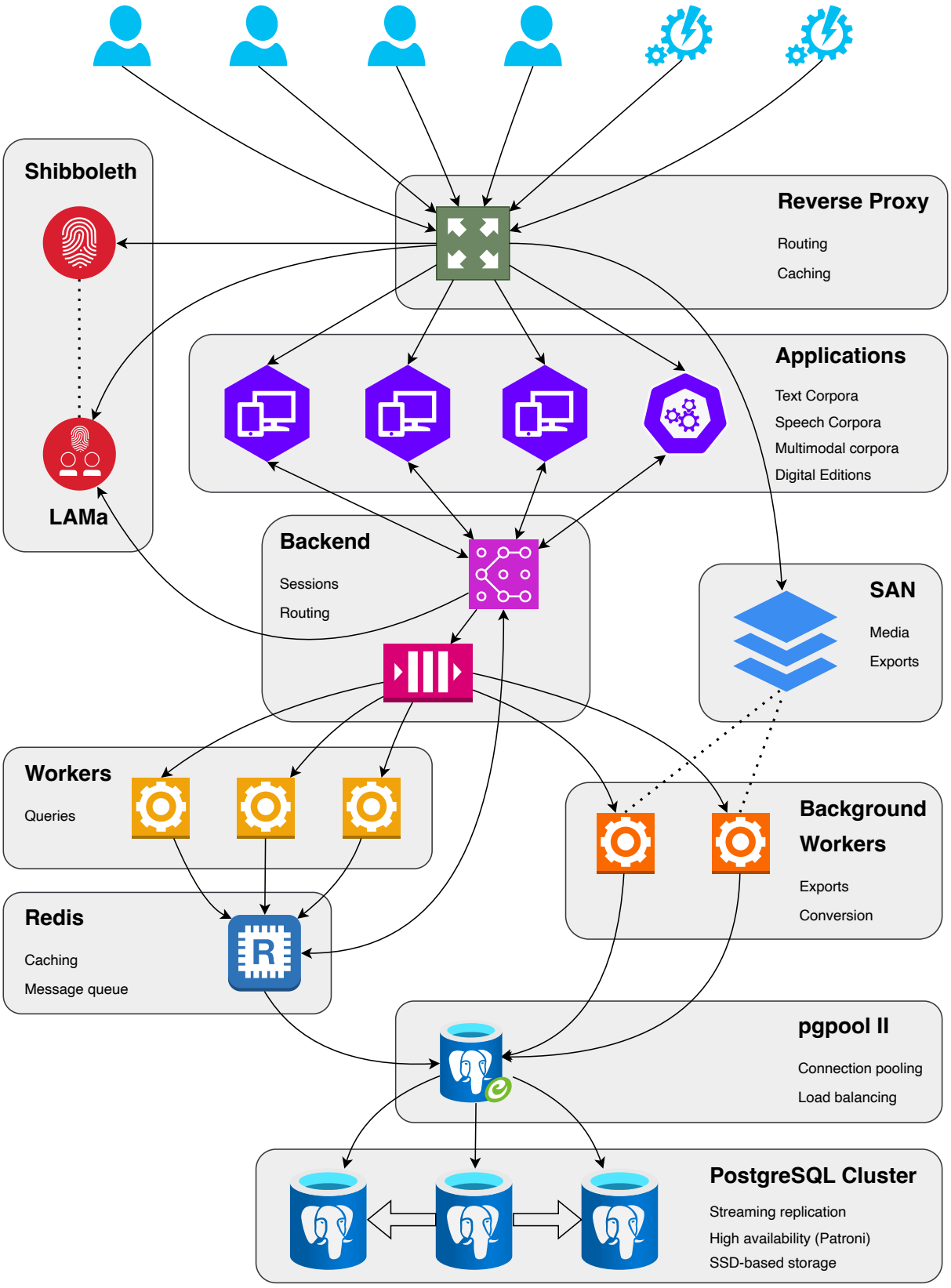


Figure 6: Architecture of LCP showing its modular design. Any number of frontend applications (purple) use a common backend, which, in turn, makes use of a variable number of workers that handle short-term or long-term requests. The database cluster can easily be extended both vertically (more disk space) and horizontally (more nodes).

the distribution of the phenomenon in question on the whole dataset, we employ a structure that allows incremental querying randomly sampled subcorpora of increasing size.

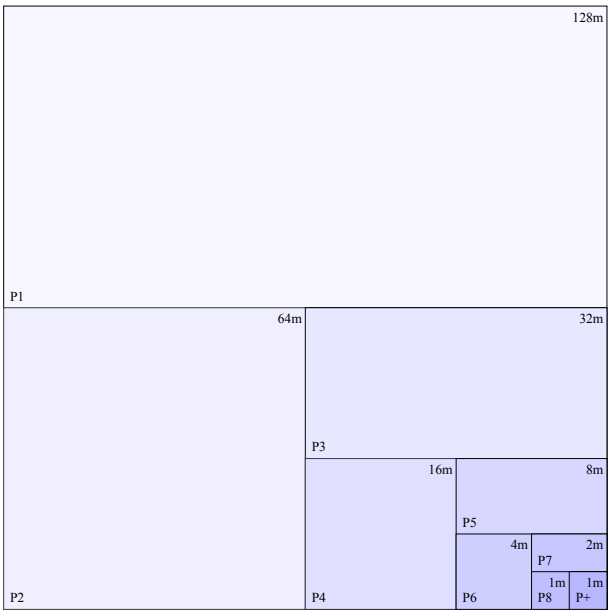


Figure 7: A corpus with 256 million segments is split into nine partitions. The last one, P+, is of approximately equal size to the second-to-last one, P8, as it contains all segments that remain after the eight-fold bisection.

To this end, we generate Universally Unique Identifiers (UUIDs) (Leach et al., 2005) for the linguistically salient units in each corpus – typically, sentences, sentence segments or utterances. We then subdivide the total space of potential UUIDs into partitions of decreasing size, always splitting one part into two halves (see Figure 7). We use Version 4 UUIDs, as they are (for the largest part) generated by a truly random algorithm,[21] so we can expect each partition to comprise a known share of the total number of units (granted a small distributional deviation that depends on the corpus size), that is, in relation to the total size, a half, a fourth, an eighth etc.

In the case of a one billion word corpus with an estimated average of ten words per sentence and a smallest partition of at least one million tokens, we would thus create eleven partitions.[22]

This logarithmic partitioning allows us to run a query on the smallest partition and extrapolating from the first result set retrieved both to the expected amount of results on the whole corpus and to the optimal partition to be queried next in order to satisfy the request for a particular number of results.

For the investigation of frequent phenomena, a small random sample is often enough. If more data is needed, our approach seamlessly scales from a quick 'pilot study' on a subset to a complete analysis of the entire corpus. Such pilot studies also allow researchers to develop and test their queries quickly, and to assess where interesting differences by the available metadata can be found (period, variety, genre, gender, etc.) in an exploratory fashion.

4.3 Indexable Vector Representations

Corpora added into our database are passed through a pipeline that computes a vector representation of each sentence, preserving the positional information of tokens; in PostgreSQL this data structure is implemented under the name of "tsvector" (Bartunov & Sigaev, 2001). This is related to the classic information retrieval task of phrase searches (Manning, 2009). In most text corpora, tokens show various layers of annotation (e.g. lemmas, part-of-speech tags, morphological features etc.). With the vector representation implemented with the help of that data structure, we can define multiple layers of infor-

[21] We actually only require the first N bits to be random, where N depends on the overall size of the corpus.

[22] 1b words / 1m words = 1000 $\approx 2^{10}$ (ten regular partitions plus the remainder).

mation per position (in a sentence). In order to take them apart when querying, we need to distinguish word forms from lemmas, tagsets and so on. To this end, we prepend each string with one character per layer to tell them apart.

The advantage of such a vector structure is that it can be efficiently indexed[23] and thus allows performant retrieval of sentences that match specific criteria. Typical search patterns like sequences of tokens with additional constraints on annotations, like a CQP-style `[pos="DET"] [pos="ADJ"] [lemma="linguist"]` can be converted to a vector query that makes use of this index and quickly finds matching sentences. This often allows us to drastically reduce the number of sentences to which the original query will be applied, hence it is used to prefilter the corpus.

4.4 Nested Sets

Syntactic analyses in the form of dependencies form a tree-structure – in the mathematical sense of the word – over sentences (Crystal & Alan, 2023, p. 137). Relational Databases do not provide a straightforward way how to represent and, especially, query such data. However, several solutions for this problem have been proposed, one of them being a representation of tree-structures in the form of nested sets, which we also chose to employ for LCP. Nested sets can conceptually be conceived as sets that are completely contained within each other. The nested set representation of a tree can be computed by traversing the tree depth-first and attaching a number when a node is passed the first and when it is passed the second time (called the left and right anchor, respectively). Figure 4.4 shows the result of this process for an example sentence.

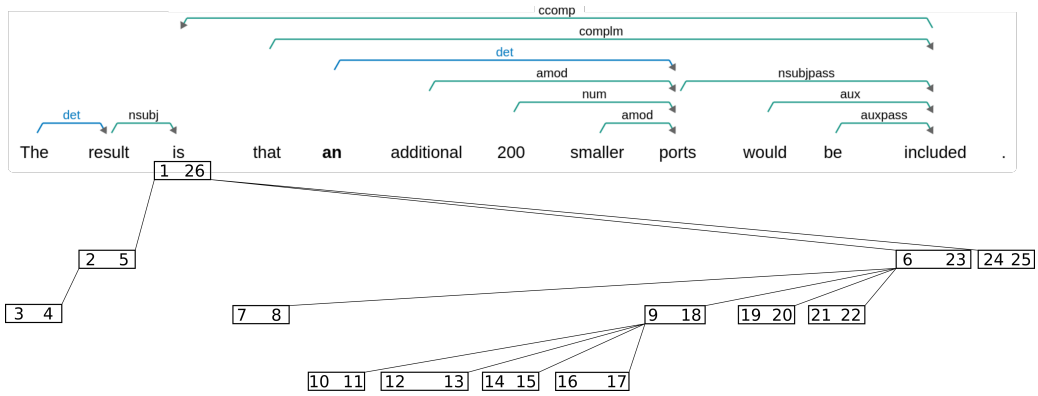


Figure 8: The dependency tree and its corresponding pairs of anchors in the nested set representation. The third word “is” is the root of the tree, and the traversal start and ends there.

As demonstrated by e.g. Celko, 2012, p. 49ff. nested sets have some useful properties: They allow, among others, e.g. (a) simple selection of subtrees, (b) the efficient computation of the height of a (sub-)tree, (c) finding paths, by restricting selections on the anchors. In other, possibly more intuitive representations (such as adjacency list, where each node points to its parent), this could only be achieved by writing more complex and computationally more expensive queries.

4.5 Linear Projection of Hierarchical Structure

Different corpora tend to be quite heterogeneous regarding the super-segmental structures that the use. They can potentially range from flat structures (e.g. a corpus consisting of recorded interviews, where no further subdivisions exists between the document level, i.e. the interview, and the utterance) to deeply nested ones (e.g. a corpus of legal texts, with a hierarchy of *Act* → *Article* → *Paragraph* → *Subparagraph* → *Point* → *Sentence*.[24] The straightforward way of representing such structures in a relational way would entail defining entities for each of the levels, with foreign-primary key relations between each pair of adjacent levels. This leads to a potentially large number of tables that need to be joined when querying

[23] https://github.com/postgrespro/rum

[24] Cf. the official style guide of the European Union for the formulation of acts: https://publications.europa.eu/code/en/en-120700.htm (2024-03-05).

the database to reach the required hierarchy level to restrict for query needs ("search for phenomenon X in sentences from only acts U, V, and W"). Our approach consists in positioning all tokens of a corpus in one virtual stream of characters[25] and compute start and end positions for all tokens, sentences, and all other hierarchy levels in this stream. This allows us to make use of PostgreSQL's native range type and operators[26], which allow for "skipping" hierarchy levels: To restrict matching sentences to only appear in certain acts, the character ranges of the respective acts can be directly compared to the indices of the sentences, without the need to first join the hierarchy chain down via the intermediate levels. In other words, this technique makes use of the fact that hierarchical containment is a transitive relation.

5 Discussion and Conclusions

We describe a new system that uses a modern database and successfully applies innovative structural and indexing solutions to the modeling and complex querying of large corpora, yielding performant retrieval of query results. Our strategy of logarithmic partitioning means that the first trends emerging from the first partitions are available quickly, which allows for prototyping on corpora of nearly unlimited size. The data model we employ in the database is designed to be as flexible as possible, allowing the representation of various corpora with very different, possibly deeply nested hierarchical structures (books, chapters, verses, parliamentary discussions, etc.).

We provide a tool for users to import corpora into LCP. It currently supports corpora that can be represented in the CoNLL-U format; additional formats will be accepted in the future to support the import of more complex corpora and of those with media files attached. Other types of corpora have been and will be imported directly by us.

In its current state, LCP comes with some limitations, of which we give both descriptions and suggestions for improvement below.

- Designing a whole new query language allowed us to make it both powerful and flexible, but its novelty also means that new users need to study it. To support them, our interfaces provides a query editor with auto-completion feature and a visual representation of the corpus structure. While we plan to provide options to translate CQP statements to DQD queries to ease the transition for users familiar with that language, DQD aims to be expressive and declarative and can unleash the full potential of our platform. The examples we provide for each corpus can easily be edited, and we are actively working on extending our documentation, as well as planning to teach workshops on LCP.

- Although CoNLL-U provides a reliable standard for corpus imports, it also comes with limitations on what information it encodes, and the importing process remains non-trivial. We will first extend support to the CoNLL-U Plus format, which allows for arbitrarily many fields, and will document common use cases such as named entities of multi-media corpora as we automatize their importing process.

- The fast and optimized architecture entails that hosting new server instances takes considerable knowledge and needs a relatively expensive infrastructure. Parties interested in deploying their own setup will find all the code of the individual modules accessible as open source software.[27]

- Our basic unit of analysis is the sentence. A consequence of this is that delivering other levels, for instance entire documents, requires extra retrieval steps. The platform compensates some of the extra cost by caching the results of recent queries, so that future queries can partially or completely reuse previous data in an optimized way.

- Negated existential queries are complex, as the entire corpus needs to be searched. Fortunately, the use of vectors allows us to query efficiently also in this situation.

[25] Similarly, we use a time axis for speech and multimodal corpora.
[26] https://www.postgresql.org/docs/current/functions-range.html
[27] The source code can be obtained from https://github.com/liri-uzh/lcp.

- We only support dependency syntax. This is a limitation compared to some other systems. However, LCP allows to define entities spanning arbitrary sequences, which are used, for instance, for the annotation of Named Entities, and could also be employed to mark phrases.

In the future, we will integrate further and more varied corpora. We have built frontends for text, speech and multimodal corpora, as well as a tool for visualizing content analysis with a purely graphical interface, allowing keyword extraction, topic modeling and time series analysis, using, among others, a large collection Swiss newspaper articles from the past decades (Graën et al., 2023).

References

Bartunov, O., & Sigaev, T. (2001). *Full-Text Search in PostgreSQL* (tech. rep.).

Celko, J. (2012). *Joe Celko's Trees and Hierarchies in SQL for Smarties*. Elsevier.

Christ, O. (1994). A Modular and Flexible Architecture for an IntegratedCorpus Query System.

Clematide, S. (2015). Reflections and a Proposal for a Query and Reporting Language for Richly Annotated Multiparallel Corpora. In G. Grigonyte, S. Clematide, A. Utka, & M. Volk (Eds.), *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA* (pp. 6–16). Linköping University Electronic Press.

Cochran, W. G. (1977). *Sampling Techniques* (3rd). Wiley.

Crystal, D., & Alan, C. (2023). *A dictionary of linguistics and phonetics*. John Wiley & Sons.

Desagulier, G. (2019). Can word vectors help corpus linguists? *Studia Neophilologica*.

Evert, S., & Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium.

Evert, S., & Hardie, A. (2015). Ziggurat: A new data model and indexing format for large annotated text corpora. *Challenges in the Management of Large Corpora (CMLC-3)*, 21–27.

Ghodke, S., & Bird, S. (2012). Fangorn: A system for querying very large treebanks. *Proceedings of COLING 2012: Demonstration Papers*, 175–182.

Graën, J. (2018). *Exploiting Alignment in Multiparallel Corpora for Applications in Linguistics and Language Learning* [Doctoral dissertation, University of Zurich].

Graën, J., Mustac, I., Rajovic, N., Schaber, J., Schneider, G., & Bubenhofer, N. (2023). Swissdox@ LiRI – a large database of media articles made accessible to researchers. *CLARIN Annual Conference Proceedings*, 111–115.

Gries, S. T., & Berez, A. L. (2017). Linguistic Annotation in/for Corpus Linguistics. *Handbook of Linguistic Annotation*, 379–409.

Hearst, M. A. (1999). Untangling Text Data Mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 3–10.

Ide, N., & Romary, L. (2004). International Standard for a Linguistic Annotation Framework. *Natural Language Engineering*, *10*(3–4), 211–225.

Ide, N., & Romary, L. (2006). Representing Linguistic Corpora and Their Annotations. *LREC*, 225–228.

König, E., & Lezius, W. (2000). A Description Language for Syntactically Annotated Corpora. *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Krause, T., & Zeldes, A. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, *31*(1), 118–139.

Lai, C., & Bird, S. (2010). Querying Linguistic Trees. *Journal of Logic, Language and Information*, *19*, 53–73. https://doi.org/10.1007/s10849-009-9086-9

Leach, P., Mealling, M., & Salz, R. (2005). *A Universally Unique IDentifier (UUID) URN Namespace* (tech. rep.).

Lehmann, H. M., & Schneider, G. (2012). BNC Dependency Bank 1.0. In S. O. Ebeling, J. Ebeling, & H. Hasselgård (Eds.), *Studies in variation, contacts and change in English, volume 12: Aspects of corpus linguistics: Compilation, annotation, analysis*. Varieng.

Manning, C. D. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.

Meurers, W. D. (2005). On the use of electronic corpora for theoretical linguistics: Case studies from the syntax of German. *Lingua*, *115*(11), 1619–1639.

Rayson, P. E., Mariani, J. A., Anderson-Cooper, B., Baron, A., Gullick, D. S., Moore, A., & Wattam, S. (2017). Towards Interactive Multidimensional Visualisations for Corpus Linguistics. *Journal for Language Technology and Computational Linguistics*, *31*(1), 27–49.

Rohde, D. L. T. (2005). TGrep2 user manual.

Schneider, R. (2013). KoGra-DB: Using MapReduce for language corpora. In M. Horbach (Ed.), *INFORMATIK 2013 – Informatik angepasst an Mensch, Organisation und Umwelt* (pp. 140–142). Gesellschaft für Informatik e.V.

Zeldes, A., Lüdeling, A., Ritz, J., & Chiarcos, C. (2009). ANNIS: A search tool for multi-layer annotated corpora.

“Hier in diesem Hause sitzen keine Idioten!” — Emotion and Concreteness in Austrian Parliamentary Discourse

Klaus Hofmann
Department of English
and American Studies
University of Vienna, Austria
klaus.hofmann@univie.ac.at

Tanja Wissik
Austrian Center for Digital Humanities and
Cultural Heritage (ACDH-CH)
Austrian Academy of Sciences (ÖAW)
tanja.wissik@oeaw.ac.at

Abstract

This study examines Austrian parliamentary discourse styles by combining utterances from the Corpus of Austrian Parliamentary Records (ParlAT; Wissik & Pirker, 2018) with a large dataset of affective norms for German (Köper & Schulte im Walde, 2016). The results suggest that parliamentary discourse styles differ significantly depending on gender, party affiliation and utterance type (regular speech vs. unauthorized utterances). The findings are discussed within the context of gendered language usage and the literature on political speech in general. In particular, we find evidence for a characteristically male right-wing populist mode of parliamentary discourse marked by negative and concrete language use and a penchant for heckling. It is also shown that discourse styles can vary over time, specifically when the parties in power change from one period to the next (e.g. a center-left/center-right coalition government following a center-right/right one).

1 Introduction and related research

Political discourse has proven a rich source for research in the humanities, the social sciences and beyond. Parliamentary discourse is often at the center of this research, as it ranks among the most prototypical and best-documented types of political speech. A growing number of machine-readable and annotated corpora of parliamentary proceedings have become available in recent years (e.g. Fišer et al., 2018: 1321; Ogrodniczuk et al., 2022), facilitating computer-based and specifically quantitative approaches. A number of initiatives and projects in the context of CLARIN are dedicated to the creation and analysis of parliamentary-based sources. These include the ParlaCLARIN recommendations for encoding parliamentary records (Erjavec & Pančur, 2021) and the multilingual comparative ParlaMint data set (Erjavec, Kopp, et al., 2023; Erjavec, Ogrodniczuk, et al., 2023).

The present study focuses on parliamentary discourse in Austria. The study is conducted within the CLARIAH-AT context, which forms part of the larger CLARIN enterprise. The aim of the study is to explore how lexical sentiment data (i.e. the emotional value and strength of words) and abstractness ratings may inform our understanding of parliamentary discourse in Austria. Specifically, we ask (a) to what degree language usage as defined by these metrics is related to factors such as gender, party membership and parliamentary role, (b) to what degree language usage is subject to change over time, and (c) whether usage differs in different utterances types, i.e. regular speeches (1) vs. unauthorized utterances (2), e.g. heckling:

(1) Abg. Dr. Josef Cap (SPÖ):
[...] Das ist die Strategie, die Sie haben, aber da machen Sie einen Fehler. Da machen Sie

Klaus Hofmann and Tanja Wissik 2023. “Hier in diesem Hause sitzen keine Idioten!” — Emotion and Concreteness in Austrian Parliamentary Discourse. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 76–88. DOI: https://doi.org/10.3384/ecp210

einen schweren Fehler! [...] Hier in diesem Hause sitzen keine Idioten, und daher werden Sie hier immer wieder von uns vorgeführt werden für die Politik, für die Sie stehen. [...]
['That is the strategy that you have, but you are making a mistake there. You are making a grave mistake. [...] There are no idiots sitting here in this House, and therefore you will be exposed for the politics that you stand for time and again. [...]'

(XX, 181st sitting, 15 July 1999, p. 148)[1]

(2) a. Abg. Mag. Karl Schweitzer (FPÖ):
Der Herr Bundeskanzler hat es schon eilig, Herr Kollege Cap!
['The Chancellor is in a hurry, dear colleague Cap!']

(XX, 181st sitting, 15 July 1999, p. 148)[2]

b. Abg. Dr. Madeleine Petrovic (Grüne):
Das haben Sie verkürzt und falsch zitiert!
['You are quoting this in an abbreviated and misleading way!']

(XXI, 9th sitting, 9 February 2000, p. 217)[3]

c. Abg. Rosemarie Bauer (ÖVP):
Aber geh!
[Come on!']

(XX, 93rd sitting, 5 November 1997, p. 45)[4]

Sentiment analysis has been applied to parliamentary speeches before (e.g. Abercrombie & Batista-Navarro, 2020). However, few studies have integrated sentiment with abstractness scores in an attempt to profile linguistic usage more broadly. Studies on parliamentary discourse in the Austrian context are generally still rare (Haselmayer et al., 2022; but see Kern et al., 2021). We propose that a focus on sentiment and abstractness metrics could prove a valuable addition to existing methods for profiling political language and we demonstrate this in the Austrian case.

The remainder of this paper is structured as follows: The data and the method are discussed in Chapter 2. The results from the quantitative analysis are presented in Chapter 3 and discussed in relation with the relevant literature in Chapter 4. Chapter 5 concludes the paper with an outlook for future research.

2 Data and method of analysis

2.1 Corpus of Austrian Parliamentary Records

With the release of the ParlaMint 3.0 dataset (Erjavec, Kopp, et al., 2023), there is now a multilingual comparable corpus of parliamentary records from across Europe, which also contains Austrian data. The present study, however, makes use of a somewhat older data set with slightly different mark-up, called the Corpus of Austrian Parliamentary Records (ParlAT) (Wissik & Pirker, 2018). The corpus contains the

[1] 181st sitting of the National Council of the Republic of Austria, XXth legislative period, 15 July 1999, Shorthand Record, p. 148. https://www.parlament.gv.at/dokument/XX/NRSITZ/181/SEITE_0148.html
[2] 181st sitting of the National Council of the Republic of Austria, XXth legislative period, 15 July 1999, Shorthand Record, p. 148. https://www.parlament.gv.at/dokument/XX/NRSITZ/181/SEITE_0148.html
[3] 9th sitting of the National Council of the Republic of Austria, XXIst legislative period, 9 February 2000, Shorthand Record, p. 217, 9 February 2000, p. 217. https://www.parlament.gv.at/dokument/XXI/NRSITZ/9/SEITE_0217.html
[4] 93rd sitting of the National Council of the Republic of Austria, XXth legislative period, 5 November 1997, Shorthand Record, p. 45. https://www.parlament.gv.at/dokument/XX/NRSITZ/93/SEITE_0045.html

parliamentary records of the National Chamber (Nationalrat), one of two chambers of the Austrian parliament, from the XXth to the XXVth legislative periods, specifically the years between 1996 and 2017. This is a longer period than is available for most countries in ParlaMint, even though the version of ParlAT used here does not include the final years leading up to the present. ParlAT is based on official transcripts (produced from shorthand). Besides being tokenized, part-of-speech tagged and lemmatized, all speeches delivered by members of parliament are annotated as utterances and each speaker is identified and marked up, so that every utterance can be linked to a specific speaker. For each speaker, additional metadata is provided, including gender and party membership. One major advantage of ParlAT over ParlaMint is that verbal interruptions, like heckling or statements of approval, are annotated with the same amount of detail as regular speeches (Wissik, 2022), so that both types of parliamentary expression can be compared with regard to their usage. Other information provided by the stenographers, like applause, interruptions, descriptions of procedures, scenes or gestures, are annotated as notes. In its entirety, the corpus contains approximately 75 million tokens representing over 600 000 word forms and 400 000 lemmas.

2.2 Dictionary of German affective norms

The dictionary of German affective norms (Köper & Schulte im Walde, 2016) contains 350 000 German lemmas (including nouns, verbs, adjectives and adverbs), automatically rated by a supervised learning algorithm on four affective dimensions, namely valence, arousal, abstractness/concreteness, and imageability. Valence refers to the value of the emotional response elicited by a word, which can range from very positive to very negative (e.g. *Geschenk* 'gift' vs. *Strafe* 'punishment'). Arousal describes the intensity of emotion provoked by a lexical stimulus (e.g. *ruhig* 'calm' vs. *gewalttätig* 'violent'). Abstractness/concreteness measures the degree to which the concept denoted by a word is accessible to the human senses (e.g. *Ball* 'ball' vs. *Theorie* 'theory'). Imageability refers to the degree to which concepts can be experienced through human vision (e.g. *Tisch* 'table' vs. *Glaube* 'belief') (Table 1).

Word	**Valence**	**Arousal**	**Abstractness/ Concreteness**	**Imageability**
feiern ('celebrate')	7.079	5.978	3.869	5.633
erschießen ('shoot dead')	1.257	8.946	5.955	7.31
variieren ('vary')	4.078	4.433	3.298	1.924
vernachlässigen ('neglect')	3.254	4.716	1.574	2.558

Table 1. Verb examples of the four affective norms (Köper & Schulte im Walde, 2016). Values range from 0 (lowest) to 10 (highest).

One major advantage of using this data set over others is its size, which the compilers achieved through propagation from human-rated seed words using deep-learning-based skip-gram embeddings (also known as word2vec, Mikolov et al., 2013). We are aware that this data set is not specifically tailored to Austrian German, but in the absence of a comparable and similarly comprehensive data set for Austrian usage we settled for it in the interest of maximizing coverage.

2.3 Data processing and method of analysis

A subcorpus including utterances by members of parliament (speeches and verbal interruptions) but excluding all procedural content (e.g. utterances by the chair) was compiled from the parliamentary

corpus. Procedural content only accounts for less than 10% of the whole corpus, so this did not reduce the amount of linguistic data much. All speaker variables, including speaker gender, party membership and parliamentary role (government vs. opposition), were linked directly to the utterances. A custom stopword list excluded function words, titles, recurrent phrases (e.g. *Bundeskanzler* 'Chancellor', *Hohes Haus* 'Parliament' [lit.: High House']) as well as party names and associated terms (e.g. *FPÖ* 'Freedom Party', *sozialdemokratisch* 'Social Democrat', *Volkspartei* 'People's Party'). The resultant utterance subcorpus was merged with the affective norms dictionary data set. Every word in the utterance subcorpus received a rating for each of the four dimensions arousal, valence, concreteness and imageability as long as an entry for the word existed in the dictionary. Since the affective norms dictionary is unique in its size and comprehensiveness, coverage was very good.

Compared to more sophisticated machine learning approaches for sentence sentiment classification, which typically only deal with one affective dimension (Haselmayer et al., 2022), analyses based on norms dictionaries is a relatively simple but also accessible and transparent method for investigating the relative impacts of a range of predictors. We do acknowledge the chosen method's likely limitations in terms of sentence classification accuracy, and accept this drawback in the interest of maximizing the breadth of affective dimensions covered. In the absence of a sufficient amount of labeled data for both sentiment and abstractness, resorting to existing lexicons is our best hope for capturing the effects of the predictors on these affective dimensions.

We opt for linear regression modeling to analyze these affective metrics. Arousal, valence, concreteness and imageability serve as dependent variables. The predictor variables include all collected speaker-related information, as well as utterance type (regular speech vs. unauthorized interruption) and legislative period. For analyzing utterance type as a dependent variable, we use binomial logistic regression, with the same set of predictors (minus utterance type itself). All calculations are carried out with R (R Core Team, 2023).

3 Results

3.1 Regular vs. unauthorized utterance types

The first set of results concerns the distribution of utterance types, operationalized here as the likelihood that a given utterance is unauthorized (e.g. heckling) or regular (i.e. parliamentary speeches). Since the output variable UTTERANCE TYPE is binary and categorical, the fitted model is a binomial regression model with four categorical predictor variables — GENDER, PARTY, ROLE, and PERIOD — added as main effects (Figure 1). The shaded areas in the graph represent 95% confidence intervals. If the confidence intervals of two values do not overlap, this can be interpreted as a significant difference.

The model reveals clear differences in the propensities to use unauthorized utterances in parliamentary sessions depending on all predictor variables. When converting the log odds ratio to simple odds, it emerges that an unauthorized utterance is about 1.5 times as likely to be uttered by a male member of parliament than a female one. Party affiliation also plays a significant role. It appears that the right-wing FPÖ and its short-lived spin-off BZÖ are the most likely to employ unauthorized utterances. In fact, their use of unauthorized utterances is about as frequent as their use of regular utterances. The Green Party, the liberal LIF/NEOS and the minor and short-lived Team Stronach are the least inclined to use unauthorized utterances, while the traditional center parties SPÖ and ÖVP are placed in the middle of this cline. The effect size is stark, with a given unauthorized utterance being three times more likely to be uttered by an FPÖ member than a LIF/NEOS member. As can be expected, members of the opposition use heckling and verbal applause to a higher degree and members of the government parties. Notably, the rate of unauthorized utterances reached its peak during parliamentary periods XXI and XXII, when the FPÖ (later as BZÖ) were in power as part of governing coalitions.

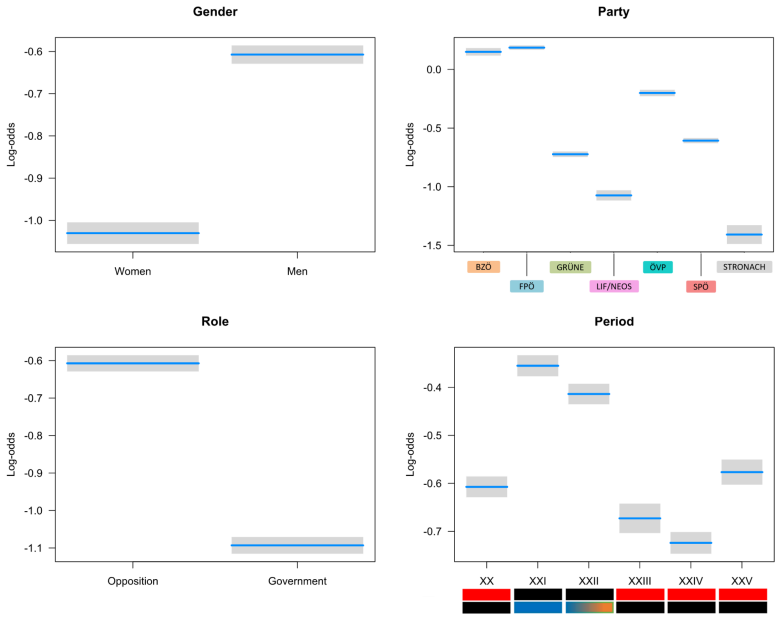


Figure 1. Effect of GENDER, PARTY, ROLE, and PERIOD on UTTERANCE TYPE. Shaded areas denote 95% confidence intervals.

3.2 Valence

Next, the first affective outcome variable to be examined is VALENCE, a metric that measures the degree of positive or negative connotation of words. VALENCE is a continuous variable, ranging from 0 (very negative) to 10 (very positive). Although VALENCE is thus bounded, few if any items actually hit the boundary values. The fitted model is therefore an ordinary linear regression model with the five categorical predictors — GENDER, PARTY, UTTERANCE TYPE, ROLE, and PERIOD — entered as main effects. To enhance comparability, the outcome is displayed as centered and scaled, that is, each element of the outcome variable is divided by the standard deviation and the mean is subtracted (Figure 2). The shaded areas in the graph once again represent 95% confidence intervals. Non-overlapping confidence intervals can be interpreted as a significant difference between two values.

Once again, all predictor variables have significant effects on the outcome. Thus, women generally use more positive language in parliament than their male colleagues. In terms of party affiliation, the right-wing FPÖ and the BZÖ once again stand out from the rest as the parties with the lowest valence values. In contrast, the most positive language is employed by the liberal NEOS, the center-right ÖVP, and the center-left SPÖ. The Green Party is positioned around the middle of the field. As may be expected, unauthorized speech is characterized by more negative language than unauthorized speech, and markedly so. The difference is the greatest between any two predictor values in the whole model, which suggests that most speech contributions of the unauthorized types can be classed as heckling rather than statements of approval. Government parties speak in more positive terms than opposition parties. Finally, it is interesting to see that valence was at its lowest point during the earliest two periods, particularly during the first right/center-right coalition, but pivoted towards more positive usage during the period that also saw the split within the right-wing FPÖ.

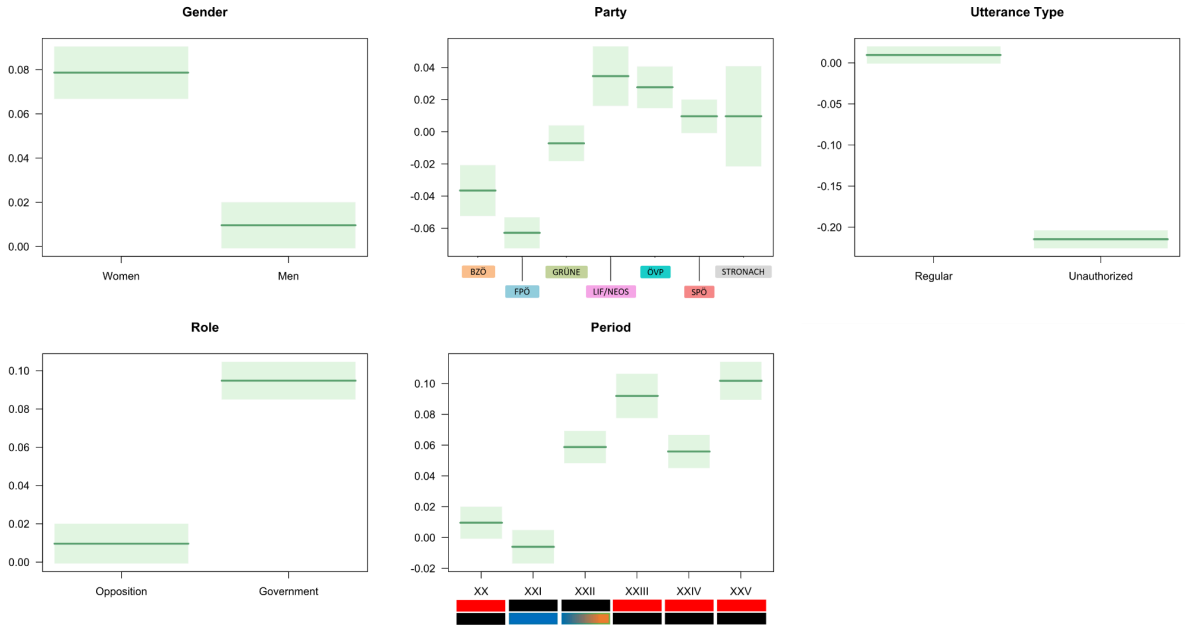


Figure 2. Effect of GENDER, PARTY, ROLE, PERIOD and UTTERANCE TYPE on VALENCE. Shaded areas denote 95% confidence intervals.

3.3 Arousal

The next affective outcome variable AROUSAL refers to the intensity of emotion associated with words. As with valence, we choose a linear model, with the same five predictors as main effects (Figure 3).

Women members of parliament seem to employ more arousing language than men. In terms of party affiliation, the different parties do for the most part not differ significantly from one another. It is worth pointing out, however, that the center-right ÖVP is situated at the lower end of the spectrum, while the FPÖ and NEOS are at the upper end. Unauthorized utterances are sharply more arousing than regular speech contributions. This is not surprising, given that the function of heckling and acclamations is to convey attitudes and positions in a succinct way without provoking a call to order from the chair for interrupting the current speaker. Regarding parliamentary role, members of government parties do use arousing speech to a slightly lower degree, even though this difference is not as marked as with valence before. Over the investigated period, language has generally become more arousing, specifically starting during parliamentary period XXI, that is, the time of the first right/center-right government coalition. It remained approximately at the same level after that.

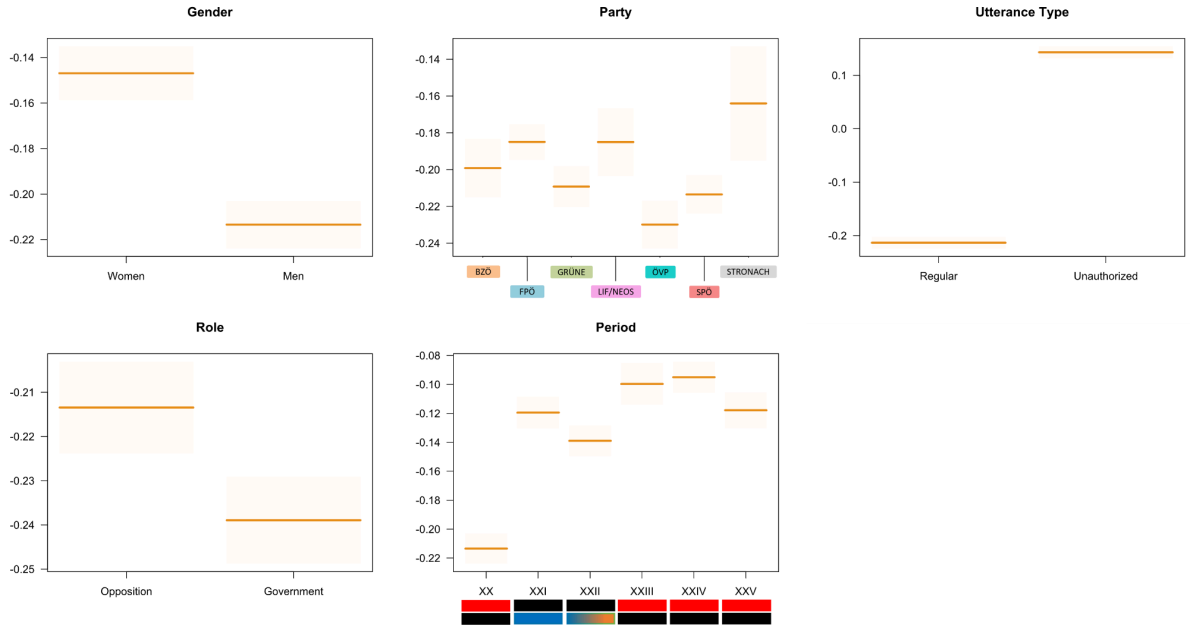


Figure 3. Effect of GENDER, PARTY, ROLE, PERIOD and UTTERANCE TYPE on AROUSAL. Shaded areas denote 95% confidence intervals.

3.4 Concreteness

Next, we turn to concreteness, a measure of how accessible the concept referenced by a word is to the five senses. The modeling procedure is the same as with the previous two outcome variables (Figure 4).

Men appear to use more concrete language, even though the difference is rather slight. More tellingly perhaps, it is once again members of the BZÖ and the FPÖ (and the short-lived Team Stronach) who employ the most concrete language, while the other end of the cline is again filled by the liberal NEOS party. The cleft between the two is striking. This is by far the predictor with the most pronounced difference between its extreme values in the whole model. The two center parties ÖVP and SPÖ are placed almost identically in a medial position of the spectrum, while the Green Party tends slightly more towards abstractness. With regard to the remaining predictors, regular speeches exhibit slightly more concrete language than unauthorized utterances and opposition parties tend to employ more concrete language. Over time, usage trended towards more abstract language.

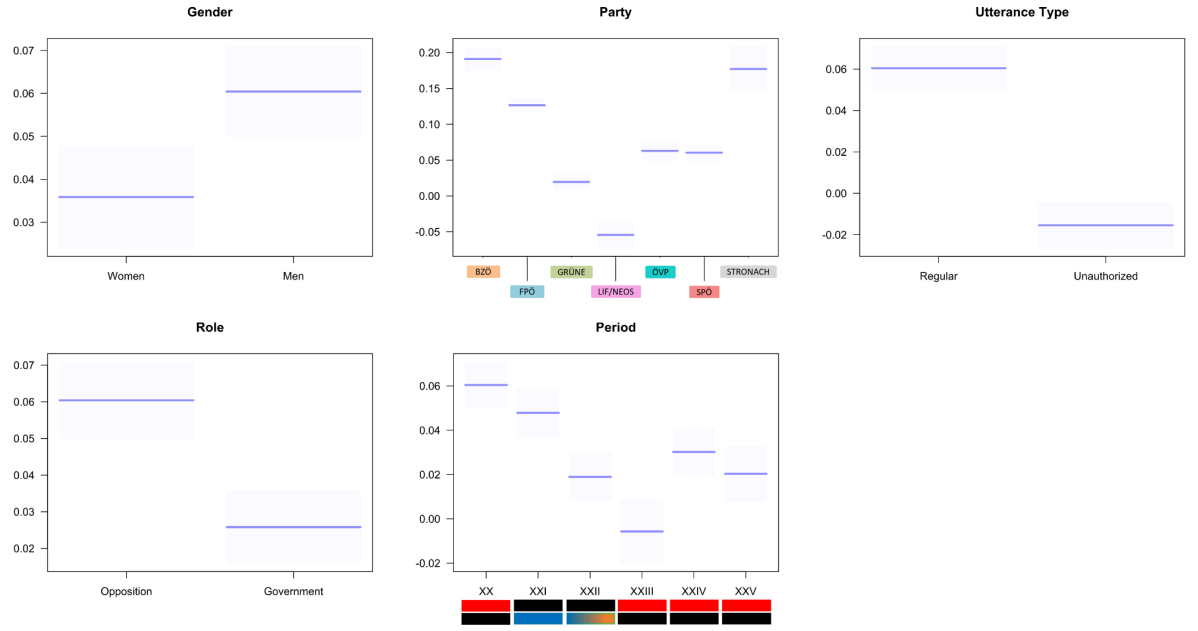


Figure 4. Effect of GENDER, PARTY, ROLE, PERIOD and UTTERANCE TYPE on CONCRETENESS. Shaded areas denote 95% confidence intervals.

3.5 Imageability

The final measure to be discussed is IMAGEABILITY (Figure 5). In contrast to CONCRETENESS, IMAGEABILITY specifically addresses the question how accessible a word's meaning is to the sense of vision, rather than all senses combined. The modeling procedure is the same as before.

It is obvious that CONCRETENESS and IMAGEABILITY behave in highly similar ways. However, the differences between predictor values with IMAGEABILITY as outcome variable are sometimes less pronounced compared to the model with CONCRETENESS as outcome variable. Regarding GENDER, for example, the estimate for men's speech does not emerge as significantly different from that for women's speech. The IMAGEABILITY estimates for the predictor PARTY are almost a mirror image of the CONCRETENESS estimates, with the right-wing parties at the high end and the liberal parties at the low end. The predictive behavior of parliamentary ROLE and PERIOD is also comparable to the CONCRETENESS model. The only predictor that behaves in a notably different way is UTTERANCE TYPE. Here the estimates for the predictor values are reversed, such that unauthorized speech is linked to higher imageability estimates than regular speeches.

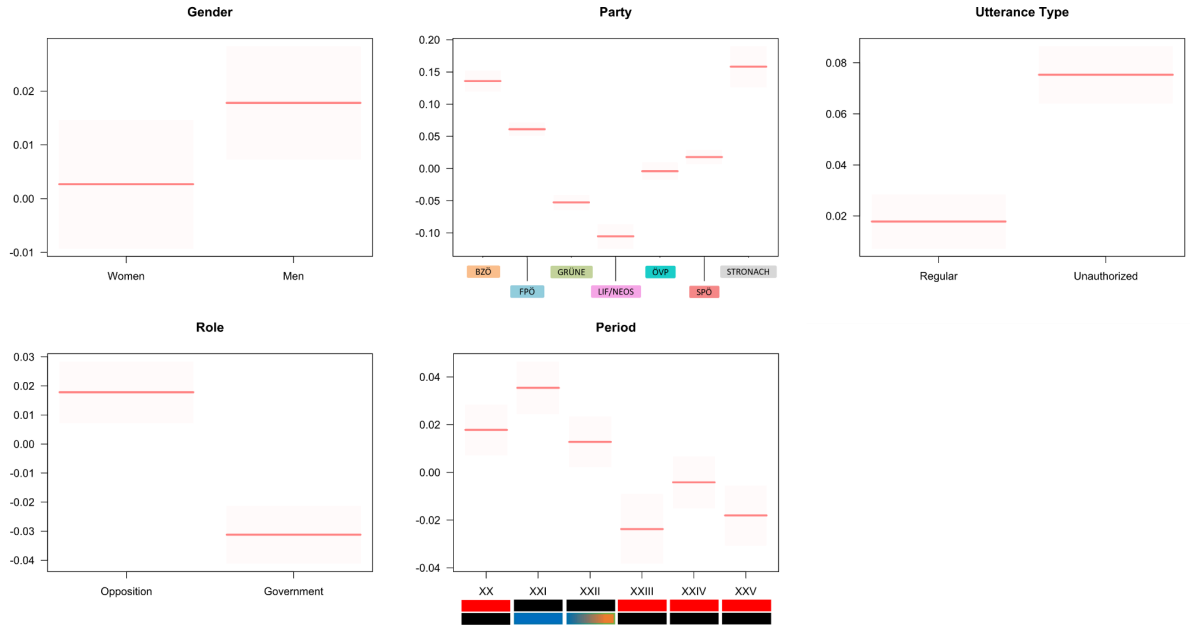


Figure 5. Effect of GENDER, PARTY, ROLE, PERIOD and UTTERANCE TYPE on IMAGEABILITY. Shaded areas denote 95% confidence intervals.

4 Discussion

Our results point to marked differences in language use across all investigated variables. In line with previous research on gendered language usage, female members of the Austrian parliament generally display more positive (Haselmayer et al., 2022; cf. Mehl & Pennebaker, 2003) and more arousing (Thomson & Murachver, 2001) language compared to their male colleagues. An oft-repeated explanation for the reported bias towards positivity is that women typically assume less contrarian interactional roles than men, instead foregrounding co-operativeness and maintaining relationships, while male speech is typically geared towards competition and domination. This behavior is ultimately a function of internalized expectations of gender-congruous behavior derived from a binaristic conception of gender based on biological sex (Eagly & Karau, 2002). It has also been pointed out that any such effects should not be seen as deterministic, neither biologically nor socially, since they are often small and can be overridden by factors such as measurement context (Anderson & Leaper, 1998; Hyde, 2005). In the present study, gender effects also tend to be small, but significant. Whatever the origin of the gendered use of positive and arousing language, Austrian parliamentary discourse seems to be one context where these aforementioned generalizations hold. Gender also plays a role with regard to utterance type, with men accounting for a much larger share of unauthorized utterances like heckling. This matches previous results in the literature, where men are found to interrupt the speech of others more often than women, especially when the type of interruption can be classed as 'intrusive' (Anderson & Leaper, 1998).

Contrary to expectation, however, men display higher concreteness scores than women do. This does not straightforwardly match previous findings that men utilize abstract communication as a signal of power , impeding women's chances to emerge as leaders (Joshi et al., 2020, 2021). At the same time, women's tendency to discuss policies in more concrete terms has been suggested to carry potential for the political process as it may serve to communicate more effectively to the wider public (Hargrave & Langengen,

2021). While concreteness does not correlate with gender in the predicted way in the present study, it does exhibit significantly increased levels in the language of opposition parties compared to government parties. This points to another function of abstractness, namely to signal power by conveying the ability to grasp the bigger picture, think strategically, and distill the general from the specific (Palmeira, 2015; Wakslak et al., 2014), in other words: to be in control of things. It is therefore not surprising that members of government parties would shift towards more abstract speech patterns when they come into power. Neither is it surprising that government parties would use more positive and less arousing language, as their continued political success depends on communicating a sense of stability and achievement. However, it has also been observed that abstract language runs the danger of being perceived as lacking the initiative and the capability to implement projects (Palmeira, 2015).

A different way of looking at these results takes into account that emotion and abstractness also vary drastically with party affiliation. Thus, the male-dominated right-wing parties use concrete language to a larger degree than liberal parties, while the traditional center parties ÖVP and SPÖ are located in the middle of this continuum. Additionally, the right-wing parties set themselves apart by more negative language compared to other parties as well as relatively high arousal scores. Finally, it is also the right-wing parties which tend to employ interruptions much more often than other, particularly liberal parties. Taken together, these results seem to coalesce around a characteristically male right-wing populist mode of parliamentary discourse, defined by negative, arousing and concrete language with a penchant for unauthorized interruptions. It contrasts with a liberal mode, which is less concrete, more positive, and more likely to be transported through regular speeches rather than irregular ones.

Populism is a multifaceted notion and as such not easy to define (De Vreese et al., 2018), but an integral part of all shades of political populism is that it pitches ‘the people’ against ‘the elite’ (Mudde, 2007). It is often assumed that populist speech appeals to the general public by using “simple” language (Canovan, 1999; Moffitt & Tormey, 2014) or by “mimic[king] the language of ordinary citizens” (Martelli & Jaffrelot, 2023). However, simplicity as a linguistic criterion has proved difficult to detect when measured in terms of lexico-grammatical complexity in the speech of well-known populist leaders (McDonnell & Ondelli, 2022). Instead, concreteness, negative valence and high arousal may serve as a combination of metrics that better fits the impressionistic descriptions of populist language as ‘simple’ or ‘direct’. Thus, the intended effect of employing this kind of language may primarily be to draw a stark contrast between the populist parties and the more abstract and positive diction of the establishment, which is thereby made to seem aloof and inert. Heckling of regular speeches further heightens the effect of resistance against an established order. It is worth noting that right-wing populist speech defined in these terms does not have a clear equivalent on the left. Either the language of left populism has different characteristics or there is no political group in the Austrian parliament that fills that end of the rhetorical spectrum.

It is perhaps telling that many of the larger diachronic shifts that we observe in the data happened at the start, during, or right after the two legislative periods when a right-wing party participated in a right/center-right coalition government. Valence shifted from its lowest levels to more positive language, arousal jumped, and concreteness started a slow descent towards relatively more abstract language. This may be a consequence of right-wing parties finding themselves in a position of power, which does not naturally fit their anti-establishment rhetoric. This may have led to a recalibration of discourse tactics towards a more moderate style. On the other hand, interruptions still peaked at the end of this time, which may reflect a generally tumultuous parliament during the XXIst and XXIInd periods, with liberal and center parties partly adopting this aspect of populist speech.

5 Conclusion and future work

We hope to have demonstrated that analyzing parliamentary speech through a combination of sentiment and abstractness metrics while also distinguishing between regular and unauthorized speech can provide valuable insights into how language is used strategically in parliament. We have also demonstrated how these metrics can be used to profile the speech of different kinds of actors in parliament, variably defined by their gender, party affiliation and parliamentary roles. We have argued for a characteristically male, right-wing mode of parliamentary expression, defined by low valence, relatively high arousal, and high concreteness/imageability coupled with a strong inclination for verbal interruptions, particularly heckling.

Using existing sentiment and abstractness lexicons is an intuitive and accessible way to approach these questions. One drawback of our approach, however, is that such lexicons classify each lexical item individually rather than utterances as a whole. Thus, rhetorical practices like irony, which is not at all rare particularly in heckling, cannot be captured reliably by our method. There is a real possibility that our valence estimates may be underestimating the effective emotional content of such utterances.

There are numerous additional avenues for profiling language usage that can be explored in future work. For example, an additional way of analyzing the data would include interaction terms in the regression models. By controlling for party membership, for example, one could test the tentative conclusion that speech strategies become more moderated when parties change from an opposition to a government role. Interactions would also make it possible to tease apart the relative impact of gender and political affiliation on concreteness values to see whether male speech is indeed more concrete across the political spectrum, as the analysis has suggested, or whether this aspect varies across different parties. Also, the degree to which various policy domains or subject matters determine the affective import and abstractness of politicians' language was not systematically taken into account in our study. Subject matter could be controlled as a factor by integrating a topic modeling element in future research (cf. Skubic et al., 2022). In addition, diachronic trends may be examined further, for example, by investigating whether the observed general increase in valence and arousal might correlate with a changing demographic composition in the parliament itself, specifically through an increase in the relative number of female representatives. Finally, it will be interesting to see whether the generalizations derived from the Austrian data are mirrored in other countries all over Europe. The ParlaMint data set and diverse sentiment and abstractness lexicons (Brysbaert, Stevens, et al., 2014; Brysbaert, Warriner, et al., 2014; Stadthagen-Gonzalez et al., 2017) can assist in broadening the scope of this line of research.

References

Anderson, K. J., & Leaper, C. (1998). Meta-analyses of gender effects on conversational interruption: Who, what, when, where, and how. *Sex Roles*, *39*(3/4), 225–252. https://doi.org/10.1023/A:1018802521676

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*, 80–84. https://doi.org/10.1016/j.actpsy.2014.04.010

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Canovan, M. (1999). Trust the people! Populism and the two faces of democracy. *Political Studies*, *47*(1), 2–16. https://doi.org/10.1111/1467-9248.00184

De Vreese, C. H., Esser, F., Aalberg, T., Reinemann, C., & Stanyer, J. (2018). Populism as an expression of political communication content and style: A new perspective. *The International Journal of Press/Politics*, *23*(4), 423–438. https://doi.org/10.1177/1940161218790035

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. https://doi.org/10.1037/0033-295X.109.3.573
Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M. del M., Calzada Pérez, M., … Fišer, D. (2023). *Multilingual comparable corpora of parliamentary debates ParlaMint 4.0*. http://hdl.handle.net/11356/1859
Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., De Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., De Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, *57*(1), 415–448. https://doi.org/10.1007/s10579-021-09574-0
Erjavec, T., & Pančur, A. (2021). The Parla-CLARIN recommendations for encoding corpora of parliamentary proceedings. *Journal of the Text Encoding Initiative*, *14*. https://doi.org/10.4000/jtei.4133
Fišer, D., Lenardič, J., & Erjavec, T. (2018). CLARIN's key resource families. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). https://aclanthology.org/L18-1210
Hargrave, L., & Langengen, T. (2021). The gendered debate: Do men and women communicate differently in the House of Commons? *Politics & Gender*, *17*(4), 580–606. https://doi.org/10.1017/S1743923X20000100
Haselmayer, M., Dingler, S. C., & Jenny, M. (2022). How women shape negativity in parliamentary speeches—A sentiment analysis of debates in the austrian parliament. *Parliamentary Affairs*, *75*(4), 867–886. https://doi.org/10.1093/pa/gsab045
Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*(6), 581–592. https://doi.org/10.1037/0003-066X.60.6.581
Joshi, P. D., Wakslak, C. J., Appel, G., & Huang, L. (2020). Gender differences in communicative abstraction. *Journal of Personality and Social Psychology*, *118*(3), 417–435. https://doi.org/10.1037/pspa0000177
Joshi, P. D., Wakslak, C. J., Huang, L., & Appel, G. (2021). Gender differences in communicative abstraction and their organizational implications. *Rutgers Business Review*, *6*(2), 145–153.
Kern, B., Hofmann, K., Baumann, A., & Wissik, T. (2021). Komparative Zeitreihenanalyse der lexikalischen Stabilität und Emotion in österreichischen Korpusdaten. Katsikadelij, C., Sellner, M., & Gassner, M. (Eds.), *Digital Lexis and Beyond*, 1–15. Austrian Academy of Sciences Press. https://doi.org/10.1553/Austrian_Corpora
Köper, M., & Schulte im Walde, S. (2016). Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 german lemmas. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2595–2598. European Language Resources Association (ELRA). https://aclanthology.org/L16-1413
Martelli, J., & Jaffrelot, C. (2023). Do populist leaders mimic the language of ordinary citizens? Evidence from India. *Political Psychology*, *44*(5), 1141–1160. https://doi.org/10.1111/pops.12881
McDonnell, D., & Ondelli, S. (2022). The language of right-wing populist leaders: Not so simple. *Perspectives on Politics*, *20*(3), 828–841. https://doi.org/10.1017/S1537592720002418
Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, *84*(4), 857–870. https://doi.org/10.1037/0022-3514.84.4.857
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in Vector Space. *arXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781
Moffitt, B., & Tormey, S. (2014). Rethinking populism: Politics, mediatisation and political style. *Political Studies*, *62*(2), 381–397. https://doi.org/10.1111/1467-9248.12032
Mudde, C. (2007). *Populist radical right parties in Europe*. Cambridge University Press.

https://doi.org/10.1017/CBO9780511492037
Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešić, N., Çöltekin, Ç., Kopp, M., & Katja, M. (2022). ParlaMint II: The show must go on. *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, 1–6. European Language Resources Association. https://aclanthology.org/2022.parlaclarin-1.1
Palmeira, M. (2015). Abstract language signals power, but also lack of action orientation. *Journal of Experimental Social Psychology*, *61*, 59–63. https://doi.org/10.1016/j.jesp.2015.07.003
R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
Skubic, J., Angermeier, J., Bruncrona, A., Evkoski, B., & Leiminger, L. (2022). Networks of power: Gender analysis in selected European parliaments. *Proccedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022), Potsdam, Germany*, 37–45. https://old.gscl.org/en/arbeitskreise/cpss/cpss-2022/workshop-proceedings-2022
Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, *49*(1), 111–123. https://doi.org/10.3758/s13428-015-0700-2
Thomson, R., & Murachver, T. (2001). Predicting gender from electronic discourse. *British Journal of Social Psychology*, *40*(2), 193–208. https://doi.org/10.1348/014466601164812
Wakslak, C. J., Smith, P. K., & Han, A. (2014). Using abstract language signals power. *Journal of Personality and Social Psychology*, *107*(1), 41–55. https://doi.org/10.1037/a0036626
Wissik, T. (2022). Encoding interruptions in parliamentary data: From applause to interjections and laughter. *Journal of the Text Encoding Initiative*, *14*. https://doi.org/10.4000/jtei.4214
Wissik, T., & Pirker, H. (2018). ParlAT beta Corpus of Austrian Parliamentary Records. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 7–12. European Language Resources Association (ELRA).

Analyses of information security standards on data crawled from company web sites using SweClarin resources

Arne Jönsson
Computer and Information Science
Linköping University
Linköping, Sweden
arne.jonsson@liu.se

Subhomoy Bandyopadhyay
Management and Engineering
Linköping University
Linköping, Sweden
subhomoy.bandyopadhyay@liu.se

Svjetlana Pantic Dragisic
Management and Engineering
Linköping University
Linköping, Sweden
svjetlana.pantic.dragisic@liu.se

Andrea Fried
Management and Engineering
Linköping University
Linköping, Sweden
andrea.fried@liu.se

Abstract

With the purpose of analysing Swedish companies' adherence and adoption of the information security standard ISO 27001 and to examine the communicative constitution of preventive innovation in organisations, we have created a corpus of corporate texts from Swedish company websites. The corpus was analysed from multiple interdisciplinary perspectives in close cooperation with management researchers and SweClarin researchers using SweClarin tools and resources as well as standard language technology tools. Some analyses require deep reading, which was performed by management researchers, often guided by results from language analyses. Initial results have been presented at a management studies conference. In this paper, we focus on presenting the research issues, the methods used in the project, the results, and the experience of SweClarin researchers supporting researchers in social sciences. Our contribution is to show how it is possible, through the integration of human insights and digital methods, to increase the credibility and validity of a digitally acquired data set and subsequent research findings. In our view, a combination of human deep reading (management researchers), contextual lexical verification (management studies) and language technology (content and sentiment analysis) can help to sensitise computational text analysis for medium-sized data sets.

1 Introduction

Today's organisations are increasingly compelled to adopt preventive innovation to tackle pressing issues. Preventive innovation differs from ordinary innovation. The innovation literature claims that the economic benefits of preventive innovation to organisations, for instance, for avoiding environmental pollution, protecting human health or ensuring information security, are mainly intangible, often time-delayed and adopted for incidents that may never occur (Rogers, 1995). Conversely, innovations that are not preventive tend to yield immediate benefits and results in the near term. For example, implementing the ISO 9001 standard for quality management can significantly boost operational efficiency, leading to better customer satisfaction and reduced cycle times and inventory levels (Lo & Chang, 2007). Organisations face a significant challenge balancing the need for preventive innovations, which address long-term concerns, with the immediate demand for tangible results (Fineberg, 2013). This tension between long-term strategic security and short-term financial accountability complicates decision-making. As a result, companies often find themselves at a crossroad, struggling to allocate resources in a way that satisfies both current demands and future necessities. To address these challenges, organisational communication is crucial to increase the potential of economic recognition for preventive innovation.

Arne Jönsson, Subhomoy Bandyopadhyay, Svjetlana Pantic Dragisic and Andrea Fried 2023. Analyses of information security standards on data crawled from company web sites using SweClarin resources. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 89–101. DOI: https://doi.org/10.3384/ecp210

Therefore, drawing theoretically on the discourse perspective of organisational communication (Orlikowski & Yates, 1994; Yates & Orlikowski, 1992) we develop a communicative approach that enables a situation- and meaning-centred understanding of preventive innovation. Using the example of the information security standard ISO/IEC 27001, which is designed to mitigate future risks such as phishing emails, exploitation of stolen credentials, and software vulnerabilities, we examine how communication of preventive innovations is shaped by its adopting organisations. We analyse texts about the information security standard ISO/IEC 27001 on Swedish corporate websites supported by computational tools for web scraping and language analyses. As a result, we first identify three communicative practices of data governance termed agency, stewardship and brokerage, and second, provide evidence that organisations' communication also depends on whether they receive direct or indirect economic recognition for their preventive innovation.

We contribute a meaningful combination of deep reading of humans (researchers), dictionary verification for a specific context (innovation research) and language technology (content and sentiment analysis) to a meaning-centred and situational understanding of preventive innovation. A similar approach was used by Saura et al. (2023) where they used textual analysis, sentiment analysis, and topic modelling to analyse social media from an open innovation perspective. For a systematic overview of social media analyses of innovation management see Geissinger et al. (2023).

Our analysis enhances Rogers' (Rogers, 1995) perspective by challenging the classification of preventive innovations as mere "isolated, static objects or practices", unveiling their dynamic interplay with organisational members — simultaneously influencing and being influenced — i.e., are enacted communicatively by organisations. Contrary to Rogers' assumption, we also provide initial evidence that preventive innovations can very well achieve economic recognition by constituting different meanings of preventive innovation.

This paper will focus on the methodology, rather than delving into the findings. We illustrate the potential of SweClarin and language technology analyses for investigating organisational communication and the production of meaning in their texts.

2 Conceptual Background

The diffusion of innovation theory fundamentally revolves around communication, outlining how innovations spread within a social system's participants (Rogers, 1995). Studies highlight the critical role of creating meaning in communication about innovations to facilitate their spread and acknowledge innovative contributions. Innovation literature highlights specific characteristics that both shape how organisations communicate about adopting innovations and assist in analysing an organisation's perception of these innovations, known as secondary innovation attributes (Downs et al., 1976; Tornatzky & Klein, 1982). In examining how organisations communicate about ISO/IEC 27001, we utilise Rogers (1995) five distinct attributes that adopters use to perceive and ascribe meaning to preventive (and other) innovations: relative advantage, compatibility, complexity, trialability, and observability.

Relative advantage is "the degree to which an innovation is perceived as being better than the idea it supersedes" (Rogers, 1995, p.212). Compatibility is "the degree to which an innovation is perceived as consistent with the existing values, past experiences, and needs of potential adopters" (Rogers, 1995, p.224). Complexity refers to "the degree to which an innovation is perceived as relatively difficult to understand and use" (Rogers, 1995, p.242)). Trialability is about the extent to which the innovation can be experimented with by adopters (Rogers, 1995), while observability covers the ability to see, imagine, or explain the innovation's results to others (Rogers, 1995).

To advance a situational and meaning-centred perspective on preventive innovation, we employ these attributes and theoretically anchor our research in the discourse perspective of organisational communication (Fried et al., 2024). This perspective emphasises communication's role in community development, such as the Swedish companies focusing on preventive innovation for information security, cf. Putnam (1999). Discourses, as communicative acts, form and reflect innovation's essence and outcome, shaping and reshaping innovation's meaning within organisations, cf. Bergquist (1993), Putnam (1999), and Taylor (1993).

3 Generating the corpus

Using ISO/IEC 27001 as an example to study the communication of preventive innovations, our research design followed three steps, see Figure 1. We first generated a dataset of Swedish corporate websites of all sectors and scraped the content for ISO/IEC 27001 related paragraphs of the text corpus. Second, we categorised the identified companies manually according to their adoption (of preventive innovation) approach. Finally, we conducted analyses on the language used in the paragraphs relating to ISO/IEC 27001 on these websites.

Regarding the first step, as a complete dataset of all websites of Swedish companies does not exist as open access, we contacted several institutions to retrieve this data. We approached Sweden's company registration office, Bolagsverket, and Statistics Sweden (SCB) to get access to company names, identification numbers, sector affiliations and innovation indicators. However, Bolagsverket and SCB could not provide a database with company URLs. We, therefore, analysed 400 company names on Nasdaq Nordic (https://www.nasdaqomxnordic.com/) through scripts that generate web addresses in order to understand how company URLs can be constructed, and used that to generate 120 million possible URLs from the 2.4 million registered companies listed on Bolagsverket. These URLs were tested to check how many of them were actual websites. These websites were then scraped in September 2020. We scraped up to 50 connected web pages of each site to grasp sufficient content (cf., Kinne and Lenz (2019)). Out of all scraped websites, we found 472 which contained the phrases 'ISO 27001', 'IEC 27001', 'IEC 270' or 'ISO 270'[1].

After we had identified the 472 websites[2], as a second step, we manually analysed each company's website by visiting their URLs to verify the scrapped data. This hands-on scrutiny of the corporate websites aimed to refine the extracted information regarding companies' certifications, business sectors, models, and value propositions. After removing duplicates and further non-Swedish companies in the dataset, we were left with 353 websites of Swedish companies. We categorised these companies according to the criteria 'certified' or 'non-certified', following a suggestion by Mirtsch et al. (2020). Their findings reveal that a third of the companies that adopt ISO/IEC 27001 do so through certification, with the remainder opting for non-certified pathways. Furthermore, our findings revealed a variety of companies: some integrated ISO/IEC 27001 consulting or training into their business models, while others, lacking certification and refraining from offering consulting or training services, solely referenced certified clients, customers, and suppliers on their websites. Based on this initial categorisation, we identified six distinct types of preventive innovation adoption, denoted as 11, 12, 21, 22, 31, and 32.[3] into which each company belongs.

In addition, two text corpora were generated from all identified company websites, one in Swedish and one in English. We use fastText (Joulin et al., 2016) to separate the paragraphs. For each company, we take each paragraph and place it in an English or a Swedish text file, i.e. a company can have two files, one with English text and one with Swedish. The English text corpus, spanning around 450 pages, underwent manual analysis through deep reading, revealing that over 50% of the dataset consisted of inconsequential noise such as ads, menus, contact details, and website cookies. As an outcome of this analysis and in pursuit of methodological rigour through Swedish sense-based sentiment analysis (elaborated upon below), companies with English-only websites were excluded from the sample, resulting in 291 companies (final sample size)[4] with websites in either Swedish or both Swedish and English. Although certain Swedish websites maintained English versions, it is noteworthy that, for analytical efficiency, the term "Swedish only" pertains solely to these 291 entities, since the English text corpus had been excluded from further analysis. Table 1 depicts the number of sentences and words for each adoption type for the 291 companies with Swedish only text on their web pages.

This resulted in a text corpus of close to 9 million words, see Table 1. Examples of paragraphs from

[1] Including variants such as iso-27001 and Iso 270.

[2] Available at https://www.ida.liu.se/~arnjo82/472_webpages

[3] The first digit (1, 2, or 3) denotes the three data governance approaches: Agents, Stewards, and Brokers, whereas the second digit (1 or 2) signifies (in)direct economic benefits resulting from preventive communication adoption, evaluated based on ISO/IEC 27001 training/consultation provision.

[4] Available at https://www.ida.liu.se/~arnjo82/291_filtered_webpages

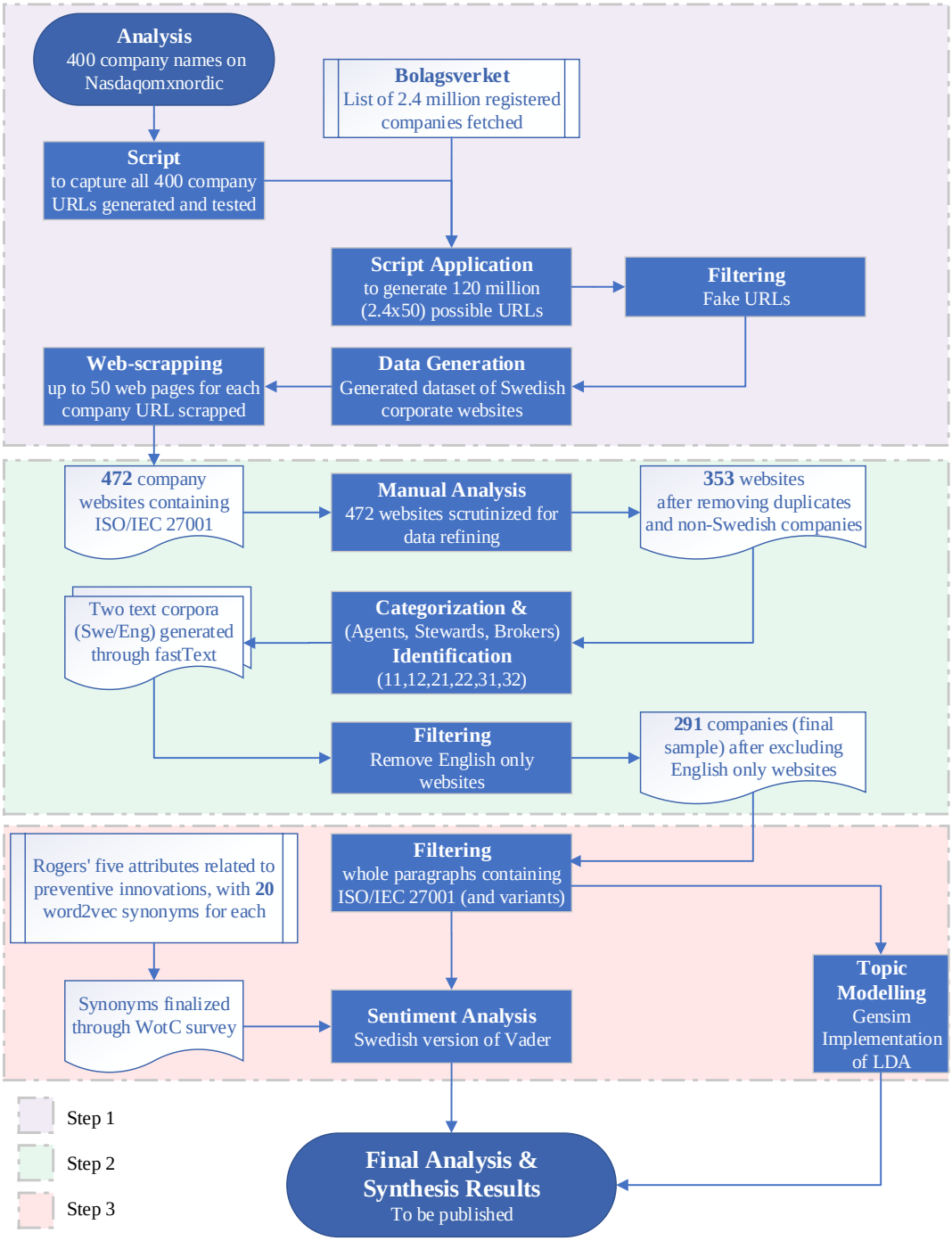


Figure 1: Overview of the process.

Adoption type	11	12	21	22	31	32
Number of companies	103	10	81	41	19	37
Number of sentences	197,131	11,225	127,880	29,404	20,462	82,390
Number of words	3,374,348	187,630	2,133,516	547,543	351,837	1,683,044
ISO paragraphs	520	88	401	133	17	372
Sentences in ISO paragraphs	8,356	1,153	4,404	2,248	561	38,817

Table 1: Descriptive statistics for the Swedish companies in each adoption type

the corpus can be seen in Figure 6. We translate all Swedish texts to English using googletrans[5], as a few management studies researchers are not fluent in Swedish. All texts are also parsed using the Sparv pipeline[6] (Borin et al., 2016) to give us lemmas, POS and word sense information, needed in our analyses.

Data from websites are very noisy containing repetitions, menu items, contact information, adverts, etc that need to be handled. Standard crawling packages provide some cleaning of the texts but there is still much that is, for instance, not syntactically correct. We identified more than 400 sentences with more than 250 words, which illustrates the noisiness of the data, in this case, the lack of proper sentence delimiters. We therefore filtered out sentences with more than 300 words and paragraphs with more than 500 sentences which resulted in a slightly smaller corpus, see Table 2.

Adoption type	11	12	21	22	31	32
Number of sentences	186,881	10,679	120,194	27,6624	19,246	77,482
Number of words	3,181,830	182,416	1,957,894	501,247	320,316	1,608,769
ISO paragraphs	373	77	203	79	10	168
Sentences in ISO paragraphs	5,943	1,088	3,028	1,620	349	16,533

Table 2: Descriptive statistics for the Swedish companies in each adoption type with sentences containing more than 300 words and paragraphs with more than 500 sentences removed

Further filtering can be done, cf. Martin et al. (2022), but despite this, we find that the SweClarin resources are robust and provide results that can be used in our analyses.

4 Analyses and reflections on usability

Content analysis on the texts was performed to demonstrate how preventive innovation is manifested within the communication of the six identified adoption approaches. To aid this analysis word clouds were created and a topic analysis was carried out.

To assess the meaning that organisations ascribe to preventive innovations along the five attributes (as suggested by Rogers (1995), see Cahpter 2), we use sentiment analysis. We want to compare the overall sentiment for each attribute and also compare the sentiment when ISO/IEC 27001 is presented.

4.1 Word clouds

Word clouds were employed as an initial step in our research to swiftly identify and visualise the most frequent words in the corpus generated through web crawling. The word clouds were created from the unprocessed scraped data, including pages unrelated to ISO/IEC 27001, accompanied by raw frequency data for each adoption type, to a holistic understanding of the content and gain insights into the overall discourse, beyond just their adherence to the ISO/IEC 27001 standard.

To build the word clouds we used the WordCloud package[7]. After filtering out stop words and a variety of other text strings, such as URLs and numbers, we used the 1000 most common words to build word

[5]https://pypi.org/project/googletrans

[6]https://spraakbanken.gu.se/sparv/#/sparv-pipeline

[7]https://pypi.org/project/wordcloud

clouds for all companies. We also built word clouds, from the same data, for each of the preventive innovation adoption types. Figure 2[8] depicts such word clouds from two different adoption types clearly showing that there are differences between the two types as well as similarities, e.g. *service, security* and *information* are important in both but also that, for instance, *risk* is more prevalent for type 22.

Figure 2: English word cloud examples, adoption type 21 left and 22 right

Word clouds were particularly beneficial for providing the management researchers with an accessible overview of the most frequent words in the corporate communication of Swedish companies about ISO27001. The visualisation of key terms and their prominence assisted in pinpointing the primary focus areas within the corpus. The word clouds served as a foundational tool, offering a snapshot of the language and terminology prevalenty used in the industry's conversations about ISO27001.

Their primary role was to orient the research process by providing a basic overview of the corpus. However, word clouds offer limited analytical depth, primarily showcasing word frequency without delving into the nuances of context, relationships, or patterns among the terms. As such, they were more of a stepping stone towards more sophisticated analysis methods rather than a source of conclusive insights. Therefore, while instrumental in the early stages of research, word clouds were ultimately superseded by more advanced analytical tools that offered deeper, more contextually rich insights suitable for the final presentation of our findings.

4.2 Topic analysis

Following the initial analysis with word clouds, topic modelling was introduced as a method to further dissect and understand the corpus. For topic modelling we use the Gensim implementation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

Topic modelling is done both on the whole text and on the paragraphs containing an ISO27001 related term, as presented in Section 3. Further filtering is done by applying a frequency threshold and a threshold for spread. For LDA each paragraph is treated as a document and we use the Sparv parsed version to only include content words, in our case words with one of the part-of-speech tags noun, adjective, verb and adverb. We trained multiple topic models with different hyperparameters and used the NPMI coherence measure (Röder et al., 2015) to find optimal topics for each adoption type. Finally we use pyLDAvis[9] to visualise the topics, see Figure 3.

We also experimented with interpreting the results from the topic analysis using ChatGPT-3.5. We used a very simple prompt, *Interprete the following topic model we internally name "indirect economic impact, grouped 112131"* and the raw data from the topic analysis, see Figure 4 which shows data from four of the twelve most frequent terms in Figure 3. The management researchers considered the results from ChatGPT more useful than the visualisations, see Figure 5.

This technique enabled us to extract abstract topics, revealing hidden thematic structures within the corporate communications of Swedish organisations on ISO27001. Topic modelling provided a more

[8]The examples in Figure 2 are from the English corpus. Used here only for illustration as we do not normally analyse the English texts.

[9]https://pypi.org/project/pyLDAvis/

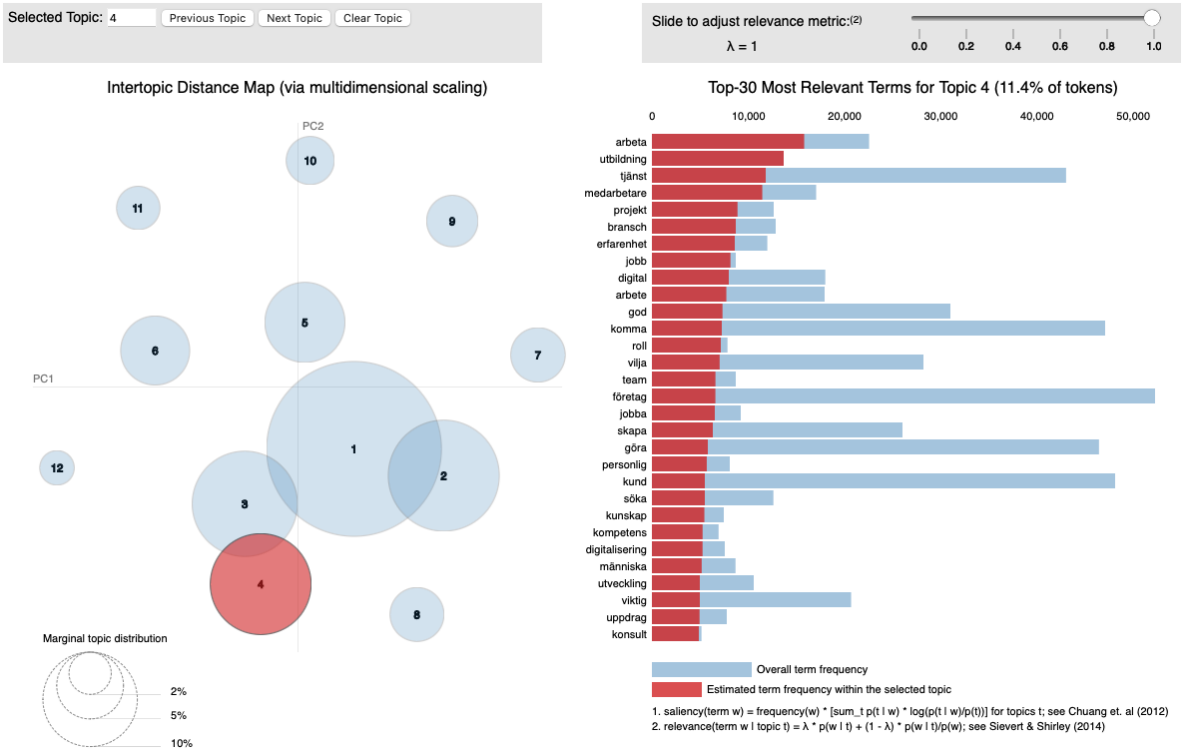


Figure 3: LDAvis visualisation of adoption type 11 for topic 4 (out of 12 topics as determined from the NPMI coherence measure). Translation of the terms from top to bottom: to work, education, service, coworker, project, sector, experience, job, digital, work, good, arrive/come, role, want, team, company, work, create, do, personal, customer, search, knowledge, competence, digitisation, human, development, important, mission and consultant.

```
    Interprete the following topic model we internally name "indirect economic
impact, grouped 112131":
LdaModel(num_terms=29019, num_topics=12, decay=0.5, chunksize=100)
[(0, '0.021*"arbeta" + 0.018*"utbildning" + 0.015*"tjänst" + 0.015*"medarbetare"
' '+ 0.012*"projekt" + 0.011*"bransch" + 0.011*"erfarenhet" + 0.011*"jobb" + '
'0.010*"digital" + 0.010*"arbete"'),
(1, '0.027*"ny" + 0.026*"år" + 0.017*"stor" + 0.017*"komma" + 0.016*"företag"
+ ' '0.015*"svensk" + 0.013*"kommun" + 0.009*"samarbete" + 0.008*"säga" + '
'0.008*"ta"'),
(2, '0.141*"kontakta" + 0.113*"följa" + 0.092*"meny" + 0.061*"söka" + '
'0.046*"beskrivning" + 0.041*"produkt" + 0.026*"ledig" + 0.024*"menya" + '
'0.023*"integritetspolicy" + 0.021*"recension"'),
(3, '0.091*"läsa" + 0.066*"språk" + 0.049*"tjänst" + 0.047*"innehåll" + '
'0.045*"välja" + 0.043*"kontakt" + 0.036*"sida" + 0.025*"nyhetsbrev" + '
'0.019*"mycken" + 0.018*"text"'),
.............
```

Figure 4: ChatGPT prompt

nuanced understanding of the corpus, uncovering the prevalent themes and areas of interest in the virtual discourse of these companies. It was especially effective in identifying commonalities across various industries, illustrating the overarching trends and concerns within the sector. This level of analysis was vital in exploring different data governance approaches – agency, stewardship, brokerage – as well as the direct and indirect economic impacts of ISO27001 on these organisations. Topic modelling thus served as a critical tool in deepening our understanding of the corpus and shaping the theoretical framework for our research. Like word clouds, topic modelling was invaluable in guiding our research trajectory and

```
Based on the provided topic model "indirect economic impact, grouped 112131," here
is an interpretation of the topics:

Topic 0: This topic seems to be related to work and education, with terms such
as "arbeta" (work), "utbildning" (education), "tjänst" (service), and "projekt"
(project) being prominent. It might indicate discussions about skills, experience,
and digital work.

Topic 1: This topic revolves around new developments and companies, including terms
like "ny" (new), "år" (year), "företag" (company), and "kommun" (municipality). It
suggests discussions about new ventures, collaborations, and possibly local economic
impact.

Topic 2: This topic focuses on contact information and following updates. It
includes terms like "kontakta" (contact), "följa" (follow), and "integritetspolicy"
(privacy policy). It might pertain to discussions about communication channels,
privacy regulations, and staying informed.

Topic 3: This topic relates to content consumption and selection, with terms like
"läsa" (read), "språk" (language), "innehåll" (content), and "välja" (choose) being
significant. It suggests discussions about reading materials, multilingual content,
and selecting relevant information.

............
```

Figure 5: ChatGPT result. Translations included in the original response from ChatGPT.

informing our analytical approach.

4.3 Sentiment analysis

To get a more targeted investigation into the attitudes and perceptions surrounding ISO27001 we used sentiment analysis along the five attributes: relative advantage, compatibility, complexity, trialability and observability, see Chapter 2. To capture various uses of the attributes, synonyms were generated for each attribute by using the Gensim package[10] (Řehůřek & Sojka, 2010). For each attribute we generated 20 synonyms using seeds, in Swedish, that reflected the various attributes. For three of the attributes, we generated a second set of synonyms using different seeds. The general applicability of these twenty computer-generated synonyms in the Swedish colloquial language was assessed through a wisdom-of-the-crowd (WotC) survey approach (Surowiecki, 2004). An online Microsoft Forms survey with these twenty synonyms was sent to eight native Swedish speaking innovation and entrepreneurship researchers at Linköping University to compile a final set of synonyms for the five attributes.

The paragraphs in the files containing 'ISO/IEC 27001', and its possible variants, were filtered out of each text to be used for sentiment analysis. We use the context in which an ISO/IEC 27001 sentence occurs, i.e. the whole paragraph, as it is scraped from the web. This filtering resulted in a considerably smaller number of paragraphs and the sentences within them (Table 1).

For sentiment analysis, we use a Swedish version of Vader (Hutton & Gilbert, 2014) that considers a word's sense. Vader is a lexicon and rule-based sentiment analyser. The lexicon in English Vader comprises 5500 lexical entries with sentiment scores between +5 and -5. We used the Swedish SenSALDO 0.2 sentiment lexicon (Rouces et al., 2019) with sentiment scores -1, 0 and +1. SenSALDO assigns different sentiment values to different senses of a word, for instance, the Swedish word *fara* can mean 'danger' or 'go (away)' where the former has a negative sentiment and the latter is neutral. SenSALDO 0.2 comprises 12287 lexical entries of which 8893 are unique words. Word sense disambiguation with the SenSALDO 0.2 lexicon is achieved by first parsing the texts using the Sparv pipeline.

Vader weights the sentiment and gives a sentiment score between -1 and 1, where $\leq$ -0,05 represents a negative sentiment, a score between $>$ -0,05 and $<$ 0,05 signifies a neutral sentiment and, a score $\geq$

[10] https://radimrehurek.com/gensim/

0,05 denotes a positive sentiment to each attribute. The higher the value the more positive the attitude, and the lower the value the more negative the attitude regarding a specific attribute.

Vader also uses booster words, such as *scarcely* (Swedish *knappast*), to further refine the sentiment analysis. The booster dictionary used in these analyses is an enhanced version of the Swedish dictionary used for sentiment analysis of e-mail conversations (Borg & Boldt, 2020) and comprises 89 items. The version used in this project, using the SweClarin SenSaldo resources, has also been used in a project on analysing Swedish official texts (Ahrenberg et al., 2022).

The mean sentiment score for each of the Roger's attributes is calculated from the sentences in paragraphs with ISO sentences containing words, and synonyms, related to the attribute, as presented above. All words generated as synonyms to the Roger's concepts, see above, are lemmatized using the Stanza pipeline (Qi et al., 2020)[11] for Swedish and can easily be matched to their occurrence in the text as the corpus is parsed using the Sparv pipeline, which includes the Stanza pipeline for Swedish lemmatization as one lemmatizer. Typical examples from the analyses can be seen in Figure 6[12].

Results from the sentiment analysis for each of the six adoption types along the five Rogers' attributes is presented in Table 3.

Table 3: Mean sentiment score for each adoption type and Roger's attribute

	11	12	21	22	31	32
Relative Advantage	0.172	0.111	0.128	0.174	0.000	0.092
Compatability	0.124	0.046	0.129	0.033	0.125	0.009
Complexity	0.156	0.135	0.110	0.077	0.122	0.091
Trialability	0.150	0.128	0.124	0.119	0.115	0.093
Observability	0.141	0.096	0.124	0.073	0.114	0.091

As can be seen in Table 3 the scores indicate that the texts are rather neutral, with adoption type 32 being more neutral than the other. Pairwise Welch t-tests for each category compared to the category's overall score also shows that this difference is significant for all Roger's attributes[13], i.e. paragraphs with ISO sentences comprising words related to any of the Roger's attributes for Broker companies that have direct economic benefits from communication adoption, adoption type 32, are more neutral.

The neutral sentiment example with Roger's attribute Relative Advantage, Table 3 is on the one hand a typical example of a neutral paragraph; it expresses no sentiment. But it also illustrates the nature of the corpus, in this case it is merely a number of statements or phrases, probably taken from a table of content or a menu on a web page. Nevertheless it has been parsed by the Sparv pipeline, showing the robustness of the SweCLARIN resources.

Table 3 further shows that no mean sentiment is negative. This is not surprising as the corpus contains companies web pages, pages where the companies present themselves.

However, as can be seen from Figure 6 even negative paragraphs are sometimes to be regarded as positive, i.e. the example from adoption type 32. That example also shows a problem with lexical sentiment analyses. When sentiment is based on averaging over words that have a positive or negative sentiment without taking context into account. Then sometimes phrases such as *increases the possibilities of identifying* in the example from adoption type 32 in Figure 6 that modifies the negative words *threats, risks, lost, stolen* are missed and the phrase is classified as negative when in fact is should be positive.

As this could potentially affect our results we have investigated the distribution of sentences with positive (≥ 0.05), negative (≤ -0.05), or neutral sentiment (between -0.05 and +0.05), as proposed by Hutton and Gilbert (2014), see Table 4.

As can be seen in Table4 the amount of sentences with positive, negative or neutral sentiment is equally distributed among the various adoption types. In each adoption type, the proportion of negative

[11] https://stanfordnlp.github.io/stanza/

[12] Translated by google translate

[13] $p < .001$ for all pairs.

Adoption type: 21
Roger's attribute: Trialability
Mean sentiment score: 0.612
Paragraph: Our NMT solution is developed in-house by our machine learning team of data scientists and engineers. Our neural machine translation is ISO 27001 compliant and all information is stored in our private data cente, so you can trust it to be secure.
Swedish original: *Vår NMT-lösning har utvecklats in-house av vårt maskininlärningsteam med datavetare och tekniker. Vår neurala maskinöversättning följer ISO 27001 och all information lagras i vårt privata datacenter, så att du kan lita på att den är säker.*

Adoption type: 32
Roger's attribute: Relative Advantage
Mean sentiment score: 0.0
Paragraph: Introduction to information security Different types of information security threats The requirements of ISO 27001 and interpretation of the requirements Exercises in interpreting and applying the standard in practice Good examples Culture and commitment in the workplace The red thread of the management system Tools for improvement Integration of systems
Swedish original: *Introduktion till informationssäkerhet Olika typer av informationssäkerhetshot Kraven i ISO 27001 och tolkning av kraven Övningar i att tolka och att tillämpa standarden i praktiken Goda exempel Kultur och engagemang på arbetsplatsen Ledningssystemets röda tråd Verktyg för förbättring Integrering av system*

Adoption typ: 32
Roger's attribute: Trialability
Mean sentiment score: -0.25
Paragraph: A certification according to ISO 27001:2017 ensures that you work in a systematic and effective way with information security. With the support of the standard you can create a framework for how you protect your most important information. Following a standard in your work with information security increases the possibilities of identifying threats and preventing risks of information being lost or stolen.
Swedish original: *Ett certifiering enligt ISO 27001:2017 säkerställer att ni arbetar på ett systematiskt och effektivt sätt med informationssäkerhet. Med stöd av standarden kan ni skapa ett ramverk för hur ni skyddar er viktigaste information. Att följa en standard i ert arbete med informationssäkerhet ökar möjligheterna att identifiera hot och förebygga risker för att information försvinner eller stjäls.*

Figure 6: Examples from the sentiment analyses

Table 4: Number of sentences with positive, negative or neutral sentiment for each adoption type

	11	12	21	22	31	32
POSITIVE	81,383	4,536	52,337	10,877	8,692	32,276
NEGATIVE	10,371	894	7,085	3,161	1,337	10,138
NEUTRAL	95,127	5,249	60,772	13,624	9,217	35,068

sentiment, relative to the total of positive and negative sentiments, ranges from 5-15%. Adoption type 11 exhibits the lowest proportion, with approximately 6% of sentences expressing negative sentiment in its entire corpus. In contrast, adoption type 32 has the highest, at around 15%. This indicates that companies generally present themselves in a positive light in their corporate presentations. We, thus, assume that the problem with sentences' sentiment is the same for all types, meaning that we can compare the sentiment of various adoption types but not necessarily give an absolute sentiment score.

5 Conclusions

In this study, we started with the idea of reviving the concept of preventive innovation given the attention this type of innovation is receiving nowadays. We have chosen to explore the adherence and adoption of the ISO/IEC 27001 information security standard as an example of preventive innovation addressing cyber security risks as one of the great challenges of our time. Using web scraping tools and a variety of computational linguistics tools, we were able to extract and analyse large amounts of text. These texts on preventive innovation ISO/IEC 27001 include communicative efforts published on the websites of companies operating in Sweden, telling us about the way these companies are adopting the standard. We have identified different adoption approaches and related modes of data governance. These results also help us understand that the original concept as introduced by Rogers (1995) need to be improved in terms of opportunities to derive economic benefits from preventive innovation. By relating the adoption approaches to the different modes of data it could be shown that a meaningful adoption of preventive innovations can already take place at an early stage.

The close cooperation between the management researchers and the SweClarin language technology researchers has been imperative for the success of this project. Based on the needs of the management researchers various analyses have been performed, and assessed. It was, for instance, initially assumed to be important to use word clouds to give an illustration of word frequencies amongst the various adoption types. Word clouds were good to get a coarse understanding of concepts used in the corpus and guided the management researchers' further investigations. Topic models were further used to guide the deep readings. However, they were also rather diverse and hard to interpret and did not form a clear characteristics of the various adoption types. The use of LLMs, in this case ChatGPT, turned out to be an interesting, and useful, complement to the classical visualisations of topic models.

In further discussions with the management studies researchers, we decided to try sentiment analysis, which turned out to give useful results on its own as well as the possibility to generate quotes from the texts for each sentiment ranked by its score. This gave management researchers the opportunity to see a quantification of the meanings to further aid the deep readings. These insights gave rise to new perspectives, which resulted in analyses along different dimensions of adoption type. Future research on sentiment analysis with the management researchers include looking into the actual sentiment analyses using new models. The Swedish BERT model (Hägglöf, 2023)[14] is one such possibility that needs to be assessed and compared to the lexical approach using the SweCLARIN SenSALDO 0.2 sentiment lexicon.

[14]https://huggingface.co/KBLab/robust-swedish-sentiment-multiclass

References

Ahrenberg, L., Holmer, D., Holmlid, S., & Jönsson, A. (2022). Analysing changes in official use of the design concept using sweclarin resources. *Proceedings of the 2022 CLARIN Annual Conference*.

Bergquist, W. (1993). *The postmodern organization: Mastering the art of irreversible change*. San Franscico: Jossey-Bass.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Borg, A., & Boldt, M. (2020). *Using vader sentiment and svm for predicting customer response sentiment, expert systems with applications* (Vol. 162). Elsevier.

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., & Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17-18 November, 2016*.

Downs, J., W., G., & Mohr, L. B. (1976). Conceptual issues in the study of innovation. *Administrative Science Quarterly!*, *21*(4), 700–714.

Fineberg, H. (2013). The paradox of disease prevention: Celebrated in principle, resisted in practice. *JAMA*, *310*(1), 85–90. https://doi.org/doi:10.1001/jama.2013.7518

Fried, A., Jönsson, A., Dragisic, S. P., & Bandyopadhyay, S. (2024). Preventive innovation revisited: Towards a sitational and meaning-based understanding of rogers' communication concept [Paper not published]. *84. Wissenschaftliche Jahrestagung des Verbandes der Hochschullehrerinnen und Hochschullehrer für Betriebswirtschaft 2024*.

Geissinger, A., Laurell, C., Öberg, C., & Sandström, C. (2023). Social media analytics for innovation management research: A systematic literature review and future research agenda. *Technovation*, *123*(102712).

Hägglöf, H. (2023). The kblab blog: A robust, multi-label sentiment classifier for swedish. https://kb-labb.github.io/posts/2023-06-16-a-robust-multi-label-sentiment-classifier-for-swedish/

Hutton, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Kinne, J., & Lenz, D. (2019). *Predicting innovative firms using web mining and deep learning* (tech. rep.). ZEW Centre for European Economic Research, Discussion Paper. 01/2019 (19-001). http://ftp.zew.de/pub/zew-docs/dp/dp19001.pdf

Lo, L. K., & Chang, D. S. (2007). The difference in the perceived benefits between firms that maintain iso certification and those that do not. *International Journal of Production Research*, *45*(8), 1881–1897.

Martin, L., Fan, A., de la Clergerie, É., Bordes, A., & Sagot, B. (2022, June). MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 1651–1664). European Language Resources Association. https://aclanthology.org/2022.lrec-1.176

Mirtsch, M., Kinne, J., & Blind, K. (2020). Exploring the adoption of the international information security management system standard iso/iec 27001: A web mining-based analysis. *IEEE Transactions on Engineering Management*, *68*(1), 87–100.

Orlikowski, W., & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Science Quarterly*, *39*, 541–574.

Putnam, L. L. (1999). Shifting metaphors of organizational communication: The rise of discourse perspectives. In P. Salem (Ed.), *Organizational communication and change* (pp. 45–65). Cresshill, NJ: Hampton Press.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora [http://is.muni.cz/publication/884893/en]. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*, 399–408. https://doi.org/10.1145/2684822.2685324

Rogers, E. M. (1995). *Diffusion of innovations*. The Free Press.

Rouces, J., Tahmasebi, N., Borin, L., & Eide, S. R. (2019). Sensaldo: Creating a sentiment lexicon for Swedish. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 4192–4198.

Saura, J. R., Palacios-Marqués, D., & Ribeiro-Soriano, D. (2023). Exploring the boundaries of open innovation: Evidence from social media mining. *Technovation*, *119*(102447).

Surowiecki, J. (2004). *The wisdom of crowds*. NY, NY: Anchor.

Taylor, J. R. (1993). *Rethinking the theory of organizational communication: How to read an organization*. Norwood, NJ: Ablex.

Tornatzky, L. G., & Klein, K. J. (1982). Innovation characteristics and innovation adoption-implementation: A meta-analysis of the findings. *IEEE Transactions on Engineering Management*, *EM-29*, 28–43.

Yates, J., & Orlikowski, W. (1992). Genres of organizational communication: A structurational approach to studying communication and media. *Academy of Management Review*, *12*(2), 299–326.

Mind the Ownership Gap?
Copyright in AI-generated Language Data

Paweł Kamocki
IDS Mannheim
Germany
kamocki@ids-mannheim.de

Toby Bond
Bird & Bird
London, UK
toby.bond@twobirds.com

Krister Lindén
University of Helsinki
Finland
krister.linden@helsinki.fi

Thomas Margoni
KU Leuven
Belgium
thomas.margoni@kuleuven.be

Aleksei Kelli
University of Tartu
Estonia
aleksei.kelli@ut.ee

Andrius Puksas
Vytautas Magnus University
Lithuania
andrius.puksas@vdu.lt

Abstract

For language scientists, a *prima facie* advantage of AI-generated data over human-created content is that AI outputs are generally regarded as free from copyright. This contribution addresses this issue in some detail.

1 Introduction

2023 was the year of the rabbit according to the lunar calendar, but in Europe it will likely be remembered as the year of Artificial Intelligence. It is safe to say that such events as the launch of Chat-GPT (in November 2022) or of GPT-4 have already revolutionized the way in which language data are generated. This revolution has not been unnoticed by the CLARIN community. The new perspective that AI opens up, is to create fully synthetic data according to the specifications of a researcher.

In branches of science where data for modeling is scarce, or access is limited by confidentiality or (usually copyright or data protection) laws, e.g., medical or behavioral sciences, the researchers can ask an AI model to generate new synthetic data for large categories, thereby avoiding the legal barriers. The model can also be used for creating more data for small categories to make the data more balanced and less biased. However, the bias reduction needs to be verified so that the additional data does more than just amplify the prejudice or bias in the original data.

Examples of synthetic data use in commercial research include Amazon using synthetic data to train Alexa's language system with available sample utterances as templates generating new data by combining and varying the templates. Google's Waymo uses synthetic visual data to extend its training data for self-driving cars with more complex but infrequent scenarios virtually adding more agents for the AI to cope with. American Express and J.P. Morgan generate statistically accurate synthetic data from financial transactions for more sophisticated fraud detection, and Roche uses validated synthetic medical data as a replacement for clinical research data to develop AI healthcare algorithms with massive amounts of personal health data, while minimizing privacy concerns[1].

[1] Types of synthetic data and 4 real-life examples (2022): https://www.statice.ai/post/types-synthetic-data-examples-real-life-examples (last access: 13.02.2024)

Pawel Kamocki, Toby Bond, Krister Lindén, Thomas Margoni, Aleksei Kelli and Andrius Puksas 2023. Mind the Ownership Gap? Copyright in AI-generated Language Data. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 102–113. DOI: https://doi.org/10.3384/ecp210

For scientists, a *prima facie* advantage of AI-generated data over human-created content is that, as it is generally agreed upon, AI outputs are not protected by copyright. This contribution addresses this issue in some detail.

The main reasons for the absence of copyright protection for AI-generated data is their lack of human authorship (Section 2), as well as – closely related – lack of originality (Section 3). However, the re-use of certain AI outputs may be in a legal grey area (Section 4). The introduction of a property right in AI outputs is seen by some as an answer to the challenges presented by the development of generative AI (Section 5), despite the fact that little evidence of this is found in the UK, where computer-generated works have been protected by a property right since 1988 (Section 6).

2 Lack of human authorship as an obstacle to copyright protection of AI outputs

The argument commonly used to refuse copyright protection of AI-generated content is lack of human authorship. The author is indeed placed at the very heart of modern copyright law, which was largely modeled after the French tradition of *droit d'auteur*, or *author's right*. The crucial role of the author in copyright law is illustrated by the fact that the author is the default holder of both economic and moral rights; moreover, the term of protection is also linked (at least in most cases) to the death of said author.

According to the general dictionary definition of the word "author", only human beings seem to be able to qualify as such. For example, the Cambridge Dictionary defines "author" as a 'a person who begins or creates something'; other dictionary definitions also seem to reserve this status to humans. But is this also the case in legal context? The question is worth asking, as legal texts often attribute specific meaning to everyday words.

The Berne Convention does not define "author", and does not expressly require human authorship for copyright-protected works. Nevertheless, upon closer inspection, this landmark international treaty is clearly based on the assumption that the author is a human being. For example, according to Article 3(2) "Authors who are not *nationals* of one of the countries of the Union but who have their *habitual residence* in one of them shall, for the purposes of this Convention, be assimilated to nationals of that country" (italics added by the authors). Moreover, Article 6bis provides that "the author shall have the right to (...) object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his *honor* or *reputation*", and adds that this right "shall, after [the author's] *death*, be maintained (...)". Finally, Article 7 defines the term of copyright protection as "*life* of the author and fifty years after his *death*", and Article 7bis further specifies that this term should be calculated "from the *death* of the last *surviving* author". Therefore, it appears clearly that under the Berne Convention only humans can be authors, as, unlike AI, they are mortal, have a nationality and a place of residence, as well as honor and reputation.

In EU law, the same conclusion can be drawn from the Copyright Term Directive (2006/116/EC), whose Articles 1 and 2 also refer to the author's death while defining the duration of copyright protection (which under EU law is longer than required by the Berne Convention, i.e. seventy years after the death of the author). Most, if not all, national laws[2] also contain similar provisions, tying copyright terms to the death of the author. Moreover, German Copyright Act (*Urheberrechtgesetz*) defines (in its §7) "author" as "the creator of a work", which also seems to reserve this status to humans.

Recently, the UK Supreme Court[3] ruled that Artificial Intelligence cannot be regarded as an "inventor" under patent law, and that only human beings can be "inventors". This, of course, is not directly related to copyright, but one can expect that if AI cannot be an "inventor", a fortiori it cannot be an "author".

In the US, the US Copyright Office (2023) also recognises that "copyright can protect only material that is the product of human creativity". In recent years, the Office has refused to register AI-generated images on the grounds of lack of human authorship: this was the case of a rather appealing image entitled *A recent entrance to paradise* (US Copyright Office (2022); *Fig. 1*) – the decision was later upheld by

[2] E.g., Article L. 123-1 of the French Intellectual Property Code, §64 of the German Copyright Act, §302 of the US Copyright Act (17 USC), s12 of the UK's Copyright, Designs and Patents Act, etc.

[3] Thaler (Appellant) v Comptroller-General of Patents, Designs and Trade Marks (Respondent) [2023] UKSC 49.

a District Court[4], who also found the image uncopyrightable and "absent any human involvement" – as well as a prize-winning image *Théâtre d'opéra spatial* (US Copyright Office (2023b); *Fig. 2*). The latter case is particularly interesting: the Copyright Review Board of the US Copyright Office found that the image, which for the most part was generated by Midjourney, lacked human authorship, and the applicant's input in generating it ('at least' 624 text prompts) was not sufficient to make him the author. Despite the fact that the modifications made by the applicant in Adobe Photoshop, if considered in isolation, might have qualified for copyright protection, the Midjourney-generated basis could not, and therefore the final image could not be registered as copyright-protected.

Figure 1: A recent entrance to paradise *(AI-generated)*

[4] US District Court of Columbia, 18.08.2023, Case 1:22-cv-01564-BAH, https://storage.courtlistener.com/recap/gov.uscourts.dcd.243956/gov.uscourts.dcd.243956.24.0_2.pdf (last access: 13.02.2024)

Figure 2: Théâtre d'opéra spatial *(AI-generated with human-made adjustments)*

At the same time, many copyright systems accept 'corporate ownership' of copyright, i.e. a situation where copyright is held *ab initio* by a legal person (a company, an employer) and not the human author. This is for example the case under the traditional anglo-saxon doctrine of *work for hire*, where copyright in a work created by an employee belongs *ex lege* to the employer[5]. Initial ownership of copyright by a corporation is therefore a well-established solution, which some praise as pragmatic and promoting investment (or even innovation). In fact, even in such an author-oriented copyright system as the French *droit d'auteur*, economic rights in a collective work (*oeuvre collective*) the creation of which is initiated and supervised by a legal person who then disseminates the work under its name[6] belong *ab inito* to the legal person, and not the actual human authors. In the field of software, Article 2 of the Directive 2009/24/EC on computer programs attributes the economic rights in software created by employees in the execution of their duties to the employer. The Article goes as far as to admit (in paragraph 1) that, where legislation of a Member State allows it, a legal person can be considered author of a computer program.

It appears, therefore, that many national laws, and even, to an extent, EU law, can tolerate a situation where initial ownership of copyright is attributed to a legal, and not a natural person. This, however, is not enough to solve the issue of AI-generated works, since AI in itself obviously has no legal personality, and attributing copyright to the company that provides a generative AI tool (such as Open AI, the provider of Chat GPT) would be a dubious solution to say the least.

3 Lack of originality as an obstacle to copyright protection of AI outputs

Another theoretical obstacle on the path to copyright protection that AI-generated works would have to face is the originality requirement.

Originality (in some copyright traditions, e.g. in Germany and in Poland, also referred to as "individuality") is the main condition for copyright protection. At the same time, it is a very elusive

[5] Cf., for example the definition of a "work made for hire" in § 101 of the US Copyright Act 1976, or s11(2) of the UK's Copyright, Designs and Patents Act of 1988.
[6] Cf. Article L. 113-2 para 3 of the French Intellectual Property Code.

concept, which for a long time was escaping any efforts toward international harmonisation. Briefly put, two approaches to originality can be distinguished: a subjective one, which emphasises the relation between the work and its author (a work is original if it carries a "personal mark" of the author) and an objective one, which focuses on elements such as skilled effort invested in the creation and novelty (absence of copy) of the resulting work.

Since CJEU's 2009 landmark decision in the *Infopaq* case[7], the EU subscribes to the subjective approach, even though it contradicts long-standing traditions of some Member States' national copyright laws. In the *Infopaq* case, the CJEU applied the definition of originality as "author's own intellectual creation", which was already present in EU law, to all copyright-protected works. Incidentally, this definition is also very close to the traditional German concept of "personal intellectual creation" (*persönliche geistige Schöpfung*)[8]. This was further elaborated in subsequent CJEU's decisions; most notably in *Painer*[9], where the Court ruled that the originality requirement is met "if the author was able to express his creative abilities in the production of the work by making free and creative choices". By making such choices, the author "stamps the work with his personal touch", so that the work "reflects his personality". At the same time, the CJEU formulated various "negative conditions" for originality, i.e. conditions that, if met, prevent copyright protection; these include situations where the expression of the work is dictated by technical considerations[10], or other rules that leave no room for creativity[11]. Moreover, the CJEU also clearly stated (in *Football Dataco Ltd*)[12] that labour and skill alone are not enough to justify copyright protection of the outcome.

It seems that autonomous AI outputs cannot meet the originality criterion as defined by the CJEU, as generative AI tools may not allow the user to make free and creative choices during the creative process, and to leave his or her "personal touch" in the work. This also seems to be the position of Advocate General Trstenjak, who in her opinion in the *Painer* case stated that only human creations can be original (in the sense of being their author's own intellectual creations) and therefore qualify for copyright protection[13]. Moreover, one could expect that mere "skill and labour" invested by the user in prompting the generative AI tool are not enough to confer originality to the output.

US copyright law has a somewhat lower (or at least: more objective) standard of originality. In order to qualify as original under US law, a work has to be independently created by the author (i.e., simply, not copied from another work) and possess a minimal degree (modicum) of creativity, a "creative spark", "no matter how crude, humble or obvious it might be" (US Copyright Office, 2021). Arguably, at least some AI-generated works may pass this test. The US Copyright Office, however, systematically refuses to register AI-generated outputs not because of their lack of originality, but because of their lack of human authorship (cf. above).

4 Grey areas related to copyright protection of AI outputs

Lacking both human authorship and (subjective) originality, AI outputs may seem safely beyond the scope of copyright protection. However, this statement is not uncontroversial, and there are circumstances where AI outputs may be argued to meet the requirements for protection.

Firstly, AI does not (yet) generate outputs autonomously; the generative process is always initiated by a human who prompts the application with an idea in their mind. At least according to the dictionary definition, this human initiator can still be referred to as 'author' ('a person who *begins* or creates something'), even though the actual expression of the work (protectable by copyright, unlike the initial idea) is generated (or at least assisted) by AI. The main obstacle to copyrightability of AI outputs may therefore lie not in the law, but in the way our culture perceives authorship – and this can evolve over time, like it did in the past (Compagno, 2012).

For decades now, copyright theorists have been distinguishing between machine-assisted and machine-generated outputs. While machine-generated works are not protected by copyright (for the reasons discussed above), machine- (computer-, AI-) assisted works are characterised by a sufficient

[7] CJEU, C-5/08, 16.07.2009 (*Infopaq*).
[8] §2(2) of the German Copyright Act.
[9] CJEU, C-145/10, 1.12.2011 (*Painer*).
[10] CJEU, C 393/09, 22.12.2010 (Bezpečnostní softwarová asociace).
[11] CJEU, joined cases C-403/08 and C-429/08, 4.10.2011 (*Football Association Premier League Ltd*)
[12] CJEU, C-604/10, 1.03.2012 (*Football Dataco Ltd.*)
[13] Opinion of Advocate General Trstenjak delivered on 12 April 2011 in Case C-145/10, para 121

degree of human intervention to qualify for copyright protection. A vast majority of works are in some way assisted by a machine, including this very article, whose creation involved modern text processing software with, among other features, an in-built automatic spellchecker. This, however, does not change the fact that the article is, by any standard, protectable by copyright.

Drawing a line between outputs with sufficient human involvement to 'deserve' copyright protection ('AI-assisted') and those without it ('AI-generated') is an extremely delicate task (cf. the 4-step test in Hugenholtz and Quintais, 2021), and courts' views on this issue are susceptible of changing over time. Such was the case with, e.g., photography, which was admitted in the realm of copyright several decades after the technology was popularized, and even today it is not recognized in the Berne convention as equal with other types of works (Art. 7(4) allows for a shorter term of protection for photographic works). In early decisions involving photographs[14] courts emphasized the role of the human photographer in, e.g., selecting the lighting, a task that is (or at least can be) fully automated in modern digital cameras, which does not seem to affect copyrightability of digital photographs (Margoni, 2014). AI outputs may follow the same trajectory, and the degree of human involvement required by courts for copyright protection may be gradually lowered. After all, since the beginning of time, almost all forms of human expression have employed some form of technology, be it very rudimentary.

In its recent policy statement, the US Copyright Office (2023a) also opted for a somewhat nuanced approach to registering AI-generated works. The key criterion seems to be whether the "traditional elements of authorship (literary, artistic, or musical expression or elements of selection, arrangement, etc.)" were "conceived and executed" by a man (assisted or not by a machine) or by a machine. In the Office's view (see above), merely prompting a machine is not enough to claim authorship in the output (no matter how elaborated or numerous the prompts, according to the Office they only function as "instructions to a commissioned artist", and the "traditional elements of authorship" are still executed by a machine). However, copyright can be claimed in cases where AI outputs are arranged by a human in a creative manner, or modified to a degree that meets the threshold of creativity.

This position was illustrated by the Office's recent decision regarding a comic book *Zarya of the Dawn* (US Copyright Office (2023c); Fig. 3) in which all images were generated by AI. The comic book as such (the plot, the dialogues) were deemed eligible for registration, although individual AI-generated images were excluded therefrom. However, the policy statement may seem inconsistent with the Office's decision concerning the image *Théâtre D'opéra Spatial* (Fig. 2, see above), which was also denied copyright protection. Although the image generated by Midjourney had been modified by the user, and the adjustments made might have been copyrightable on their own, the final result as submitted to the Office was not deemed eligible for copyright protection (Roose, 2022).

[14] See esp. Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53 (1884)

Figure 3: Panels from Zarya of the Dawn *(AI-generated images, texts by Kashtanova)*

Secondly, AI tools do not create outputs *ex nihilo*. Therefore, another gray area regarding copyright in AI outputs is linked to the relationship of these outputs with the data that were used to train the underlying model. Although in EU law the use of copyright-protected content to train AI models seems generally (under certain conditions) allowed under the exceptions for Text and Data Mining (Kamocki et al., 2018; Kelli et al., 2022), the copyright status of AI outputs remains rather unclear. Carlini et al. (2021) have shown that under certain conditions "training data extraction attacks" can be performed on GPT-2 which result in the model outputting text which resembles training material. The existence of

these techniques contributes to a lack of legal certainty regarding the copyright status of such outputs, especially considering that according to the CJEU excerpts as short as 11 consecutive words may be protected by copyright (Infopaq, C-5/08).

Even without regurgitating verbatim copies of training data, some (e.g., Gervais, 2022) have argued that AI outputs are derivatives, derived from the training material, which would also impact their copyright status. This lack of legal certainty is illustrated by recent US lawsuits; e.g., Getty Images sued Stability AI for allegedly using their images to train an AI model[15], and NY Times sued Open AI and Microsoft for allegedly using their articles for this purpose (O'Brien, 2024). A group of 17 authors, including John Grisham and George R. R. Martin, went as far as to sue Open AI for "systematic theft on a mass scale" (Italie, 2023). As a matter of fact, there are a number of other lawsuits brought by authors against AI companies (see, Setty, 2023) or content producers accused of using AI techniques (Khalid, 2024). According to recent media reports, a Chinese court found a provider of an AI text-to-image tool guilty of copyright infringement; the tool (when prompted accordingly) generated images of Ultraman, a popular cartoon character, that were substantially similar to the original artwork (Costigan, 2024).

The opinion according to which AI-generated outputs are in fact infringing copyright in the data used to train the underlying model remains to be tested by European and US courts, and rightholders have so far struggled to consistently identify outputs which bear a resemblance to items of training data without wilfully contriving circumstances intended to create such resemblance[16].

Finally, it has also happened that, for fear of a successful copyright infringement lawsuit, platforms removed AI-generated content when pressured by rightholders (Snapes, 2023). Such content was, therefore, assumed to infringe copyright.

5 Towards (Property) Rights in AI Outputs?

In February 2023 it was reported that ChatGPT is listed as author or co-author of over 200 books available on Amazon (Nolan, 2023). One can only imagine the number of books and other texts that were 'secretly' generated by AI and passed as human creations. As purely AI-generated texts are generally in the public domain, they can fall victim to 'copyfraud', i.e. a false copyright claim (e.g., by simply signing an AI-generated text with one's name, as a pretended human author).

In fact, the Berne Convention (Article 15(1)) provides that 'in order that the author of a literary or artistic work protected by this Convention shall, in the absence of proof to the contrary, be regarded as such, and consequently be entitled to institute infringement proceedings in the countries of the Union, it shall be sufficient for his name to appear on the work in the usual manner. This (…) shall be applicable even if this name is a pseudonym, where the pseudonym adopted by the author leaves no doubt as to his identity'. The same principle is repeated in the EU Directive 2004/48/EC on the enforcement of IP rights (Article 5). Both legal instruments establish a presumption of ownership for those whose name 'appear on the work in the usual manner'.

It seems, therefore, that it is enough for a user of a generative AI tool to sign his or her name on the output in order to benefit from a strong presumption of authorship, and become *de facto* enabled to sue others for copyright infringement. In this context, the act of signing automatically-generated content with one's name may appear controversial from the ethical standpoint, and it does constitute an act of copyfraud, but is not effectively punishable in the current state of the law.

One way out of this conundrum is the introduction of a transparency obligation, according to which all AI outputs would have to be clearly labeled as such. The proposed AI Act (European Commission, 2021) aims at addressing this issue in its Article 52, which on the one hand requires the providers of AI systems to design those systems in such a way as to inform users that they are interacting with AI, and on the other hand, obliges the users of image-, audio- or video-generating AI systems to disclose that the content resembling existing persons had been artificially generated. In the proposal, however (unlike in the French reform proposal discussed below), this obligation does not apply to AI-generated texts. Considering that such texts are practically indistinguishable from human-written ones (Casal & Kessler, 2023), such a requirement would meet serious evidence-related obstacles. It is conceivable to make the providers, and not the users, responsible for ensuring transparency of AI-generated text, e.g. by an

[15] Getty Images (US), Inc. *v.* Stability AI, Inc. (1:23-cv-00135).
[16] Cf. Open AI's rebuttal of NY Times' accusations: https://openai.com/blog/openai-and-journalism (last access: 13.02.2024)

imposed implementation of watermarking techniques. In the current state of the art, however, the feasibility of watermarking, particularly in shorter texts, seems doubtful.

If "enclosing" AI-generated content with a transparency obligation proves unworkable, other solutions to fill the perceived "void of ownership" (US Copyright Office, 2023b) would be to either extend the scope of copyright to include such works, or to create a new (property?) right to protect them.

Rather surprisingly, the authors of a proposal recently submitted to the French legislator opted for extending the scope of copyright. A copyright reform proposed in September 2023 (Assemblée Nationale, 2023) aims at introducing a series of rather revolutionary measures to protect the interests of creators against the influx of AI-generated creations. Firstly, this would include an express provision according to which rightholders' permission would be necessary to integrate a copyright-protected work "in an AI system", which at least *prima facie*, contradicts TDM exceptions (a part of EU *acquis*). Secondly, another new provision would state that copyright in content generated by AI without direct human intervention should belong to the authors of works that "enabled" the generation of the content. This not only contradicts the general lack of copyright in AI-generated works (due to lack of originality and human authorship), but is also extremely difficult to apply in practice, as it is rarely possible to determine a limited group of authors whose works "enabled" the AI system to generate a specific output; furthermore this may be regarded as a violation of EU law, as according to the CJEU originality (understood as "the author's own intellectual creation" – see above) is the only condition ("necessary and sufficient") for copyright protection[17]. Thirdly, the proposal also includes a system of collective rights management for AI-generated works. A designated collective rights management organisation would represent holders of rights in AI-generated works (i.e., according to the proposal, authors of works that AI used to generate content), perceive remuneration on their behalf and redistribute it among them. Fourthly, the proposal also aims at including a transparency obligation: all AI-generated outputs would not only have to be labeled as such, but also carry the names of all the authors that have enabled their creation (and who, therefore, would hold copyright in the work, as per the proposal). Especially in the case of longer AI-generated texts, this obligation is rather impossible to meet, because, among other reasons, it is often difficult to determine authorship in data obtained via web crawling. However, the French proposal also anticipates a situation where it is impossible to determine the *origin of works* (which, we believe, should be interpreted as authorship) that were used by an AI system; in such cases a levy (tax) would have to be paid to the abovementioned designated collective management organisation.

Although the proposed system of collective rights management and levies for AI-generated works may seem both controversial and impracticable, it has advocates among Europe's most renowned Intellectual Property scholars. In his recent article Senftleben (2023) argues that an output-oriented levy system, in contrast to remuneration for AI training activities, "does not weaken the position of the European AI sector and the attractiveness of the EU as a region for AI development. Even more importantly – Senftleben continues – an output-oriented AI levy system can be combined with mandatory collective rights management".

The creation of an entirely new exclusive right in AI outputs would be another possibility. As early as the 1960s it was argued (Demsetz, 1967) that technological progress will necessarily be accompanied by the creation of new property rights, mostly to guarantee legal certainty of transactions and to prevent market failure. Indeed, in the last decades new property rights have been created, such as the *sui generis* database right, or the right in computer-generated works in the UK (see below).

Already in 2020 the European Parliament took the view that AI-outputs 'must' be protected under Intellectual Property Rights in order to encourage investment and improve legal certainty, and called the Commission to reform EU law accordingly. Such statements from the Parliament should, however, be regarded as devoid of any legal meaning. However, in a recent response[18], the Commission stated that 'the issue of AI-generated works does not deserve a specific legislative intervention'. Moreover, many European IP scholars criticize the idea of introducing new property rights (Bulayenko et. al, 2022).

On the other hand, in recent years the Commission was active in proposing governance-based (as opposed to property-based) regimes for data, including AI-generated data. This follows an attempt to introduce a data producers' right (Gangjee, 2022). These regimes, introduced, e.g., by the Data

[17] CJEU, C-683/17, 12.09.2019 (*Cofemel*), para 30.
[18] https://www.europarl.europa.eu/doceo/document/E-9-2023-000479-ASW_EN.pdf (last access: 13.02.2024).

Governance Act or the Data Act, are focused on rights of users, enabling access and portability of data (that companies want to keep 'secret'), rather than on recognizing monopolies (property rights) in the data (Margoni & Kretschmer, 2022). This can be a novel approach to regulating AI, both at the input end (e.g., by recognizing 'artist data', distinct from copyright in literary, artistic and scientific works), and at the output end.

For now, the re-use of AI outputs is mostly regulated by contracts, especially Terms and Conditions of related online services, which tend to vary significantly. For example, Terms of Use of ChatGPT allow for the generated content to be reused for any purposes, including commercial ones ('such as sale or publication'), with an important exception: the use of ChatGPT outputs to develop models that compete with OpenAI is prohibited. A similar prohibition can be found in Bard's Terms of Service. Bing's Terms of Use for its consumer-focused product only allow for the generated content to be reused 'for personal and non-commercial purposes'.

It should be noted here that if the outputs of these applications are not protected by copyright, copyright exceptions, including the TDM exceptions, cannot apply to them, and so the above-mentioned Terms and Conditions cannot be overridden by such exceptions, as long as the contracts are enforceable.

Some language models, such as BERT or GPT-2, are also available under open source licences (Apache 2.0 and MIT, respectively), which impose no restrictions on the use of their outputs. However, more recent versions of GPT, starting from GPT-3, are publicly available only through a web API (i.e., subject to Terms and Conditions), and this trend is likely to continue with subsequent iterations of the most performant language models.

6 UK's Experience with Protection of Computer-generated Works

UK's Copyright, Designs and Patents Act of 1988 contains (since its adoption) a provision on computer-generated works (s9(3)). These works, defined as works 'generated by computer in circumstances such that there is no human author of the work', are protected by copyright (which, in the continental tradition, would be classified as a 'related' or 'neighbouring' right rather than copyright *stricto sensu*) for 50 years following their creation (s12(7)). The right belongs to 'the person by whom the arrangements necessary for the creation of the work are undertaken' (referred to as 'author'). Somewhat paradoxically, in order to qualify for protection, computer-generated works, like all other works, have to meet the criterion of originality (which historically was understood in the UK as involving a degree of 'labour, skill and judgement', but under the influence of the CJEU, a more author-centric approach to originality, presented above, was adopted). Similar provisions exist also in Ireland, New Zealand and South Africa.

Although it seems tempting to use this provision, adopted with the intention to regulate re-use of works such as satellite photographs, to AI-generated content, this has never been done by UK courts. In fact, case law involving this provision is extremely scarce, and the provision has been described as 'unclear and contradictory'. In a recent public consultation, the UK Intellectual Property Office listed computer-generated works as one of the issues to be addressed by the legislator. In its 2022 response, however, the government stated that, as there is no evidence that the provision is harmful, and 'any changes could have unintended consequences', especially given that the development of AI is still in its early stages. In the same statement, the government also declared that they will keep the provision under review and may remove, replace or amend it if the evidence supports this[19].

7 Conclusion

AI-generated outputs which do not involve human creative input should, in principle, remain copyright-free, as they cannot meet the traditional criteria for copyright protection. However, AI-generated outputs may bear strong similarities to copyright-protected works, which causes significant tensions between the interests of authors, generative AI tools providers, and users of such tools. These tensions resulted in a series of lawsuits, and in the coming years some landmark court decisions are expected both in Europe and in the US.

Meanwhile, copyright scholars and legislators are pondering the possibility of extending the scope of copyright, or even introducing a new related right, to balance the interests at stake. The result of these

[19] https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/artificial-intelligence-and-intellectual-property-copyright-and-patents (last access: 13.02.2024).

debates should not, or at least not before several years, affect the possibility for scholars to use synthetically generated data for their research.

However, one should not lose sight of the fact that generative AI tools are generally available via web APIs, governed by Terms and Conditions, which are likely to regulate the way the tool can be used and the allowed uses of the outputs; since these outputs are not protected by copyright, copyright exceptions (e.g., for research or TDM) do not apply.

We do live in interesting times, certainly for copyright scholars.

References

Assemblée Nationale (2023), Proposition de loi visant à encadrer l'intelligence artificielle par le droit d'auteur, No. 1630, https://www.assemblee-nationale.fr/dyn/16/textes/l16b1630_proposition-loi.

Bulayenko, O., Quintais, P. J., Gervais, D. & Poort, J. (2022). *AI Music Outputs: Challenges to the Copyright Legal Framework*. ReCreating Europe Report. https://doi.org/10.5281/zenodo.6405796

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A. & Raffel, C. (2021). Extracting Training Data from Large Language Models. *arXiv: 2012.07805*. https://doi.org/10.48550/arXiv.2012.07805

Casal, J. E. & Kessler, M (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing, *Research Methods in Applied Linguistics*, Volume 2, Issue 2023, https://doi.org/10.1016/j.rmal.2023.100068.

Compagno, D. (2012). Theories of Authorship and Intention in the Twentieth Century: An Overview. *Journal of Early Modern Studies*, 2012, 1 (1), pp.37-53. hal-01846362

Costigan, J. (2024). China Rules AI Firm Committed Copyright Infringement. *Forbes*, February 29, 2024. https://www.forbes.com/sites/johannacostigan/2024/02/29/china-rules-ai-firm-committed-copyright-infringement/

Demsetz, H. (1967). Toward a Theory of Property Rights. *The American Economic Review*, 57, 2, 347-359.

European Parliament. (2020). *Resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies* (2020/2015(INI))

European Commission (2021). *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (AI Act) and Amending Certain Union Legislative Acts*. COM(2021) 206 final.

Gervais, D. J. (2022). AI Derivatives: the Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines. *Seton Hall Law Review*, 53, 1111-1136. http://dx.doi.org/10.2139/ssrn.4022665.

Gangjee, D. S. (2022). The Data Producer's Right: An Instructive Obituary. [in:] Lim, E. & Morgan, P. (eds), *The Cambridge Handbook of Private Law and Artificial Intelligence*, Cambridge University Press.

Hugenholtz, P.B., & Quintais, J.P. (2021). Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output? *International Review of Intellectual Property and Competition Law*, 52, 1190–1216. https://doi.org/10.1007/s40319-021-01115-0

Italie, H. (2023). 'Game of Thrones' creator and other authors sue ChatGPT-maker OpenAI for copyright infringement. *Associated Press News*. https://apnews.com/article/openai-lawsuit-authors-grisham-george-rr-martin-37f9073ab67ab25b7e6b2975b2a63bfe

Kamocki, P., Ketzan, E., Wildgans, J. & Witt, A. (2018). New exceptions for Text and Data Mining and their possible impact on the CLARIN infrastructure. *Selected papers from the CLARIN Annual Conference 2018*

Kelli, A., Tavast, A., Lindén, K. (2022). Building a Chatbot: Challenges under Copyright and Data Protection Law. In: Martin Ebers, Cristina Poncibò, Mimi Zou (Ed.). *Contracting and Contract Law in the Age of Artificial Intelligence*. (115−134). Hart Publishing. http://dx.doi.org/10.5040/9781509950713.ch-007

Khalid, A. (2024). Amazon's Road House reboot is accused of copyright infringement — and AI voice cloning. *The Verge*, February 28, 2024. https://www.theverge.com/2024/2/27/24085264/amazon-road-house-reboot-lawsuit-ai-cloning-copyright-infringement

Margoni, T. (2014). The Digitisation of Cultural Heritage: Originality, Derivative Works and (Non) Original Photographs (December 3, 2014). http://dx.doi.org/10.2139/ssrn.2573104

Margoni, T. & Kretschmer, M. (2022). A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. *GRUR International*, 71(8), 685–701. http://dx.doi.org/10.2139/ssrn.3886695

Nolan, B. (2023). More than 200 books in Amazon's bookstore have ChatGPT listed as an author or coauthor. *Business Insider*, February 23, 2023. https://www.businessinsider.com/chatgpt-ai-write-author-200-books-amazon-2023

O'Brien, M. (2024). ChatGPT-maker braces for fight with New York Times and authors on 'fair use' of copyrighted works, Associated Press News, https://apnews.com/article/openai-new-york-times-chatgpt-lawsuit-grisham-nyt-69f78c404ace42c0070fdfb9dd4caeb7.

Roose, K. (2022). An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

Senftleben, Martin, Generative AI and Author Remuneration (2023). International Review of Intellectual Property and Competition Law (IIC) 54 (2023), Forthcoming, Available at SSRN: https://ssrn.com/abstract=4478370 or http://dx.doi.org/10.2139/ssrn.4478370

Setty, R. (2023). Sarah Silverman, Authors Hit OpenAI, Meta With Copyright Suits. https://news.bloomberglaw.com/ip-law/sarah-silverman-authors-hit-openai-meta-with-copyright-suits

Snapes, L. (2023). AI song featuring fake Drake and Weeknd vocals pulled from streaming services. The Guardian. https://www.theguardian.com/music/2023/apr/18/ai-song-featuring-fake-drake-and-weeknd-vocals-pulled-from-streaming-services

US Copyright Office (2021). Copyrightable Authorship: What Can Be Registered, https://www.copyright.gov/comp3/chap300/ch300-copyrightable-authorship.pdf.

US Copyright Office, Copyright Review Board (2022). Second Request for Reconsideration for Refusal to Register A Recent Entrance to Paradise. https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf

US Copyright Office. (2023a). *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence*. 16190 Federal Register, vol. 88, no. 51, 37 CFR Part 202.

US Copyright Office, Copyright Review Board (2023b). Second Request for Reconsideration for Refusal to Register Théâtre D'opéra Spatial. https://copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf

US Copyright Office (2023c). Zarya of the Dawn (Registration # VAu001480196). https://www.copyright.gov/docs/zarya-of-the-dawn.pdf

The SSH Open Marketplace and CLARIN

Alexander König
CLARIN ERIC
alex@clarin.eu

Laure Barbot
DARIAH-EU
laure.barbot@dariah.eu

Cristina Grisot
CLARIN-CH
DARIAH-CH
cristina.grisot@uzh.ch

Michael Kurzmeier
Austrian Centre for Digital Humanities
michael.kurzmeier@oeaw.ac.at

Edward J. Gray
DARIAH-EU
IR* Huma-Num
edward.gray@dariah.eu

Abstract

This paper showcases the SSH Open Marketplace, which is a discovery portal which pools and contextualises resources for Social Sciences and Humanities research communities, and its tight connections with the CLARIN infrastructure. The proposal presents how the SSH Open Marketplace can provide insights into the use of tools, methods and standards in the Social Sciences and Humanities communities in general, and for the CLARIN community in particular. The paper also describes how the SSH Open Marketplace can increase serendipity in the discovery of new methods and standards, by interlinking the resources and describing workflows. As contextualisation is provided between the items of the catalogue, it is easy to understand and assess the usefulness of a resource.

1 Introduction

In the context of Open Science, infrastructures, catalogues and discovery portals play a crucial role for enabling open research data and for increasing the degree of FAIRness (findability, accessibility, interoperability and reusability) of research data. The Social Sciences and Humanities Open Marketplace (SSH Open Marketplace) - marketplace.sshopencloud.eu - is a discovery portal which pools and contextualises resources for Social Sciences and Humanities research communities: **tools, services, training materials, datasets, publications and workflows**. The SSH Open Marketplace showcases solutions and research practices for every step of the research data life cycle. In doing so, it facilitates discoverability and findability of research services and products that are essential to enable sharing and re-use of workflows and methodologies.

The SSH Open Marketplace, conceptualized and implemented during the Social Sciences and Humanities Open Cloud (SSHOC) project [1], is one of the pieces of the bigger puzzle called the European Open Science Cloud (EOSC)[2]. The vision for EOSC is to create an environment for hosting and processing research data to support EU science, which provides seamless, Europe-wide access to research data and tools across scientific or thematic disciplines and geographical borders. The SSHOC project, along with the other four thematic clusters, namely ESCAPE (astronomy and particle physics), ENVRI (environmental sciences), panosc (materials, health, energy, physics) and EOSC-Life (life sciences)[3], supported the integration and consolidation of thematic e-infrastructure platforms in preparation for connecting them to the EOSC.

In this vein, the overall objective of the SSHOC project was to realise the Social Sciences and Humanities component of EOSC. As a domain-oriented discovery portal and the aggregator of the SSHOC project, the SSH Open Marketplace, contributes directly to the EOSC, supplementing existing services

[1] See the SSHOC project description: https://cordis.europa.eu/project/id/823782

[2] See EOSC description on the European Commission Directorate-General for Research and Innovation website: https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en

[3] See EOSC portal about the five ESFRI cluster projects in the EOSC panorama: https://eosc-portal.eu/news-and-events/news/five-new-esfri-cluster-projects-eosc-panorama

Alexander König, Laure Barbot, Cristina Grisot, Michael Kurzmeier and Edward J. Gray 2023. The SSH Open Marketplace and CLARIN. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 114–125. DOI: https://doi.org/10.3384/ecp210

such as the EOSC Catalogue and Marketplace, and facilitating the fluid exchange of tools, services, data, and knowledge. As a continuation of the SSHOC project and to sustain its outputs, 5 ESFRI Landmarks, CESSDA, CLARIN, DARIAH, ESS and SHARE, have signed a Memorandum of Understanding for the establishment of the **SSH Open Cluster**, and were later joined by 14 other national or European institutions and/or research infrastructures. All ESFRI projects and landmarks from the Social and Cultural Innovation domains [4] are currently members of the SSH Open Cluster. This cluster acts as an umbrella for the SSH Open Marketplace organisation and activities. More generally, the collaboration between the SSH Open Marketplace stakeholders (funders, providers, moderators or contributors) ensures that these cataloguing and contextualising efforts are meaningful, notably because they are undertaken by and serve humanities researchers.

The SSH Open Marketplace is one of the 33 Key Exploitable Results of the SSHOC project, and CLARIN, DARIAH and CESSDA decided to ensure the sustainability of the service after the end of the project. They act as a Governing Board for the SSH Open Marketplace and define the Marketplace strategic policy with regards to scientific, technical and managerial matters. In that context, two institutions act as service providers on behalf of these ERICs: the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)[5] of the Austrian Academy of Sciences providing hosting and maintenance for the service; and the Poznan Supercomputing and Networking Center (PSNC)[6], affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences, providing the data ingestion pipeline as well as maintenance for the service. The SSH Open Marketplace can also count on an Editorial Board, composed of 17 members[7], to ensure the day-to-day maintenance and (meta)data quality. Liaising with service providers and the end-users of the service (SSH researchers and support staff for researchers), the Editorial Board ensures the technical running of operation, the effectiveness of the curation process and the editorial policy's successful implementation.[8].

In sum, CLARIN has been heavily involved in the SSHOC project and is a founding partner for the continuation of the project in the form of the SSH Open Cluster. The SSH Open Marketplace is one of the key elements of research empowerment and discovery with which CLARIN is concerned. The special focus on contextualistion of the resources in the Marketplace can act as a complementary discovery tool to CLARIN's Virtual Language Observatory (VLO) - which includes a much larger number of items, but presents a lot less context - and the CLARIN Resource Families - which contain a much smaller number of items, but therefore can be even more extensively curated and contextualised.

2 Presentation of the SSH Open Marketplace

Initiated in the Digital Humanities (DH) context, inspired by the DiRT directory (Dombrowski, 2014), TAPoR[9] or the Standardization Survival Kit (Riondet & Romary, 2018) and aggregating their data for its initial population, the SSH Open Marketplace now acts as one of the thematic entry doors into EOSC. The SSH Open Marketplace was influenced by Dombrowski's "directory paradox", according to which DH tool registries should be community-led despite the organisational and infrastructural challenges it brings to such projects. (Dombrowski, 2021) This paradox is the foundation of one of the three Guiding principles which govern the SSH Open Marketplace, namely *Community* (see next section).

2.1 Guiding principles

While planning and building the SSH Open Marketplace three main pillars were identified, and these remain essential for its ongoing operation and future development. These pillars are:
Curation - The service thrives on a curation process that makes it easy to discover the most appropriate

[4] See the European Strategy Forum on Research Infrastructures website: https://roadmap2021.esfri.eu/projects-and-landmarks/

[5] see ACDH-CH website: https://www.oeaw.ac.at/acdh/acdh-ch-home

[6] see PSNC website: https://www.psnc.pl/

[7] see this page on the Marketplace website: https://marketplace.sshopencloud.eu/about/team

[8] For a detailed version of the sustainability plan, the report on Marketplace governance (Petitfils et al., 2021) can be consulted

[9] TAPoR 3: https://tapor.ca/. The Text Analysis Portal for Research is a project led by Geoffrey Rockwell and Milena Radzikowska.

and up-to-date results for each request, so that researchers can discover the best resources for the digital aspects of their work. The curation process relies on three components: automatic ingest and update of data sources; continuous curation of the information by the editorial team and – most important – contributions from users, the SSH research community.

Community – The content available in the SSH Open Marketplace and its contextualisation is the result of collaborative work that is characterised by a user-centric approach. Features that allow contributions are implemented to ensure that the portal mirrors real research practices.

Contextualisation – The portal puts all items into context: each solution suggested is linked to other related resources (e.g. a tutorial showing how to use a tool, a tool used in a workflow, a publication presenting research results produced using a given service). This contextualisation enhances the usefulness of the SSH Open Marketplace by showing how all these parts of the research process intertwine, and ensures users receive the maximum possible benefit from all its contents.

These three guiding principles describe the essence of the SSH Open Marketplace and contribute to increasing its usefulness for the target research communities and its sustainability in time.

2.2 Inclusion criteria

In order to guide users who wish to add resources to the SSH Open Marketplace, the following inclusion criteria and related guiding questions are enforced:

The relevance of the resource. The question to ask is: *will this resource be relevant to the SSH scientific community?* Thus, to be selected, any resource must fulfil at least two criteria: (1) scientific relevance and usefulness for SSH research and researchers and (2) pertinence to the digital methodologies used within the SSH landscape.

The technical status of the resource. The question to ask is: *is the resource current, supported, and ideally open?* The SSH Open Marketplace favours the uptake of Open Science workflows and open research practices. Software resources are preferably built upon open source solutions. Nonetheless, given that the SSH Open Marketplace seeks to mirror actual research practices, commercial or non-current resources are also referenced where these are relevant for the scientific community.

The degree of compliance of the resource with Open Science requirements. The question to ask is: *is the resource FAIR – Findable, Accessible, Interoperable and Re-usable - or contributing to the uptake of Open Science best practices?* The SSH Open Marketplace maximises the findability and re-use of data, and guides users towards tools, services or training materials that can help them in their FAIRification of workflows [10].

The uniqueness of the resource. The question to ask is: *is the resource already in the Marketplace?* If yes, there is no need to add it again, either as an individual item or with a source. Users are invited to enrich these existing items instead. However, when duplicates exist they can be identified as part of the curation activities (see Moderation and Curation section below). Currently, automatic checks of duplicates based on the name and the URL of the resources are regularly performed to merge the identified records. More parameters could be used in the future to reinforce the ”uniqueness of the resource” criterion.

Thanks to these inclusion criteria, the quality and the relevance of the resources added on the SSH Open Marketplace are guaranteed. This is an essential advantage for researchers who use the SSH Open Marketplace to discover resources which originate not only in their discipline but also from outside their own discipline. For example, a scholar who studies history and who uses digital methods to examine old documents, can easily discover on the SSH Open Marketplace the *Jupyter notebooks for Europeana newspaper text resource processing with CLARIN NLP tools*. [11] This training material presents, in a Jupyter notebook format, a three-chapter tutorial - (1) XML and CMDI introduction, (2) Data selection and resource access, (3) NLP processing - allowing interested trainees to interact with the Europeana

[10] While the FAIR principles are very well suited to data, using them to evaluate software, services or other resource types does not make much sense. This is why these principles are mentioned here as criteria alongside other dimensions of Open Science, in order to encompass all resource types covered by the SSH Open Marketplace.

[11] Jupyter notebooks for Europeana newspaper text resource processing with CLARIN NLP tools. Version 1 Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/training-material/duVII1

newspaper collection using, for example, Named Entity Recognition tools from the CLARIN environment.

A lot of interesting resources are created to support SSH researchers in the digital aspects of their work. The multiplicity of resources in a fast-changing environment makes it sometimes difficult to keep track of the most relevant ones. The inclusion criteria are here to provide a generic framework for the items selected and published in the SSH Open Marketplace, and make sure that pertinent resources are populating the platform.

2.3 Item types

Taking into consideration user requirements and developments of the EOSC data model, 5 main content types have been identified to structure the SSH Open Marketplace resources (Barbot et al., 2021). These 5 item types are considered to be representative for the large array of digital resources that can be found on this discovery platform.

Tools and services which refer to services and products, such as software, applications, programs, websites, programming libraries and APIs. The trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files *UDPipe*[12] is an example of a tool provided by CLARIN.

Training materials are tutorials, lessons or didactic resources explaining how to perform an action or highlighting the potential learning outcomes gained from using that material. For example, the *CLARIN Hands-on Tutorial on Transcribing Interview Data* [13]focuses on the role of automatic speech recognition – what are the opportunities, what are the pitfalls and where can it be applied successfully.

Workflows are sequences of steps that one can perform on research data during their lifecycle. Workflows can be created by using diverse tools, resources and methods, and useful resources are connected to each step. For example, *Intertextuality phenomena in European drama history* [14]is a workflow composed of 4 steps useful for analysing the relationships between the characters in a drama based on monologue/dialogue.

Datasets are defined as an organised collection of data. They are generally associated with a unique body of work, typically covering one topic at a time and are treated as a single unit by a computer. The SSH Open Markeptlace indexes CLARIN Resource Families datasets, for example the *DK-CLARIN Reference Corpus of General Danish* [15]

Publications are defined as research results published in academic journals or non-peer-reviewed publication repositories such as Zenodo. The SSH Open Marketplace references only publications that can be connected to other resources (i.e. tools and services, training materials, workflows or datasets). For example, you can find a paper on *Using TEI, CMDI and ISOcat in CLARIN-DK*[16] or the *Dublin Core Metadata Schemas* [17] on the SSH Open Marketplace.

Despite the fact that these five content types are equally (re)presented in both the SSH Open Marketplace data model (Ďurčo et al., 2021) and the front-end of the service, metadata quality for some of the types and user interactions these last two years have led to a focus on tools and services, and how they can be contextualised thanks to the other content types, and as part of workflows. Training materials, datasets and publications are used and considered in relation to other items, while tools and services are more and more seen as the primary content type or ”first class citizen” of the SSH Open Marketplace. For example, when interacting with users during hands-on sessions to create or enrich items, one of the main questions is often ”which tool do you use to perform your research?”. This then leads to an elaboration

[12]UDPipe. Retrieved Apr 27, 2023 from https://marketplace.sshopencloud.eu/tool-or-service/F7K42P

[13]SSHOC Webinar: CLARIN Hands-on Tutorial on Transcribing Interview Data. Retrieved Apr 27, 2023 from https://marketplace.sshopencloud.eu/training-material/ITNpCC

[14]Intertextuality phenomena in European drama history. Retrieved Apr 27, 2023 from https://marketplace.sshopencloud.eu/workflow/DMJlzG

[15]DK-CLARIN Reference Corpus of General Danish. Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/dataset/XR876U

[16]Dorte Haltrup Hansen, Lene Offersgaard, Sussi Olsen (2022): Using TEI, CMDI and ISOcat in CLARIN-DK. Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/publication/4jQvZ5

[17]DCMI Schemas. Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/publication/6kYac0

on the context of use, giving shape to the creation or enrichment of the most relevant resources in the catalogue.

2.4 Moderation and Curation

With a population of approximately 7000 items, aggregated from more than 15 trusted sources, the SSH Open Marketplace relies on community curation - i.e. contributions from the research communities in SSH and from the Editorial Board - to ensure the catalogue entries remain up-to-date and useful for SSH researchers, the end-users of the portal.

Contributions from the research community can take the form of creation of new items or enrichment of existing ones. In both cases, contributors suggest changes that are then passed on to the moderators, i.e. the members of the Editorial Board, who accept or reject the suggestions. Rejection usually goes hand in hand with contacting the contributor and asking further questions or suggesting options to revisit the initial approach. Moderating is not reviewing, and this is why, especially in the case of workflow moderation, the checks performed are limited to editorial control rather than peer reviewing.

Furthermore, curation routines, mixing automatic and manual tasks, are set up to ensure and continuously improve (meta)data quality. Indeed, in order to gain an overview of the SSH Open Marketplace data and to perform some analysis to prioritise the curation tasks and improve the Marketplace data quality, a Python library and a set of Jupyter notebooks have been created[18]. The flexible scripts allow moderators and administrators to query the SSH Open Marketplace with advanced parameters and filters and, in some cases, to write back to the system to flag some items for curation in the editorial dashboard.

Close to the inclusion criteria, a set of quality criteria have been established - general entry requirements; non-redundancy; completeness of item description; verification of conformity and relevance; interlinking - to guide the improvement of the metadata quality[19]. Based on these criteria, quality metrics have been derived and are used as a basis for the checks performed via the notebooks. In practice, the curation tasks performed these last two years revolved around data monitoring, bug fixing and data enrichment. Particular attention has been paid to underlying elements such as the (controlled) vocabularies[20], actor curation[21] or relations between items. For instance, Editorial Board members have worked on the consolidation of the *keyword* vocabulary, an open vocabulary (or 'folksonomy') in which users are allowed to add candidate concepts, and to which a wide variety of metadata from Marketplace sources have also been mapped, resulting in a 2000-concepts vocabulary. The curation exercise consists, in this case, in automatically mapping concepts belonging to other existing vocabularies in the Marketplace, or merging values when they are variations of the same concept. The work on the keyword vocabulary has also led to fruitful exchanges with other catalogues from the SSH domains as to how these services deal with "topical vocabularies". [22]

An important aspect to highlight here is that the manual curation work could not be done without the help and expertise of the members of the Editorial Board. As the SSH Open Marketplace covers a range of disciplines, the Editorial Board needs to mirror this diversity to be able to appropriately assess and make decisions on discipline-specific issues such as keywords, time periods or intended audience. This is why the approach of mixing automatic and manual curation is seen as a powerful one, that has proven its efficiency for a domain-oriented catalogue such as the SSH Open Marketplace in which rich metadata is essential to ensure findability of the resources.

[18] This library and the set of notebooks have been created by Cesare Concordia (CNR-ISTI) and are available at: https://github.com/SSHOC/marketplace-curation

[19] see SSH Open Marketplace moderator guidelines:https://marketplace.sshopencloud.eu/contribute/moderator-guidelines

[20] the SSH Open Marketplace currently counts 13 vocabularies, see:https://marketplace.sshopencloud.eu/contribute/metadata-guidelines

[21] In the Marketplace an Actor is the entity representing persons or institutions involved in the creation or maintenance of a resource. The SSH Open Marketplace counts around 7000 actors.

[22] see the TRIPLE project event *Use of vocabularies for metadata curation and quality assessment in Social Sciences and Humanities.* https://campus.dariah.eu/resource/events/use-of-vocabularies-for-metadata-curation-and-quality-assessment-in-social-sciences-and-humanities

3 The SSH Open Marketplace and CLARIN

3.1 CLARIN resources within the SSH Open Marketplace

The SSH Open Marketplace has been populated from a wide variety of sources, among which two come from the CLARIN world. The linguistic tools from the Language Resource Switchboard (LRS) (Zinn, 2018) and the tools, corpora and lexical resources collected in the CLARIN Resource Families (CRF) (Fišer et al., 2018). In both cases the original metadata has been mapped to the Marketplace Data Model (Ďurčo et al., 2021). As both the LRS and the CRF are very active, which means that items are constantly being added or updated (and in some cases also removed), the SSH Open Marketplace team has decided for a continuous ingest, i.e. to regularly re-harvest them to reflect changes at the source in the Marketplace.

Among the advantages of having the CLARIN resources listed in the SSH Open Marketplace is their increased discoverability by scholars from the SSH field who are not used to working with language data, who have not considered using language data in their research or who are not aware of the fact that CLARIN is an infrastructure which offers data and tools to support research that goes well beyond the linguistic domain. Indeed, language as cultural and social data is of interest for scholars from numerous other SSH disciplines. A historian, for instance, may use the SSH Open Marketplace discovery platform to find, in one search, language resources and tools (from CLARIN), digital resources from arts and humanities (from DARIAH) and social sciences resources (from CESSDA). Leveraging the SSH Open Cluster position in the EOSC to push these resources, useful beyond their initial target user group, to other audiences while maintaining the quality and richness of the accompanying metadata, even when processed through aggregation pipeline(s) is a challenge *per se* on which we elaborate in the last section of this paper.

Another important type of resource present on the SSH Open Marketplace is that of training materials. To increase their discoverability, such materials are being added to the SSH Open Marketplace, either manually or via the SSH Training Discovery Toolkit[23], which is also a source for the Marketplace. Giving space to and increasing the visibility of training and education is essential when it comes to increasing the access to open educational resources on various topics, and this focus is aligned to lines of action already present in the underlying ERICs, such as CLARIN and DARIAH. In what concerns CLARIN, it has recently strengthened its focus on training, especially in the context of the UPSKILLS project[24], which resulted in the creation of an important number of training resources, the creation of the CLARIN Learning Hub[25], and the creation, in collaboration with DARIAH, of the DH Course Registry[26].

3.2 Metadata and Interoperability

Metadata in the SSH Open Marketplace is subject to a process of mediation between the data available from the ingested source and the data requirements of the Marketplace users. The goal of this process ideally is to represent all important metadata taken from the ingested source while still maintaining a universal metadata structure in the Marketplace.

To illustrate, a tool from a source such as the CLARIN Language Resource Switchboard [27] can only be represented in the Marketplace if the input and output data formats can be recorded in the metadata describing the tool. While this is easy to achieve on a technical level, the universal approach on the Marketplace adds some additional steps to this process. For one, the chosen approach should be universally applicable to future ingests as well as manual curation of existing items. This likely means to create a controlled vocabulary from the data formats, so that different descriptions of the same format (for example, *XML file* and *TEI file* may be incorrectly used to describe the same format) can be unambiguously mapped to the canonical one. This new metadata then needs to be added as a metadata field in the back end and as searchable fields in the Marketplace front end. Additionally, the curation guidelines need

[23] see https://training-toolkit.sshopencloud.eu/entities?search=clarin
[24] see https://upskillsproject.eu/
[25] see https://www.clarin.eu/content/learning-hub
[26] see https://dhcr.clarin-dariah.eu/
[27] see https://switchboard.clarin.eu/

Figure 1: SSHOMP item detail with metadata

to reflect this new field and the user interface for manual data entry must highlight this important new attribute.

The described steps are to ensure that the new metadata field for data in- and output can be applied across the Marketplace. This in turn means employing resources to create the controlled vocabulary, changing the back and front end and adding additional steps to the item curation process. The example of the CLARIN Language Resource Switchboard serves to illustrate how the SSH Open Marketplace metadata structure can be adapted to different sources, however this process must always be undertaken with a view towards the generalist approach of the Marketplace.

Following the general process outlined in this example, the metadata schema of the SSH Open Marketplace can be adapted to accommodate new data sources as well as new types of interactions. In the example of the CLARIN Language Resource Switchboard, these new interactions are toolchains which enable multiple tools to be connected through shared data formats. In other cases, such as the SSHOC Conversion Hub [28], these interaction are conversions. Both require unambiguous metadata regarding the data input and output formats.

As can be seen in the above figure, metadata records in the SSH Open Marketplace are partially visible in the user interface, where they help to contextualize items. In the above example, all visible metadata fields except *Terms of Use* are clickable links through which the user can find other tools with the same attributes. This example further shows how important metadata records are for creating interoperability between items. Especially since the Marketplace combines different sources, universal metadata records create higher interoperability between items from different sources.

Being able to adapt the metadata structure of the Marketplace is essential to be able to represent ingested data from a variety of sources. The metadata structure also helps to express relations or common-

[28] see https://conversion-hub.sshopencloud.eu

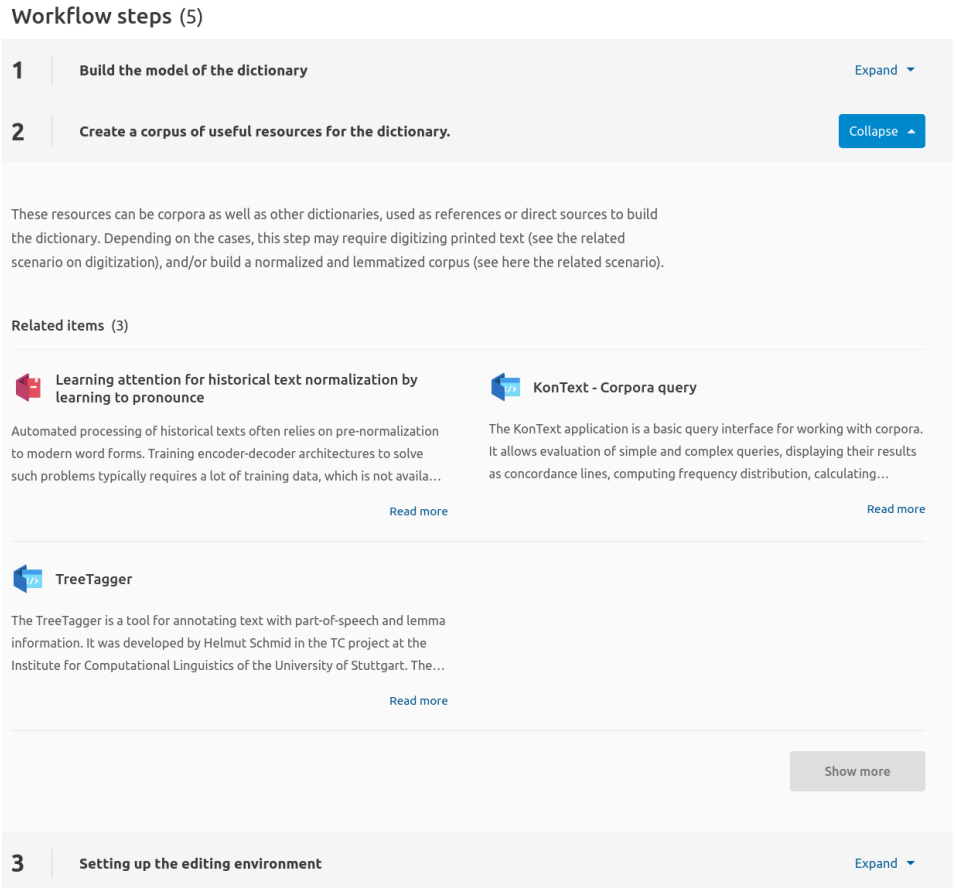


Figure 2: Workflow detail

alities between items, such as the shared data formats, and thereby can help users find appropriate tools through an increase in searchable attributes. Decisions on new metadata fields are generally made with the involvement of the technical, curation and ingest teams of the SSH Open Marketplace (i.e. service providers and Editorial Board members).

3.3 Workflows

Workflows are, as described earlier, sequences of steps utilizing different SSH Open Marketplace items such as tools, training materials and datasets. Workflows have important functions in the SSH Open Marketplace, which will be described here insofar as they are in relation to CLARIN.

Connecting multiple items, workflows play an important role in contextualizing resources throughout the SSH Open Marketplace. Along all marketplace item types, workflows have by far the highest number of connected items, making them essential in contextualizing resources. Through this contextualisation work, previously unconnected items from different data ingests can be integrated into rich workflows.

Workflows allow researchers to present and share their methodology in a unique way. Through workflows, researchers can document every step of a methodology in accurate detail, thus providing a reproducible and transparent way of documenting research. Regarding research data management and FAIR principles, workflows can offer a compatible way to document and share research. Workflows are highly connected within the marketplace, and thus they allow for expert-led serendipity in the sense of discovering related resources that have been curated by the workflow author.

As the above Figure shows, workflows are composed of steps, which relate to different item types. Users can add as much detail as they like to workflows and workflow steps, making the workflow type

ideal for documenting complex methodologies.

Because workflows combine practical application of individual tools with subject-matter expertise concerning the larger research project, they enable the SSH Open Marketplace to go beyond the representation of technical aspects about research tools and to provide the researcher with methodologies. This helps the SSH Open Marketplace to keep the focus on research and treat tools as utilities for said research workflows.

For instance, Marongiu et al., 2024 describe two multilingual workflows for semantic change research built on the SSH Open Marketplace and using CLARIN resources. The workflows proposed by Marongiu et al., 2024 aim to support research in lexical semantic change, i.e. the phenomenon by which words change their meaning over time. The workflows each consist of a series of steps required to detect words that have undergone semantic change as evidenced by a corpus and cover a range of user scenarios, including lexicology, historical research, and legal studies. The workflows are both research domain- and language-independent, and present the advantage of "simplifying access to relevant language resources and tools scattered across different repositories and platforms" (Marongiu et al., 2024, section "Reuse potential"). As put by the authors, the SSH Open Marketplace was chosen to be used as an environment for the creation of these workflows thanks to: (i) its possibility to create links to various resources for each step and even at the workflow level itself, provided that the resource in question is part of the platform, (ii) its broad scope across the whole SSH domains and its robust infrastructure, and (iii) its anchoring in three ERICs: CLARIN, DARIAH and CESSDA. The example of Marongiu et al., 2024 brings into light the great potential of the SSH Open Marketplace and its *workflow* type of item for increasing the findability and reusability of CLARIN resources.

At the time of writing, workflows in the SSH Open Marketplace are linear sequences of steps, and as such do not in all cases represent real-life research activities. A possible future extension of the Marketplace may therefore be the introduction of more flexible workflow types, including elements such as decision points, iterations and multiple end points. Workflows and their possible future extensions are discussed in more detail in a forthcoming publication. (see Barbot et al., 2024) These additions will increase the usability of the workflow item type in documenting existing research practices and will help to improve the usability of provided training resources.

3.4 Multilinguality of user interface and of records

Especially in the context of CLARIN resources, support for multilingual content in the SSH Open Marketplace is crucial. The SSH Open Marketplace currently features only an English user interface, with the majority of records presented in English. Operating as a monolingual discovery platform has its advantages and drawbacks. On the positive side, there's enhanced discoverability of resources when users employ English keywords and the utilization of an English-controlled vocabulary. Conversely, the drawback lies in the necessity to translate names and descriptions of resources from non-English languages, recognizing that these resources may not be valuable to users unfamiliar with the resource's language. Some of the vocabularies used in the SSH Open Marketplace, such as TaDiRAH, the Taxonomy of Digital Research Activities in the Humanities (Borek et al., 2021), are available in multiple languages and the Marketplace could also rely on them to bring more multilinguality to its interface.[29]

To evaluate the impact and potential expansion of the SSH Open Marketplace through multilinguality, it is crucial to comprehend the different levels of multilinguality in this context. Given that the marketplace serves as an aggregator and discovery platform, its content is metadata, not the resource itself, the resource content, documentation, and associated materials would still remain in its original language, most often English. Envisioning a scenario where users find translated metadata useful while the resource itself is solely available in English is challenging. For a more cohesive user experience, providing metadata in languages other than English becomes essential. However, as a resource aggregator, the SSH

[29] In that regard, the work done on the Triple Vocabulary (Triple Project Consortium, n.d.) based on a subset of the Library of Congress Subject Headings and with concepts translated in Greek, French, Polish, German, Italian, Portuguese, Spanish and Croatian is an interesting line of work to build on. See also the freely, openly available data a set of multilingual metadata concepts and an automatically extracted multilingual Data Stewardship terminology developed and delivered by Gamba et al., 2022

Open Marketplace currently cannot facilitate such translation work.

A solution would be to integrate an automatic translation system which would provide multilingual descriptions of the records. This would likely result in imperfect translations, which would need to be revised afterwards. This track was explored during the SSHOC project by Gamba et al., 2022, who tested the application of Natural Language Processing and Machine Translation approaches in view of providing resources and tools to foster multilingual access and discovery to SSH content across different languages. They tested state-of-art machine translation tools, such as Deep-L[30] or Google Translate[31] for the translation of metadata concepts and their definitions into Dutch, French, Greek, and Italian. Their exploratory works revealed both advantages and drawbacks to integrating automatic translation systems into the SSH Open Marketplace. On the one hand, the tested tools proved to be useful for the translation of metadata as the decrease of translation quality is minimal compared to the gain in terms of time and effort needed for traditional human translation. On the other hand, the results of the automatic translation must be revised and validated by humans who are experts in the domain and who have high language proficiency. In conclusion, current machine translation technologies do have limitations and cannot completely replace manual revision.

Furthermore, the Editorial Board encourages the inclusion of records in other languages, offering suggestions to enhance the search experience for non-English records. Specifically, it is recommended to use or at least add English names for entries to aid discoverability. If an English name doesn't exist or translating it doesn't make sense, including an English description in the title is recommended (e.g., Portal xx, Corpus xx). Resource descriptions can be in a language other than English, but including a short English description is advised for broader discoverability. Additionally, the use of English keywords is strongly encouraged for consistent discoverability, and specifying the language of the resource in the dedicated metadata field is highly recommended.

Finally, multilinguality takes on yet another meaning as it can refer to languages a tool or service is capable of processing. This broader perspective underscores the importance of accommodating diverse linguistic needs within the SSH Open Marketplace. While acknowledging the importance of multilinguality, the Editorial Board defers this aspect to future work for the SSH Open Marketplace.

3.5 Towards a better integration

First, investigating how a better connection between the CLARIN Resource Families and the SSH Open Marketplace could take shape is one of the main lines of work for the Editorial Board and the CLARIN team involved. At the moment, indeed, CRF records are harvested by the SSH Open Marketplace, using manually curated files on GitHub as a source. This has been proven challenging for the continuous ingest pipeline. And that is why a more structured workflow is now under investigation, opening up opportunities to also join the (manual) curation efforts of both teams leading to better metadata description of the CRF records.

Second, we will look into improving the connection of the Marketplace with the VLO, which is a vast discovery portal including almost a million metadata records harvested from 47 CLARIN Centres and various non-CLARIN sources like Europeana or ELRA. Indeed, the SSH Open Marketplace by its nature extends beyond the CLARIN world and it could be interesting to investigate in what way the SSH Open Marketplace could complement the VLO, the CLARIN Resource Families and the tools in the Language Resource Switchboard both from a technical point of view (i.e. mutual harvesting) as well as from the points of view of increasing the findability and accessibility of language data.

Third, the SSH Open Marketplace extends CLARIN's quite complex infrastructure of discovery portals, which already includes the VLO, the Language Resource Switchboard and the CLARIN Resource Families. This multitude of discovery portals can be confusing for researchers or developers that would like to include information about their resource into the CLARIN infrastructure and want to ensure the maximum visibility for the community. The same could be the case for those researchers who aim to discover resources. These users currently have at their disposal more and more discovery platforms useful

[30] see https://www.deepl.com/translator

[31] see https://translate.google.com

to access resources as data, tools and services. To diminish the risks of confusion, it is therefore planned to create a guide that clearly outlines the various options of discovery portals, with their similarities and complementarities, to better inform and guide users who want to share their resources, as well as users who search resources. For instance, while the SSH Open Marketplace allows to discover the same language data and tools as the CLARIN discovery portals, it presents a number of differentiation points, as follows:

- The SSH Open Marketplace showcases all its resources, including CLARIN resources, in a contextualised manner, namely via the relations that can be created at the metadata level with other SSH Open Marketplace resources which can be: ”relates to”, ”is related to”, ”documents”, ”is documented by”, ”mentions”, ”is mentioned by”, ”extends”, ”is extended by”).

- Through contextualization, the SSH Open Marketplace enables putting in relation different types of resources, for instance, services and tools can be put in relation with training materials, datasets, publications and/or workflows. This increases the level of informativeness and usability of resources by allowing users to easily finding and accessing related information about the initial resource of interest.

- The SSH Open Marketplace allows searching, in addition to the classical search by keywords, by type of activity, search that draws on the TaDiRAH, the Taxonomy of Digital Research Activities in the Humanities. This increases the probability of finding CLARIN resources, especially for users who are not acquainted with language data or with the CLARIN VLO, the CLARIN Resource Families and the Language Resource Switchboard.

- Having CLARIN resources in the SSH Open Marketplace broadens the type of potential users that may discover and use them through the fact that scholars from social sciences and humanities disciplines, who do not regularly use language data, use the SSH Open Marketplace to discover new resources relevant for their research.

- In opposition to all other discovery engines, the SSH Open Marketplace proposes a unique type of resource, which becomes more and more important when it comes to reproducible research: workflows. This resource type represents an interesting niche that some CLARIN users have already adopted (as explained in the workflows section above) and that will be further promoted among the CLARIN community through a CLARIN Café and hands-on workshops.

Finally, new EOSC-related projects in which CLARIN is involved, such as Open Science Plan-Track-Assess Pathways (OSTrails)[32], will also investigate if and how the SSH Open Marketplace could play a role, alongside other research resource catalogues, in a wider (SSH) knowledge graph. These developments need to be closely monitored, keeping in mind that the curation work currently happening in the SSH Open Marketplace and the records enrichment should continue to serve the research community, no matter which form the resource catalogues will take in the future.

4 Conclusion

As one of the founders of the SSH Open Marketplace, CLARIN ERIC supports the SSH Open Marketplace and its development for the benefits of the CLARIN community and to increase the findability and re-usability of CLARIN resources beyond its target user community. Such a joint enterprise proposing shared services between ERICs has the strong potential to better serve the SSH communities. This is true especially in the EOSC context, where the Cluster approach seems to be an appropriate level to join forces for various scientific domains. Having shared and flagship services, such as the SSH Open Marketplace, for the Cluster demonstrates collaboration capacities between ERICs, while ensuring long-term sustainability, for the benefits of the research community. The process of integrating CLARIN resources into the SSH Open Marketplace exemplifies some of the overall challenges discovery portals such as the

[32] see https://cordis.europa.eu/project/id/101130187

marketplace face. These challenges are generally expressed in the difficult balance between an overly universal and overly specific approach to the data structure. As every new item potentially increases the curation work and requires new, subject-specific knowledge on the side of the curation team, it is important that both the technical and curatorial side are involved in future development of the marketplace. The SSH Open Marketplace can benefit from the CLARIN community expertise in curation work to guarantee metadata quality, especially in aggregation context, based on the knowledge of the VLO and CRF experiences.

References

Barbot, L., Dolinar, M., Gray, E. J., Grisot, C., Illmayer, K., Kurzmeier, M., & McGillivray, B. (2024). Contextualizing research tools & services through workflows in the SSH open marketplace. *Journal of Open Humanities Data*, *10*(1). https://doi.org/10.5334/johd.192

Barbot, L., Moranville, Y., Fischer, F., Petitfils, C., Ďurčo, M., Illmayer, K., Parkoła, T., Wieder, P., & Karampatakis, S. (2021, February). SSHOC D7.1 System Specification - SSH Open Marketplace. https://doi.org/10.5281/zenodo.4558302

Borek, L., Hastik, C., Khramova, V., Illmayer, K., & Geiger, J. D. (2021). Information Organization and Access in Digital Humanities: TaDiRAH Revised, Formalized and FAIR. Universität Regensburg. https://doi.org/10.5283/EPUB.44951
Other Session 5: Knowledge Representation.

Dombrowski, Q. (2014). What Ever Happened to Project Bamboo? *Literary and Linguistic Computing*, *29*(3), 326–339. https://doi.org/10.1093/llc/fqu026

Dombrowski, Q. (2021). The directory paradox. In *People, practice, power: Digital humanities outside the center* (pp. 83–98). University of Minnesota Press. Retrieved April 28, 2023, from http://www.jstor.org/stable/10.5749/j.ctv2782dmw.9

Ďurčo, M., Barbot, L., Illmayer, K., Karampatakis, S., Fischer, F., Moranville, Y., Ocansey, J. T., Probst, S., Kozak, M., Buddenbohm, S., & Yim, S.-B. (2021, December). 7.2 marketplace – implementation. https://doi.org/10.5281/zenodo.5749465

Fišer, D., Lenardič, J., & Erjavec, T. (2018). CLARIN's Key Resource Families. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*. https://aclanthology.org/L18-1210

Gamba, F., Frontini, F., Broeder, D., & Monachini, M. (2022). Language technologies for the creation of multilingual terminologies. lessons learned from the sshoc project. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 154–163. https://aclanthology.org/2022.lrec-1.17

Marongiu, P., McGillivray, B., & Khan, A. F. (2024). Multilingual Workflows for Semantic Change Research. *Journal of Open Humanities Data*, *10*, 15. https://doi.org/10.5334/johd.179

Petitfils, C., Dumouchel, S., Larrousse, N., Gray, E. J., Barbot, L., Roi, A., Ďurčo, M., Illmayer, K., Buddenbohm, S., & Parkola, T. (2021, October). D7.5 marketplace - governance. https://doi.org/10.5281/zenodo.5608487

Riondet, C., & Romary, L. (2018). The standardization survival kit: For a wider use of metadata standards within arts and humanities. *Archives et Bibliothèques de Belgique-Archief-en Bibliotheekwezen in België*, *106*, 55–62.

Triple Project Consortium. (n.d.). Triple Vocabulary: An SSH multilingual vocabulary based in LCSH [Institution: National Documentation Centre]. https://doi.org/10.12681/semantics.gr/SSH-LCSH

Zinn, C. (2018). The Language Resource Switchboard. *Computational Linguistics*, *44*, 1–13. https://doi.org/10.1162/coli_a_00329

MWE-Finder: An evaluation through three case studies

Martin Kroon
Institute for Language Sciences
University of Utrecht, the Netherlands
m.s.kroon@uu.nl

Jan Odijk
Institute for Language Sciences
University of Utrecht, the Netherlands
j.odijk@uu.nl

Abstract

In this paper we showcase and evaluate MWE-Finder, a system that allows users to search for occurrences of an MWE in a large Dutch text corpus. To this end, we conduct three small case studies, and discuss the results in detail. We make use of the MWEs *0geen *+haan zal naar iets kraaien* 'no one will say anything about something', *iemand zal 0dat *+varken wassen* 'someone will deal with that problem' and *iemand zal iemand het hemd van het lijf vragen* 'someone will want to know all the ins and outs of something from someone', which are all in canonical form following Odijk (2023) and Odijk and Kroon (2024).

The results show that MWE-Finder is very accurate in retrieving the target MWEs, reaching an accuracy of 93.7%, and an F_1-score of 95.2%. The case studies additionally lay bare points of improvement of MWE-Finder, specifically concerning the enrichment of syntactic parses by making the object relation explicit in certain constructions.

1 Introduction

Many multiword expressions (MWEs) are flexible in the sense that their components can have different forms, can occur in different orders, or may not be contiguous, with other words appearing between elements of the MWE. This makes searching for such MWEs in large text corpora difficult. To this end, Odijk et al. (2024) developed MWE-Finder for Dutch.

MWE-Finder is a system that allows users to search for occurrences of an MWE in a large Dutch text corpus. It automatically generates three queries based on an input in canonical form. These three queries are increasingly less strict, and allow the user to investigate potential variation easily. MWE-Finder is implemented in the newest version (v5) of GrETEL (Augustinus et al., 2012), and ships with several pre-configured corpora, e.g. SONAR (Oostdijk et al., 2013), parts of Lassy-Large, Lassy-Small (van Noord et al., 2013), the Spoken Dutch Corpus (Oostdijk et al., 2002), and Mediargus.[1] It also offers the user over 11k Dutch MWEs in a canonical form from the DUCAME resource (Odijk, 2023).[2] MWE-Finder is particularly effective, as it takes into account the syntactic structure of the MWEs and the sentences over which it queries using the Alpino parser (van Noord, 2006).

MWE-Finder is intended as a research tool for any linguist or lexicographer interested in research into multiword expressions, in particular *flexible* multiword expressions. It is therefore a natural part of the CLARIN research infrastructure, and it is factually part of it because it is embedded in the CLARIN web application GrETEL. The three queries can lay bare potential variation, leading to the updating of the canonical form in a lexicon for MWEs such as DUCAME, and a more complete and accurate description of the MWE. The system does *identification* (in the sense of (Constant et al., 2017)) of candidate occurrences of MWEs. Though it cannot determine whether an expression is used literally or as an MWE, most occurrences are found to be instances of the MWE, compliant with the observations by other researchers (Savary et al., 2019).

[1] A large treebank with Flemish newspaper text created by Kris Heylen from KU Leuven in 2009.

[2] https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP

Martin Kroon and Jan Odijk 2023. MWE-Finder: An evaluation through three case studies. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 126–142. DOI: https://doi.org/10.3384/ecp210

Whereas Odijk et al. (2024) describe how MWE-Finder works in detail and explain certain development choices, this paper acts as a showcase of how MWE-Finder can be used to determine properties of Dutch MWEs on the basis of large corpus data and as an evaluation of MWE-Finder's performance when querying for MWEs, identifying potential points of improvement for the system. To this end, we conduct three small case studies, and discuss the results in detail. We make use of the following Dutch MWEs, which are in canonical form:

(1) 0geen *+haan zal naar iets kraaien
no rooster will to something crow
'no one will say anything about something'

(2) iemand zal 0dat *+varken wassen
someone will that pig wash
'someone will deal with that problem'

(3) iemand zal iemand het hemd van het lijf vragen
someone will someone the shirt from the body ask
'someone will want to know all the ins and outs of something from someone'

The organisation of this paper is as follows. We begin with a brief introduction of the notion multiword expression (Section 2), followed by a dicussion of the notion *canonical form* for an MWE (Section 3). We continue with a brief description of MWE-Finder in Section 4. Section 5 introduces general characteristics of the evaluation performed here. In Subsections 5.1 through 5.3 we discuss the results from the small case studies involving the MWEs listed above. We briefly discuss related work in Section 6 and we conclude in Section 7.

2 Multiword expressions

An MWE is a word combination with linguistic properties that cannot be predicted from the properties of the individual words or the way they have been combined by the rules of grammar (Odijk, 2013).[3] A word combination can, for example, have an unpredictable meaning (*de boeken neerleggen*, lit. 'to put down the books', meaning 'to declare oneself bankrupt'), an unpredictable form (e.g. *ter plaatse* 'on location', with idiosyncratic use of *ter* and *e*-suffix on the noun),[4] or it can have only limited usage (e.g. *met vriendelijke groet* 'kind regards', used as the closing of a letter). In a translation context, it can have an unpredictable translation (*dikke darm*, lit. 'thick intestine', 'large intestine'), etc.

Note that it is not always easy to determine whether a combination of words is an MWE, because we do not always know the exact properties of the individual component words or what the grammar rules of a language are exactly. So this may require a substantial amount of research.

Words of an MWE need not always be fixed. This can be illustrated with the Dutch MWE *de boeken neerleggen* 'to declare oneself bankrupt'. The verb *neerleggen* in (4) can occur in all of its inflectional variants (e.g., past participle in (4a), infinitive in (4b), and past tense singular in (4c) and (4d)), and with the separable particle *neer* attached to it (4a, 4b) or separated (4c, 4d). MWEs do not necessarily consist of words that are adjacent, and the words making up an MWE need not always occur in the same order. This expression allows a canonical order with contiguous elements (as in (4a)), but it also allows other words to intervene between its components (as in (4b)), as well as permutations of its component words (as in (4c)), and combinations of permutations and intervention by other words that are not components of the MWE (as in (4d)):

(4) a. Saab heeft gisteren **de boeken neergelegd**.
Saab has yesterday the books down.laid
'Saab declared itself bankrupt yesterday.'

b. Ik dacht dat Saab gisteren **de boeken** wilde **neerleggen**.
I thought that Saab yesterday the books wanted down.lay
'I thought Saab wanted to declare itself bankrupt yesterday.'

[3] For a similar but slightly different definition, see (Sag et al., 2001).

[4] *Ter plaatse* actually concerns a fossilization of an old dative form, which is no longer productive in Dutch.

c. Saab **legde de boeken neer**.
Saab laid the books down
'Saab declared itself bankrupt.'

d. Saab **legde** gisteren **de boeken neer**.
Saab laid yesterday the books down
'Saab declared itself bankrupt yesterday.'

In addition, certain MWEs allow for (and require) controlled variation in lexical item choice, e.g. in expressions containing bound anaphora such as *zijn geduld verliezen* 'to lose one's temper', where the possessive pronoun varies depending on the subject (cf. *Ik verloor mijn/*jouw geduld; jij verloor *mijn/jouw geduld*, etc.), exactly as the English expression *to lose one's temper*. Of course, not every MWE allows all of these options, and not all permutations of the components of an MWE are well-formed (e.g. one cannot have **Saab heeft neergelegd boeken de.* lit. 'Saab has downlaid books the.').

This flexible nature of such MWEs makes it difficult to reliably search for such expressions in text corpora. Standard search engines such as Google do not enable the user to systematically search for different word forms of the same lemma. Search applications such as OpenSoNaR (de Does et al., 2017; van de Camp et al., 2017) or Nederlab (Brugman et al., 2016) can do this, but it is difficult to formulate a query allowing different orders and interspersed irrelevant words, and the results of such a query will be unreliable. At best, one will find all instances but at the same time also many cases where all the component words occur but do not make up an MWE. One should be able to search for flexible MWEs in such a way that their grammatical structure is taken into account. This can be done in a treebank, and MWE-Finder enables searching for MWEs in a treebank.

3 Canonical Form

A canonical form for an MWE is a unique representation for a set of variants of this MWE that differ only in grammatical properties. A canonical form for MWEs is necessary because many MWEs are flexible, i.e., their component words can occur in different forms, in different orders, or need not always be adjacent. Odijk and Kroon (2024) describe related work on canonical forms for MWEs. The canonical forms assumed here have several unique features: (1) they are defined by very detailed requirements (Odijk, 2023; Odijk & Kroon, 2024); (2) they form well-formed utterances of the language; (3) they can be enriched with annotations; (4) the definition of canonical form has been tested on more than 11k Dutch MWEs in the DUCAME resource; and (5) there are explicit rules on how one can generalise from the canonical form to other forms, which have been implemented in MWE-Finder by a mechanism for the automatic generation of multiple queries for searching for the MWE in large text corpora.

4 MWE-Finder

MWE-Finder enables a user to search for occurrences of an MWE in a treebank based on an example MWE. The example MWE must be in canonical form. The canonical form for MWEs has been defined in (Odijk, 2023) and (Odijk & Kroon, 2024). Canonical forms of MWEs can contain annotations to describe properties of their component words. We will not say anything about this here, but we do provide a list of the most important annotations in Table 1.

MWE-Finder is available in a first version since the end of 2022 as part of the web application GrETEL 5.[5] Thanks to this integration, MWE-Finder has access to all GrETEL features, and supports all treebanks that are included in GrETEL as well as its possibility of uploading one's own text corpora. MWE-Finder partially mimics the structure of GrETEL's main *query-by-example* search functionality. It distinguishes the following steps: *Canonical Form* (cf. GrETEL's *Example* step), *Treebanks*, *Results*, and *Analysis*.

Just like GrETEL, MWE-Finder enables the user to enter an MWE example, though it must be in its canonical form. The user thereby implicitly formulates a hypothesis about the properties of this MWE. The annotations on the example (or their absence) specify how the system should generalise from this example, so these annotations can be seen as a different way of implementing GrETEL's *Matrix*.

[5] https://gretel5.hum.uu.nl

notation	interpretation
**word*	*word* is modifiable/determinable
+word	*word* is inflectable
=word	*word* must occur in the MWE as given
!word	*word* is not modifiable/determinable
dd:[*word*]	*word* must be a definite determiner
<text>	*text* is interpreted as a freely replaceable argument
0*word*	*word* is not part of the MWE

Table 1: Notational devices for annotating a canonical form. The code + can also be combined with * or *!* (in any order).

The MWEs contained within DUCAME[6] have been included in a drop-down list and are directly searchable within MWE-Finder. The user can also enter a new MWE, provided that it complies with the conventions for MWE canonical forms.

After the MWE has been selected or entered, the system automatically generates three queries to search for occurrences of this MWE in a treebank. The three queries correspond to different levels of agreement between the MWE and the sentences of the corpora. They are the *major lemma query*, the *near-miss query*, and the *MWE query*. The query generation process has been explained in detail in (Odijk et al., 2024).

The MWE query (MEQ) searches for sentences that contain an occurrence of the MWE. The near-miss query (NMQ) searches for sentences in which the major words[7] of the MWE occur in the grammatical configuration in which they occur in the MWE. It allows the presence of determiners and modifiers that are not expected on the basis of the MWE's canonical form. The results of the NMQ are a superset of the results of the MEQ. MWE-Finder enables the user to inspect the difference between the results of the NMQ and the results of the MEQ. The Major Lemma Query (MLQ) searches for sentences in which the major words of the MWE occur in any grammatical configuration. The results of the MLQ are a superset of the results of the NMQ.

Next, the user can select the treebank that the query should be applied to. Once selected, the application switches to the *Results* view where query results are displayed as they arrive from the server. In that view, the user can also switch between the different queries for the chosen MWE or choose to exclude results of finer-grained queries. It is also possible to inspect or manually change the automatically generated XPath queries and retrieve new results. In the *Results* view, users can also look at the parse trees for results or toggle extra context (one preceding sentence, one following sentence) to better analyze the occurrences found, just like in GrETEL.

Finally, there is the analysis step, which we will not describe here and for which we kindly refer to (Odijk et al., 2024).

Odijk et al. (2024) illustrated MWE-Finder using the example MWE *de dans ontspringen* (lit. 'the dance escape', 'to escape the nasty consequences'). Here we will carry out small case studies for three different MWEs to show the potential of MWE-Finder and to identify opportunities for improving it.

5 Evaluating MWE-Finder

In this paper we evaluate MWE-Finder by assessing the performance of three different MWEs. In such an evaluation multiple factors play a role:

- The quality of the Alpino parses;
- The quality of the canonical form;

[6] DUtch CAnonical form Multiword Expressions, https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP.

[7] The major words of an MWE are the content words if there is more than one, and the content and function words otherwise.

- The quality of the automatic query generation.

We focus on the quality of the automatic query generation in this paper, but the other aspects will be relevant in some cases as well.

For each generated query we have specific expectations. We expect that the MEQ finds examples that satisfy all lexical, morphological and syntactic requirements that are encoded in the canonical form. Savary et al. (2019) found that when syntactic conditions necessary for an idiomatic reading are fulfilled, this reading occurs in 96% to 98% of the cases. Therefore we expect most of the results found by MEQ to have the idiosyncratic reading of the MWE.

We expect that the results of NMQ minus the results of the MEQ do not contain instances of the MWE. If it does contain them, it could mean that the canonical form for the MWE was too strict (see an illustration of this with the MWE *de dans ontspringen* 'to escape the nasty consequences' in (Odijk et al., 2024)). In such a case, the user can adapt the canonical form and evaluate with this revised canonical form. It can also mean that Alpino parsed the sentence incorrectly, or preferred one parse over another in case of an ambiguity, or that MWE-Finder's query generation mechanism contains errors or omissions.

As for the MLQ, it is the expectation that it finds all sentences in which the lemmas of the major words of the MWE occur. This is surely the case, but we have no guarantee that Alpino assigns the right lemmas to the words of an MWE in a particular sentence. This may happen if a word is ambiguous (e.g., *bommen*) and Alpino analyses it in a particular sentence as a noun ('bombs') rather than as a verb ('to concern') as required by the MWE. Or when a combination of the verb *passen* with the word *aan* is not analysed as a verb + preposition combination ('to fit to'), as required by the MWE, but as a verb + separable particle combination ('to adapt'). We are aware that such examples exist, but we will not deal with this in this paper. We are working on an even more general query (the *Related Word Query*, RWQ) to cover such cases, and we hope to describe this in more detail in future work.

We also expect that the results of the MLQ minus the results of the NMQ do not contain instances of the MWE. The latter can be the case, however, especially when Alpino parsed the sentence incorrectly, or when MWE-Finder's query generation mechanism contains errors or omissions.

We will deal with each of these aspects in the sections on specific MWEs. We will give links to the queries for each of the examples in the tables summarising the search results. For one MWE we have also listed the queries in the Appendix.

5.1 0Geen *+haan zal naar iets kraaien

In this section we illustrate MWE-Finder with the MWE canonical form in (5), which we will call the *target MWE*.

(5) 0geen *+haan zal naar iets kraaien
no rooster will to something crow
'no one will say anything about something'

We search in the treebank for the SoNaR corpus (Oostdijk et al., 2013), known in MWE-Finder as *Sonar4*. It contains more than 40 million sentences.

The canonical form states that the word *haan* can be modified and inflected, and that the word *geen* is not a component of this MWE. The queries automatically derived from this canonical form yield the results given in (6):

(6) Results of the queries derived from *0geen *+haan zal naar iets kraaien*[8]

Query	**Matches**
MEQ	331
NMQ	379
MLQ	804

We discuss the results of the three queries in separate subsections, the MLQ results in section 5.1.1, the NMQ results in section 5.1.2, and the MEQ results in section 5.1.3.

[8]The query names in the table are links to the query results in the application. Beware that there may have been updates to MWE-Finder since the time of writing, influencing results.

5.1.1 Analysis of the MLQ results

In order to analyse the MLQ results we only consider the MLQ results without the NMQ results. The NMQ results will be analysed in section 5.1.2.

The analysis of the MLQ results minus the NMQ results (804 – 379 = 425) has been summarised in Table 2. We discuss it here in detail.

MWE?	Cause	Details	Total
no	other MWE	*iemand zal victorie kraaien*	1
		iemands haan zal koning kraaien	5
			363
	variant-P + R-pronouns	*om*	17
		over	6
		achter	1
		Total	**393**
yes	*daar*-drop	(blank)	1
	dialect		1
	no P		3
	wrong sentence		1
	wrong parse	*naar* as ADJ, no `pc`	5
		PP attachment	1
	typo		1
	R-pronoun		17
		Total	**30**
unclear	no P		2
		Total	**2**
		Grand Total	**425**

Table 2: Analysis of the MLQ results minus the NMQ results (804 – 379 = 425) for *0geen *+haan zal naar iets kraaien*.

We first focus on example sentences that do not involve the target MWE. There are 393 such cases. 363 sentences contain forms of *haan* and *kraaien* but must be interpreted literally.

There are variants with other prepositions than *naar*: *om* (17), *over* (6), *achter* (1). For the authors all of these are ill-formed but *om* and *over* are probably correct variants of local dialects, given the number of occurrences. One would expect these as a result of the NMQ, but they are not there because the NMQ allows the absence of *naar* but still requires the presence of a prepositional complement, or because they contain R-pronouns,[9] which were not treated correctly yet at the moment of the measurement. We believe that it is necessary to reformulate the procedure to generate the NMQ so that it does not require an adpositional complement if its head need not be present.

In some cases the sentences contain a different MWE in which *haan* and *kraaien* happen to occur, e.g. *iemand zal victorie kraaien* (1 occurrence), and *iemands haan zal koning kraaien* (5 occurrences), an expression that the authors do not know (but it occurs in (Kruyskamp, 1974)):[10]

(7) different MWEs:

a. De Franse haan **kraait** dus **victorie**.
the French rooster crows so victory
'Thus the French celebrate the victory.' (WR-P-E-H_part00004.data.dz:4330)

[9] A limited set of pronouns with special syntax that all happen to contain the character *r*, e.g. *er* 'there', *waar* 'where', *hier* 'here'.

[10] The actual sentences are often very long. For long sentences we selected the relevant part and used '...' to indicate that we left out a part of the full sentence.

b. ... en voor de rest van de dag **kraait jouw haan koning**.
... and for the rest of the day crows your rooster king
'... and for the rest of the day you will have your way.' (WR-P-P-B_part00129.data.dz:7676)

There are 5 occurrences without a preposition, 3 of which are intended as the target MWE. For 2 it is unclear whether they are intended as the target MWE or as the literal expression (*unclear* in Table 2). The possibility for the adposition to be absent suggests that a variant without a prepositional complement must be included in DUCAME.

There are several occurrences in which *naar* is analysed as an adjective (meaning 'nasty') instead of a preposition, and no adpositional complement (pc) is present (5 occurrences).

One case involves topic drop of the complement of the preposition *naar*, most probably *daar*, a construction that is not in accordance with the norms of the standard language but that occurs frequently in colloquial language. MWE-Finder cannot currently handle this construction properly, so the example is in the MLQ results rather than in the MEQ results.

(8) ... **kraait** geen **haan naar**
... crows no rooster to
'... no one will say anything about that.' (WR-U-E-A_part00227.data.dz:9310)

17 cases indeed involve the target MWE but contain R-pronouns or pronominal adverbs not dealt with correctly yet by MWE-Finder at the time, which is why we find them among the results of the MLQ.

Finally, some have not been identified as the target MWE because the Alpino parse is completely wrong, in one case because of the strong dialectical nature of the sentence, in other cases caused by typos or grammar errors.

5.1.2 Analysis of the NMQ results

We analyze the NMQ results minus the MEQ results. This involves 48 examples. Most do not involve the target MWE. There are 2 occurrences of a different MWE (*iemands haan zal victorie kraaien*), 3 examples with the preposition *over* instead of *naar*, and 42 examples involve a literal interpretation of the words *haan* and *kraaien*. Only one case does involve the target MWE, but here *naar* has been parsed as an adjective, which is why we find this example among the NMQ results rather than among the MEQ results.

5.1.3 Analysis of the MEQ results

In example (9) Alpino analysed a directional PP incorrectly as a prepositional complement to the verb *kraaien*. It is clearly not an instance of the target MWE, because the sentence does not contain a negative polarity licensor (see below).

(9) ... en er kraaide een haan .- (14:68) naar het voorportaal ...
... and there crowed a rooster .- (14:68) to the vestibule ...
'... and a rooster crowed .- (14:68) to the vestibule ...' (WR-P-P-B_part00165.data.dz:4493)

All other results found among the MEQ results are indeed instances of the target MWE. A special case to mention is (10), in which the word *haan* is the head of a noun phrase that contains a relative clause that contains the verb *kraaien* with the relative pronoun *die*. MWE-Finder can correctly handle such constructions.

(10) Geen **haan** die **ernaar kraaide**.
no rooster that there.to crowed
'(There was) no one who said anything about it.' (WR-P-E-A_part00099.data.dz:8178)

As for form variants of *haan*, we mostly find *haan*, but occasionally also the plural form *hanen*. We encountered no diminutive forms.

As far as determination is concerned, *geen* 'no' was the most frequent determiner, but the indefinite article *een* 'a' occurred approximately 20 times (including the dialectical variant *ne*) and *weinig* 'few' had 2 occurrences.

The MWE is a negative polarity construction, i.e. it requires a negative polarity licensor (NPL) in the sentence (and in the right position). The determiners *geen* 'no' and *weinig* 'few' are such NPLs. In the examples where the determiner *een* occurs we observed NPLs such as *zonder* (as in (11a)) 'without', *nauwelijks* 'hardly', *amper* 'hardly', *nooit* 'never' and yes-no questions (as in (11b)):

(11) Some examples with negative polarity licensors:

a. ... zonder dat er in het Westen een **haan naar kraaide** ...
... without that there in the West a rooster to crowed ...
'... without anyone saying anything about it in the West ...'
(WR-P-E-A_part00199.data.dz:2034)

b. ... is daar een **haan** die daar **naar kraait**?
... is there a rooster that there to crows
'... is there anyone who says anything about that?' (WR-P-E-A_part00021.data.dz:1667)

c. ... zonder dat er ook maar een diplomatieke **haan naar kraait**.
... without that there also but a diplomatic rooster to crows
'... without any diplomat saying anything about it.' (WR-P-P-H_part00272.data.dz:2312)

MWE-Finder currently does not check for the presence of NPLs, but this surely could be a useful extension of the MEQ. In fact, it is crucial if we would allow *0geen haan zal kraaien* (without the PP) as a canonical form: if there is no check for a polarity licensor any sentence with *haan* in the subject of *kraaien* will match.

Modifiers of *haan* were most often absent, but we encountered the adjectives *diplomatieke* 'diplomatic' (as in 11c), *Europese* 'European', and *rode* 'red',[11] and relative clauses.

5.2 Iemand zal 0dat *+varken wassen

For another illustration of MWE-Finder we use the MWE as in (12), which we shall again refer to as the *target MWE*.

(12) iemand zal 0dat *+varken wassen
someone will that pig wash
'someone will deal with that problem'

The canonical form of the target MWE indicates that *dat* 'that' is not part of the MWE, and that *varken* can be modified, determined and inflected.

For this example, we query the Mediargus corpus (containing around 103 million sentences), which is pre-configured within MWE-Finder. The queries automatically derived from the target MWE yield the results given in (13):

(13) Results of the queries derived from *iemand zal 0dat *+varken wassen*:

Query	**Matches**
MEQ	537
NMQ	537
MLQ	615

As can be seen in (13), the MEQ and the NMQ yield the same hits.[12] All 537 but 2 were instances of the target MWE, with the 2 sentences without the MWE requiring a literal reading of *varken* and *wassen*. The fact that the MEQ and the NMQ yield the same results, suggests that the canonical form is correct as is. Indeed, *dat* is not a part of the MWE: in many instances of the MWE *dat* is replaced by another determiner such as the definite article *het*, or no determiner is present at all. Furthermore, *varken* is often attested in inflected form (it often occurs in its diminutive form *varkentje*, but not exclusively) and modified (it is possible to have an adjective modify the noun). All this is illustrated in one single hit, given in (14).

[11] Most probably intended as 'socialist' or 'communist'.

[12] Remember that the MEQ results are a subset of the NMQ results; when the sets are of equal size, they must contain the same items.

(14) Ik denk dat jonge progressieven wel andere **varkentjes** te **wassen** hebben.
I think that young progressive POS other pigs.DIM to wash have
'I think young progressives have other things to deal with.' (DM_20041213_01.data.dz:1685)

The MLQ, however, found more than the MEQ and NMQ. 78 more sentences were retrieved containing the "major lemmas", i.e. the content words, of the target MWE, regardless of their grammatical relation. Of these 78 hits, 38 were clear instances of the target MWE, while 35 clearly were not, for instance where *varken* was not even a direct object of *wassen*. The remaining 5 hits were more difficult to judge and required closer inspection of the surrounding sentences in the corpus, a feature that MWE-Finder offers. These 5 hits were found to indeed be instances of the target MWE.

Of the 43 (= 38 + 5) misses, 20 can be attributed to a wrong parse, for instance caused by missing punctuation (e.g. (15), which would have been found if it had an extra comma between *gewassen* and *dacht*, as that would have resulted in a correct parse by Alpino), by Alpino's treatment of quotes, the words of which are all put in a flat list of nodes under a `mwu` (multiword unit) phrase (e.g. (16)), or complex, long-distance dependencies (e.g. (17), in which Alpino takes *uit dit administratieve varkentje* as PP, while *uit* should be analyzed as a particle belonging to the main verb *dagen* 'to challenge'). These parsing mistakes are highlighted because they occur more than once.[13]

(15) $ **Varkentje gewassen** dacht iedereen, behalve de bezoekers.
pig.DIM washed thought everybody except the visitors
'Problem dealt with, everybody thought, except for the visitors.'
(NB_19990308_01.data.dz:11479)

(16) $ ["We hadden twee uur nodig om dit **varkentje** te **wassen**",]$_{\texttt{mwu}}$ is de Wevelgemse trainer Marnix
we had two hour needed for this pig.DIM to wash is the Wevelgem trainer Marnix
Pattyn vol lof over de Markse tegenstander.
Pattyn full praise about the Marke opponent
'"We needed two hours to deal with them," Wevelgem trainer Marnix Pattyn praised his opponents from Marke.' (NB_19981202_01.data.dz:12777)

(17) $ Landbouwers dagen staatssecretaris van Administratieve Vereenvoudiging, Vincent Van
farmers challenge Secretary of State of administrative simplification Vincent Van
Quickenborne (VLD), [uit dit administratieve **varkentje**]$_{\text{PP}}$ te **wassen**.
Quickenborne PRT this administrative pig.DIM to wash
'Farmers challenge Secretary of State for Administrative Simplification, Vincent Van Quickenborne (VLD), to deal with this administrative problem.' (NB_20040602_01.data.dz:160)

The remaining 23 results did contain instances of the target MWE, but were missed. It concerns 7 cases of finite relative clauses, with *varken* as antecedent and the verb *wassen* inside the relative clause. While MWE-Finder can deal with relative clauses correctly (cases of relativization were found with the MEQ and NMQ), these cases were missed due to the relative pronoun *dat* being wrongly tagged as a subordinating conjunction (`vg` instead of `vnw`), such as in (18). It is interesting, however, to note that the Mediargus corpus was parsed with an older version of Alpino, and that the wrong parse cannot be recreated using the newer version embedded within MWE-Finder.[14]

(18) Woluwe is een **varkentje** dat$_{\texttt{vg}}$ jullie moeten kunnen **wassen**?
Woluwe is a pig.DIM that 2.PL must can wash
'Woluwe is something that you should be able to deal with?' (NB_20050930_01.data.dz:3975)

Another 13 can all be attributed to the object relation between *varken* and *wassen* not being explicitly labelled by Alpino, or indeed recognized at all, in certain specific non-frequent constructions. During the

[13] With the notation '$', we mean that while the sentence is grammatical (ignoring normative spelling and punctuation rules), the parse associated with the sentence is wrong.

[14] We do not count these tagging errors towards wrong parses, because in certain cases it is correct to analyse the relative pronoun as a `vg` (though not in this one), and MWE-Finder should be robust against this, which these examples showed us it is not. For instance in *de dag dat jij naar Leuven ging* 'the day you went to Louvain', *dat* cannot be a `vnw`, as it does not agree with its antecedent *dag*.

development of MWE-Finder, these constructions were not accounted for, and should be implemented in a future version. These can be broken down into:

- 4 cases of nonfinite relative clauses, with an infinitive and a zero object. In these cases the noun *varken* is modified by an infinitival clause with *wassen* accompanied by the particles *om* and *te*, in which *varken* is understood as the object of *wassen*, e.g. (19).

  (19) Een **varkentje** om te **wassen**
  a pig.DIM for to wash
  'A problem to deal with', lit. 'A piglet to wash' (DM_20060429_01.data.dz:4833)

- 2 cases of a reflexive middle construction involving the permissive verb *laten* 'to let' (Broekhuis et al., 2016), such as in (20). Note that the construction is itself embedded in a relative clause, but that is not what causes the sentence to be missed by the MEQ and NMQ.

  (20) Het probleem van Carestel is geen **varkentje** dat zich snel **laat wassen**.
  the problem of Carestel is no pig.DIM that SELF quickly lets wash
  'The problem of Carestel is not one quickly dealt with.' (DS_20031129_01.data.dz:5618)

- 2 cases of a deverbal noun causing the object *varken* to end up in a PP headed by *van* 'of', and therefore losing its object relation to the verb *wassen*, e.g. (21), which also contains the MWE *de kous zal af zijn* 'it will be settled'.

  (21) Of dachten de heren en dames van de partijbesturen dat met het **wassen** van het
  or thought the gentlemen and ladies of the party.leaderships that with the washing of the
  Antwerpse **varkentje** de kous af is?
  Antwerp pig.DIM the stocking off is
  'Or did the ladies and gentlemen of the party leaderships think that dealing with Antwerp would solve everything?' (LN_20030405_01.data.dz:937)

- 2 cases of *varkentjes* being deeply embedded, e.g. (22). The problem here is that *varkentjes* is a modifier to a quantity noun (*aantal* 'number') and is itself modified by an adjective, resulting in an NP inside an NP and the loss of the object relation between *varkentjes* and *wassen*. Alpino can analyse constructions such as these in two ways: one where it takes *aantal* as the head of the NP (which is what happens in this parse in question); the other where it takes *varkentjes* as the head of the NP with *een flink aantal* as a complex determiner, in which case MWE-Finder would have been able to find this instance of the target MWE. Whether or not Alpino disambiguated wrongly between these two analyses, MWE-Finder could be made more robust for constructions such as these, so that it can find instances with the former analysis, too.

  (22) Maar wie dit jaar Brad Pitt of Julia Roberts van nabij wil monsteren, zal eerst [een flink
  but who this year Brad Pitt or Julia Roberts from near wants inspect will first a large
  aantal [bureaucratische **varkentjes**]$_{NP}$]$_{NP}$ moeten **wassen**.
  number bureaucratic pigs.DIM must wash
  'But those who want to inspect Brad Pitt or Julia Roberts from up close this year, must deal with quite a few bureaucratic hurdles first.' (DM_20020206_01.data.dz:2169)

- 2 cases of an adjectivized past participle, e.g. (23), which were missed because Alpino does not explicitly label the object relation between the head noun and the attributive past participle.

  (23) Allicht, want Beringen, zou een snel **gewassen varkentje** worden.
  obviously for Beringen would a quickly washed pig.DIM become
  'Obviously, for Bering would quickly be dealt with.' (NB_20010417_01.data.dz:9265)

- one case of a modal infinitive, expressing some notion of ability or obligation (Broekhuis & Keizer, 2015), as in (24).

(24) ≪Nochtans moeten wij ons niet zenuwachtig maken want zowel Neerlandia als Hoboken 2000
nevertheless must we us not nervous make for both Neerlandia as Hoboken 2000
zijn te **wassen varkentjes**.
are to wash pigs.DIM

'Nevertheless, we must not get nervous, for both Neerlandia and Hoboken 2000 are to be dealt with.' (LN_20061209_01.data.dz:15625)

Finally, there were 3 cases of a passive perfect without a copula, e.g. (25). In Dutch the passive perfect uses the copula *zijn* 'to be' as auxiliary, which in specific cases may be absent. 2 out of the 3 hits were additionally embedded within an accusativus cum infinitivo (AcI) construction, e.g. (26). For us the authors, all three hits are only marginally grammatical, but they may be considered more grammatical in Belgian Dutch.[15]

(25) ? Eens dat **varkentje gewassen**, heeft Frank de handen vrij en dat zal niet alleen zijn om zich in
once that pig.DIM washed has Frank the hands free and that will not only be for SELF in
het nachtleven te werpen.
the nightlife to throw

'Once that has been dealt with, Frank will have his hands free, and that won't be just to immerse himself into the nightlife.' (DM_20050223_01.data.dz:4554)

(26) ? Bij Antonia dacht men het **varkentje gewassen**.
at Antonia thought one the pig.DIM washed

'Those at Antonia thought the problem dealt with.' (NB_19981014_01.data.dz:3343)

These 78 misses therefore do not suggest an adjustment of the canonical form, but rather point to possible points of improvement for MWE-Finder's algorithm, which will require some form of enrichment of the parses given by Alpino or an extension of the query generation mechanism. Several infrequent Dutch constructions lead to the loss of the object relation of *varken* to *wassen* in Alpino's parse, and indeed of any direct object part of a verbal MWE. The implementation of these constructions in MWE-Finder remains a *varkentje om te wassen* in itself.

5.3 Iemand zal iemand het hemd van het lijf vragen

As a final illustration we search for the MWE in (27) as target MWE:

(27) iemand zal iemand het hemd van het lijf vragen
someone will someone the shirt from the body ask

'someone will want to know all the ins and outs of something from someone'

We search in Mediargus. The results of the three queries are summarised in (28):

(28) Results of the queries derived from *iemand zal iemand het hemd van het lijf vragen*:

Query	**Matches**
MEQ	30
NMQ	32
MLQ	43

All 11 examples from the MLQ results minus the NMQ results contain an occurrence of the target MWE. 10 sentences have been parsed completely wrongly by Alpino. In one sentence Alpino produces a different parse than expected, which prevents recognition of the target MWE, but additionally, one of the MWE components is modified (29):

[15] It could be argued that (26) is missing some form of punctuation between *men* and *het*. This would render the analysis of the sentence rather different with a speech tag at the start of the sentence, and would allow MWE-Finder to find this MWE with the MEQ or NMQ. The reason to argue for this is the sentence's marginal grammaticality for the authors if it is analysed as an AcI, however more instances of an AcI embedded under the verb *denken* were attested in the corpus, suggesting that it may be considered more grammatical in Belgian Dutch and corroborating the AcI analysis. Alpino, though, remains unable to parse it as an AcI.

(29) ... **vroegen** ze Marc **het** bezwete **hemd van het lijf**.
... asked they Marc the sweating shirt from the body
'... they wanted to know all the ins and outs from sweating Marc.'
(NB_20010104_01.data.dz:5254)

It is not clear that we should change the canonical form for the target MWE based on this single example. It seems that the modifier does not really semantically modify *hemd* (this sentence can be uttered even if Marc is not wearing a shirt) but rather *Marc*.

In example (30) a variant of the expression occurs, with *de hemd* instead of *het hemd*. It is unclear to us whether this is a performance error or a real variant of the expression. The example is also in passive, and there is agreement between the indirect object and the finite verb. The latter is ill-formed according to normative grammar though it occurs often in colloquial speech. Despite the fact that Alpino cannot analyse this aspect correctly, the parse is sufficiently good to recognize this variant of the target MWE.

(30) ... werden de klanten **de hemd van het lijf gevraagd** ...
... were the customers the shirt from the body asked ...
'... they wanted to get to know all ins and outs from the customers...'
(LN_20030821_01.data.dz:10761)

Of the 2 results in the NMQ results minus the MEQ results, one sentence is parsed incorrectly, and in the other sentence (31) a variant of the expression occurs with *zijn lijf* 'his body' instead of *het lijf* 'the body'. We find this expression well-formed, so we should include it in a new version of DUCAME.

(31) Ik **vroeg** hem **het hemd van zijn lijf**.
I asked him the shirt from his body
'I wanted to know all ins and outs from him.' (NB_19990710_01.data.dz:14448)

All sentences from the MEQ results contain the target MWE. There are rather complex sentences among them, e.g. one with a large verb cluster (32a) and one in passive voice (32b).

(32) a. ... een journalist die je al zo vaak **het hemd van het lijf** heeft proberen te **vragen**.
... a journalist who you already so often the shirt from the body has tried to ask
'... a journalist that has already often tried to get to know all ins and outs from you.'
(DM_19980206_01.data.dz:305)

b. ... telkens weer wordt je **het hemd van het lijf gevraagd**.
... each time again is you the shirt from the body asked
'... again and again they want to get to know all ins and outs from you.'
(DS_20010905_01.data.dz:4124)

6 Related Work

There has been abundant research in developing automatic tools to reliably search for occurrences of MWEs (MWE Identification in the sense of (Constant et al., 2017). An excellent survey of recent work on MWE identification is (Ramisch et al., 2023). It mostly describes and evaluates work done in the context of the 2016 SEMEVAL DiMSUM Shared Task[16] (Schneider et al., 2016) or PARSEME Shared Tasks (Ramisch et al., 2018, 2020; Savary et al., 2017).

We want to highlight here the commonalities and differences of the approach taken in this paper in comparison to most other work.

First of all, all works on MWE identification describe software tools to annotate text for MWEs. None of the works describe a software application with a dedicated user interface and an identified target user group. MWE-Finder is a web application with a user interface targeted at linguists and lexicographers that do research on MWEs or try to describe them.

Second, none of the works on MWE identification target the Dutch language. To our knowledge, MWE-Finder is the first piece of software to identify Dutch MWEs. So far, linguists and lexicographers

[16] https://dimsum16.github.io/

		found by MEQ?				
		Yes	**No**	**Total**	**Acc.**	93.7%
	Yes	895	87	982	**Prec.**	99.7%
contains MWE?	**No**	3	475	478	**Rec.**	91.1%
	Unclear	0	2	2	$\mathbf{F_1}$	95.2%
	Total	898	562	1462		

Table 3: Results of the MEQs for the three case-study MWEs. Note that **No** also includes literal readings of the target MWEs.

investigating Dutch MWEs defined queries in OpenSoNaR or similar interfaces (Nederlab (Brugman et al., 2016), Corpus of Contemporary Dutch[17]) to search for the lemma's of an MWE in each other's neighborhood within a sentence, basically equivalent to the major lemma query of MWE-Finder. These corpus search applications provide no special facilities for searching for MWEs.

Though MWE-Finder has been developed for Dutch, the general approach is not restricted to the Dutch language. However, this is not the place to discuss this in detail. Instead we refer to (Odijk et al., 2024), which describes what would be needed to create an MWE-Finder for other languages.

Ramisch et al. (2023, p. 106) classify approaches in a number of categories (paradigms): syntactic parsing, compositionality prediction of MWE candidates, and sequence annotation. Our work falls in the syntactic parsing paradigm, and is in this respect comparable to the work by Nagy T. and Vincze (2014) focusing on English verb particle constructions and Constant and Nivre (2016), who developed a general strategy for MWE identification and tested it on French and English data.

Many works focus on specific MWE subclasses, e.g. the PARSEME-related work focused on verbal MWEs, Nagy T. and Vincze (2014) focused on verb particle constructions. MWE-Finder aims to cover all MWE classes, especially the most difficult class of flexible MWEs.

Finally, in MWE-Finder currently nothing special is done to distinguish the MWE-interpretation of an expression from a literal interpretation of the expression.

7 Discussion and conclusion

In this paper we have showcased MWE-Finder's performance through the use of three small case studies of Dutch MWEs.

We observed that MWE-Finder is very accurate in retrieving the target MWE using the MEQ. Of all sentences in the corpus containing at least the relevant lemmas of the content words (i.e., the results of the MLQs), the MEQ correctly found 895 of them to contain the target MWE while it also correctly identified 475 not to contain the target MWE. Out of the total 1462 (remember that there were 2 hits for which it remained unclear whether they did or did not contain the MWE; see Table 2), this results in an accuracy for the MEQ of 93.7%. In terms of precision and recall, the MEQ shows a precision for these three case studies of 99.7%, finding only 3 sentences that did not concern an instance of the target MWE. Recall is slightly lower at 91.1%, missing 87 cases of the target MWE. An F_1-score is then calculated as 95.2%, see Table 3. Note that these numbers also include literal readings, and our numbers therefore nicely reflect (Savary et al., 2019), which found that when syntactic conditions necessary for an idiomatic reading are fulfilled, this reading occurs in 96% to 98% of the cases.

Most false negatives, i.e. the instances of the target MWEs that the MEQ missed, can be attributed to wrong parses, an issue that will likely remain to impair the results of MWE-Finder as it depends on Alpino and the cleanliness of the data. While not many, the false positives, i.e. the sentences that the MEQ did retrieve despite them not containing the target MWE, can be attributed to the fact that MWE-Finder was not designed to distinguish between a metaphorical and literal reading of the components of the MWE.

The results of the NMQ showed its effectiveness in guiding the user in the formulation of a better,

[17] https://ivdnt.org/corpora-lexica/corpus-hedendaags-nederlands/

more complete description of the MWE in a canonical form. For instance, it suggested the addition of the canonical form of *iemand zal iemand het hemd van het lijf vragen*, given that it retrieved a well-formed variant of the MWE with the possessive pronoun *zijn* 'his' instead of *het* before *lijf*. It may be expected that the performance of MWE-Finder's MEQ will be even higher when using the updated canonical forms.

The MLQ proved useful in identifying points of improvement for MWE-Finder's system. While the wrong parses produced by Alpino may not be mitigated in the near future, especially the case study on *iemand zal 0dat *+varken wassen* laid bare that in quite a few (though infrequent) constructions it fails to recognise the object relation between *varken* and *wassen*, a problem likely to arise in other MWEs as well. MWE-Finder can thus be improved by, e.g., enriching the Alpino parses by making the object relation in such constructions explicit. Also the issue concerning the labelling of relative pronouns (`vg` vs. `vnw`) is an obvious point of improvement, as well as the reformulation of the way the NMQ is generated and how it deals with adpositional complements. Another valuable addition can be found in a way to check for the presence of NPLs.

There is, though, one caveat with regards to the performance as presented in Table 3: it may be – and probably is – the case that the MLQ has missed cases of the target MWEs, for instance where a homonym received a wrong POS tag or was lemmatised wrongly, as was already pointed out in Section 5. The MLQ specifically searches for the lemmas of the content words in the MWE, along with their POS tags as derived from the canonical form, and will miss any such case that was tagged otherwise. The results presented in Table 3 may therefore be somewhat skewed, but the argument that MWE-Finder works well, we believe, still holds, also considering that the potential cases missed by the MLQ would be the result of Alpino's parse, not MWE-Finder's algorithm directly. As said before, however, we are working on the formulation of a fourth query (the Related Word Query, RWQ) that is even more relaxed than the MLQ in order to retrieve candidates with wrongly tagged content words.

Using MWE-Finder does sometimes entail closer, manual inspection of the data – also in the case studies presented in this paper. This closer inspection can be streamlined with a special analysis component built into MWE-Finder. A preliminary version of such a component already exists, but will require more testing before being definitively implemented into the application.

In the future MWE-Finder's performance may be evaluated more formally using the Dutch MWE-corpus currently being produced by Bouma et al. (2024). For now, we believe that the current case studies presented in this paper provide a simple evaluation, showcasing and roadmap for further development of MWE-Finder.

References

Augustinus, L., Vandeghinste, V., & Van Eynde, F. (2012, May). Example-based treebank querying. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC 2012)* (pp. 3161–3167). European Language Resources Association (ELRA).

Bouma, G., Odijk, J., & Tiberius, C. (2024, February). Towards a Dutch Parseme corpus. In *Proceedings of the second general UniDive meeting, Naples, Italy*. UniDive Project. https://unidive.lisn.upsaclay.fr/doku.php?id=meetings:general_meetings:2nd_unidive_general_meeting

Broekhuis, H., Corver, N., & Vos, R. (2016, January). 3.2.2.5. The reflexive middle construction [Retrieved January 29, 2024 from https://taalportaal.org/taalportaal/topic/pid/topic-14406719669673231]. https://taalportaal.org/taalportaal/topic/pid/topic-14406719669673231

Broekhuis, H., & Keizer, E. (2015, December). 3.2.III. Modal infinitives [Retrieved January 29, 2024 from https://taalportaal.org/taalportaal/topic/pid/topic-14286557605977117]. https://taalportaal.org/taalportaal/topic/pid/topic-14286557605977117

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Tjong Kim Sang, E., & van den Bosch, A. (2016, May). Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard,

J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* (pp. 1277–1281). European Language Resources Association (ELRA).

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, *43*(4), 837–892. https://doi.org/10.1162/COLI_a_00302

Constant, M., & Nivre, J. (2016, August). A transition-based system for joint lexical and syntactic analysis. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 161–171). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1016

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating research environments with BlackLab. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 245–257). Ubiquity. https://doi.org/http://dx.doi.org/10.5334/bbi.20

Kruyskamp, C. (1974). *Groot woordenboek der Nederlandse taal* (10th ed.). Martinus Nijhoff.

Nagy T., I., & Vincze, V. (2014, April). VPCTagger: Detecting verb-particle constructions with syntax-based methods. In V. Kordoni, M. Egg, A. Savary, E. Wehrli, & S. Evert (Eds.), *Proceedings of the 10th workshop on multiword expressions (MWE)* (pp. 17–25). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-0803

Odijk, J. (2013). Identification and lexical representation of multiword expressions. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch. Results by the STEVIN-programme* (pp. 201–217). Springer. http://link.springer.com/content/pdf/10.1007

Odijk, J. (2023). A canonical form for Dutch multiword expressions (version 1.0) [Part of the DUCAME documentation]. https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP

Odijk, J., & Kroon, M. (2024, May). A canonical form for flexible multiword expressions. In *Proceedings of LREC-COLING 2024*. European Language Resources Association (ELRA).

Odijk, J., Kroon, M., Baarda, T., Bonfil, B., & Spoel, S. (2024). MWE-Finder: Querying for multiword expressions in large Dutch text corpora. In V. Giouli & V. B. Mititelu (Eds.), *Multiword expressions in lexical resources. Linguistic, lexicographic and computational perspectives*. Language Science Press.

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The construction of a 500 million word reference corpus of contemporary written Dutch [http://link.springer.com/book/10.1007/978-3-642-30910-6/page/1]. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch: Results by the STEVIN-programme* (pp. 219–247). Springer.

Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., & Baayen, H. (2002). Experiences from the Spoken Dutch Corpus project. In *Proceedings of the third international conference on language resources and evaluation (LREC-2002)* (pp. 340–347). ELRA.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., … Walsh, A. (2018, August). Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In A. Savary, C. Ramisch, J. D. Hwang, N. Schneider, M. Andresen, S. Pradhan, & M. R. L. Petruck (Eds.), *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (LAW-MWE-CxG-2018)* (pp. 222–240). Association for Computational Linguistics. https://aclanthology.org/W18-4925

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., & Xu, H. (2020, December). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In S. Markantonatou, J. McCrae, J. Mitrović, C. Tiberius, C. Ramisch, A. Vaidya, P. Osenova, & A. Savary (Eds.), *Proceedings of the joint workshop on multiword expressions and electronic lexicons* (pp. 107–118). Association for Computational Linguistics. https://aclanthology.org/2020.mwe-1.14

Ramisch, C., Walsh, A., Blanchard, T., & Taslimipoor, S. (2023, May). A survey of MWE identification experiments: The devil is in the details. In A. Bhatia, K. Evang, M. Garcia, V. Giouli, L. Han, & S. Taslimipoor (Eds.), *Proceedings of the 19th workshop on multiword expressions (MWE 2023)* (pp. 106–120). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.mwe-1.15

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2001). Multiword expressions: A pain in the neck for NLP. *LinGO Working Paper*, *2001-03*. http://lingo.stanford.edu/csli/pubs/WP-2001-03.ps.gz

Savary, A., Cordeiro, S., Lichte, T., Ramisch, C., Iñurrieta, U., & Giouli, V. (2019). Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics*, *112*(1), 5–54. https://doi.org/doi:10.2478/pralin-2019-0001

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., & Doucet, A. (2017, April). The PARSEME shared task on automatic identification of verbal multiword expressions. In S. Markantonatou, C. Ramisch, A. Savary, & V. Vincze (Eds.), *Proceedings of the 13th workshop on multiword expressions (MWE 2017)* (pp. 31–47). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1704

Schneider, N., Hovy, D., Johannsen, A., & Carpuat, M. (2016, June). SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 546–559). Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1084

van de Camp, M., Reynaert, M., & Oostdijk, N. (2017). WhiteLab 2.0: A web interface for corpus exploitation. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 231–243). Ubiquity. https://doi.org/http://dx.doi.org/10.5334/bbi.19

van Noord, G. (2006). At last parsing is now operational. In P. Mertens, C. Fairon, A. Dister, & P. Watrin (Eds.), *TALN06 verbum ex machina. Actes de la 13e conférence sur le traitement automatique des langues naturelles* (pp. 20–42).

van Noord, G., Bouma, G., Van Eynde, F., de Kok, D., van der Linde, J., Schuurman, I., Tjong Kim Sang, E., & Vandeghinste, V. (2013). Large scale syntactic annotation of written Dutch: Lassy. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch* (pp. 147–164). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30910-6_9

Appendix: Queries

In this appendix we list the queries for the MWE *iemand zal 0dat *varken wassen*. For this particular example, the MEQ and NMQ queries are identical.

MEQ
```
//node[
    node[@rel="obj1" and @cat="np" and
        node[@lemma="varken" and @rel="hd" and
             @pt="n" and @ntype="soort" and
             (@genus="onz" or @getal="mv")]] and
    node[@lemma="wassen" and @rel="hd" and @pt="ww"]]
```

NMQ
```
//node[
    node[@rel="obj1" and @cat="np" and
        node[@lemma="varken" and @rel="hd" and
             @pt="n" and @ntype="soort" and
             (@genus="onz" or @getal="mv")]] and
    node[@lemma="wassen" and @rel="hd" and @pt="ww"]]
```

MLQ
```
//node[@lemma="varken" and @pt="n"]/ancestor::alpino_ds/
    node[@cat="top" and
         descendant::node[@lemma="wassen" and @pt="ww"]]
```

Protective Measures for Sharing the Finnish Dark Web Marketplace Corpus (FINDarC)

Krister Lindén
Department of Digital Humanities
University of Helsinki, Finland
krister.linden@helsinki.fi

Teemu Ruokolainen
Faculty of Information Technology
and Communication Sciences
Tampere University
Tampere, Finland
teemu.ruokolainen@tuni.fi

Lasse Hämäläinen
Faculty of Information Technology
and Communication Sciences
Tampere University
Tampere, Finland
lasse.hamalainen@tuni.fi

J. Tuomas Harviainen
Faculty of Information Technology
and Communication Sciences
Tampere University
Tampere, Finland
tuomas.harviainen@tuni.fi

Martin Matthiesen
CSC - IT Center for Science
Espoo, Finland
martin.matthiesen@csc.fi

Mietta Lennes
Department of Digital Humanities
University of Helsinki, Finland
mietta.lennes@csc.fi

Abstract

We discuss the archiving procedure of a corpus comprising posts submitted to Torilauta, a Finnish dark web marketplace website. The site was active from 2017 to 2021 and during this time one of the most prominent online illegal narcotics markets in Finland. A reduced version of the corpus, Finnish Dark Web Marketplace Corpus (FINDarC), has been archived in the Language Bank of Finland. In the current work, we focus on the protective measures for storing the data and how researchers can apply for access rights to the corpus under the CLARIN RES licence.

1 Introduction and Background

Torilauta was a dark web marketplace website. The site was active from 2017 to 2021 and during this time one of the most prominent online illegal narcotics markets in Finland. Functionally, the site consisted of discussion imageboards where vendors and customers were able to set up instances of face-to-face trading, typically with the assistance of instant messaging software such as Wickr or Telegram. The original, unmodified data set comprising 3,104,976 posts was collected and handed over to the ENNCODE consortium[1] by the site administration to be archived and shared for research purposes, as permitted by the site's Terms of Service. To promote the FAIR data principles, a reduced version of the corpus comprising 3,104,515 posts, referred to as the Finnish Dark Web Marketplace Corpus (FINDarC), has been deposited in the Language Bank of Finland, a language resource service coordinated by the national FIN-CLARIN consortium formed by Finnish universities and other research organizations. Researchers can contact the Language Bank and apply for permission to access the corpus under the CLARIN RES license offering a time-restricted personal license to re-use the data according to an approved research plan.[2]

While the dark web online market places, including Torilauta, emphasize user anonymity, the posts submitted to such sites can nevertheless contain personal information, such as unique usernames and personal names, enabling data subject re-identification. Therefore, as described in previous papers (Lindén et al., 2023a, 2023b), we have made our best effort to assess and identify the type and amount of personal information in the original unmodified data set, to assess and implement viable data anonymization/reduction approaches, to assess privacy and security measures implemented by the Language Bank

[1] Consortium website: https://research.tuni.fi/enncode/

[2] Permanent link to the corpus: http://urn.fi/urn:nbn:fi:lb-2022062221

Krister Lindén, Teemu Ruokolainen, Lasse Hämäläinen, J. Tuomas Harviainen, Martin Matthiesen and Mietta Lennes 2023. Protective Measures for Sharing the Finnish Dark Web Marketplace Corpus (FINDarC). *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 143–152. DOI: https://doi.org/10.3384/ecp210

of Finland, and to put in place a future corpus management plan coordinated by the Language Bank of Finland. In this paper, we add information on the protective measures that are applied when storing the data in the Language Bank of Finland and what researchers need to do to get access to the data.

Those carrying out future research based on the corpus are encouraged to implement appropriate ethical proofreading measures (see e.g. (Harviainen et al., 2021)) in order to further mitigate any potential harm from access to the material, to both the researchers and the studied populations.

2 Related Work

In this section, we discuss previously published studies using the Torilauta site as a data source and related corpora in Sections 2.1 and 2.2, respectively. For a discussion of the privacy-utility trade-off within the data privacy literature, see Alvim et al. (2011) and Li and Li (2009).

2.1 Previous Studies on Torilauta

Prior to the data publication presented here, Torilauta was utilized in the ENNCODE project as a data source in multiple linguistic and social science studies (Haasio et al., 2020; Hämäläinen & Ruokolainen, 2021; Hämäläinen et al., 2021; Harviainen et al., 2020; Karjalainen et al., 2021). In particular, Haasio et al. (2020) examined information needs of drug users using a sample of 9,300 posts.[3] Harviainen et al. (2020) studied cultural and socioeconomic aspects of drug traders using the same 9,300 post sample. Hämäläinen and Ruokolainen (2021) studied narcotic substance vocabulary based on a sample of 3,000 posts. Hämäläinen et al. (2021) studied a sample of 1,654 usernames extracted from posts submitted to the site. Karjalainen et al. (2021) examined the availability of illegal narcotics during the first wave of the COVID-19 pandemic using a sample of 535 posts.

It is notable that none of the previous studies attempted to share their data sets with the research community in a systematic manner. This practice negatively impacts the replication and verification of the published studies and potentially discourages further research on the topic. On the other hand, given the sensitive and potentially incriminating nature of the data, not releasing the data is an understandable approach since preparing and managing such a resource gives rise to multiple technical, ethical, and legal challenges. The purpose of this paper is to describe and discuss these challenges and how we approached them.

2.2 Related Corpora

To the best of our knowledge, there exist relatively few published dark web corpora or text data sets. Three notable exceptions include the Dark Net Market archives (2013–2015) (Branwen et al., 2015), a collection covering 89 dark net markets and over 37 related forums (1.6TB uncompressed) scraped during 2013-2015, DUTA (Nabki et al., 2017), a set of 7,000 text samples formed by sampling the Tor network for two months, and CoDa (Jin et al., 2022), a set of 10,000 web documents tailored towards text-based dark web analysis. All three corpora comprise primarily English texts and are either publicly downloadable (Dark Net Market archives) or available to researchers upon request (DUTA, CoDa).

Existing Finnish web forum corpora include texts collected from the Ylilauta imageboard (Ylilauta, 2016) and Suomi24 social networking site (City Digital Group, 2021). While emphasizing user anonymity, both Ylilauta and Suomi24 forums operate on the clear web and strictly forbid illegal content. Both corpora are available for research purposes via the Language Bank of Finland under a Creative Commons (CC BY-NC 4.0) license.

The Finnish Internet Parsebank (Laippala & Ginter, 2014) is a large-scale syntactically analyzed text collection created using plain text webpage data made available by the Common-Crawl2 Internet crawl project. Due to the employed web crawling approach to data collection, the corpus is likely to contain web forum content.

In a recent study, Leedham et al. (2021) discussed their work on archiving a hard-to-access WiSP corpus consisting of texts written by social work professionals describing their work practices. Due to the potentially sensitive nature of the texts, Leedham et al. (2021) created two versions of the corpus:

[3] In their paper, Haasio et al. (2020) refer to *Torilauta* using its other commonly used name *Sipulitori*.

one for the research project and an anonymized/reduced version for archiving. In a similar vein, our work presented here aims to provide an extensive discussion of the process of preparing a corpus of potentially sensitive texts for archiving and sharing.

3 Data Set Description

This section provides an overview of the data as a whole and discusses the post lifespans in more detail.

3.1 Overview

Table 1: Data fields comprising a single post. The column titled *missing (%)* indicates the portion of all posts where the field value is not available.

	description	example	missing (%)
boardUri	board identifier	roi	0.0
creation	post creation datetime (UTC)	2020-01-14T17:51:24.714Z	0.0
deletion	post deletion datetime (UTC)	2020-01-27T16:49:03.663Z	2.8
threadId	thread identifier	27961	0.0
postId	post identifier	28069	29.8
name	poster name	example-name	54.1
subject	message subject	Example message subject	46.2
message	message text body	Example message text body	0.0

Each post in the corpus is represented as a data structure with 8 fields as shown in Table 1. Each field belongs to one of the following three types: a string, an integer, or a date. All dates are in timezone UTC+0 (GMT+0). Note that throughout this paper, we refer to a set of these 8 data fields as *post*, whereas the content of the text data field within a single post is referred to as *message* or *text body*.

The original data set received by the consortium included all posts submitted to Torilauta between 2019-09-11 and 2020-05-20 (1,863,639 posts in 251 days) and 2020-06-17 and 2020-10-31 (1,099,710 posts in 136 days). In addition to the posts collected during these active collection periods, the data contained "residue" posts submitted between 2017-11-02 and 2019-09-11 (141,627 posts in 678 days). Meanwhile, posts submitted between 2020-05-20 and 2020-06-17 were missing completely. Therefore, the original unmodified corpus consisted of 3,104,976 posts in total.

The data is grouped by boards. Of the 32 boards, the board with the highest activity measured by the total number of submitted posts and threads was the market board dedicated to narcotics transactions within the city of Helsinki (*/hki*). The total number of posts submitted to this board was 787,459 corresponding to 25.4% of all posts in the data. Meanwhile, in total 96.5% (2,997,624) of all posts were submitted to the 16 boards dedicated to transactions.

Missing values have different meanings depending on the field. For deletions, a missing value means the post was never deleted. The first post of a thread, referred to as the original post (OP), always has a missing postId value and is instead identified by its (boardUri, threadId) pair. The subject field is missing for 46.2% of all posts since it was common practice to omit the subject. Similarly, the poster name field is missing for 54.1% of all posts which is in line with the anonymous nature of the site and since any optional contact information, such as an instant messenger username, was often included in the message text body instead.

The posts submitted to Torilauta optionally contained an attached image. However, no images were included in the originally deposited data set. Moreover, the data fields comprising a post did not include information on whether the post had contained an image or not.

3.2 Post Lifespans

Submitted posts were deleted from the site for three main reasons. First, the site hosted a fixed number of threads on each board at a given time and so inactive threads were regularly removed by an automatic pruning mechanism to make room for new, active threads. Second, posts which violated the site rules (e.g. spam) were removed by the site administration. Third, the site interface did not provide users with means to edit messages and, therefore, the only way to correct erroneous message content (e.g. typos,

updates) was to delete the post and resubmit. Lastly, a small portion of posts were ”pinned” by the site administration, that is, they were meant to stay available on the site indefinitely.

There are two known caveats related to the deletion timestamps. First, while the data set included the creation and deletion times of posts, it unfortunately did not include information about the reason for the deletion. Second, all posts deleted during the pause in collection 2020-05-20 - 2020-06-17 had their deletion value marked as missing and, therefore, appeared as if they were not deleted. The amount of these potentially erroneous missing values was, however, relatively small and 97.22% of all posts (3,104,976) in the data had reliable deletion time information.

Finally, we estimated the median lifespans of submissions to the market and non-market boards to be 23 and 238 hours, respectively. The difference was mainly due to the lower posting frequency and consequently lower thread pruning frequency of the non-market boards. For noise filtering purposes, we are mostly interested in the messages with short lifespans. To this end, we note that 5% of all messages had a lifespan of less than 32 minutes. Since posts with such short lifespans were likely removed by the user and resubmitted after minor modifications, they may be discarded as noise.[4]

4 Data release

Text anonymization approaches proposed in the literature commonly utilize automatic named-entity recognition (NER) as a part of the processing pipelines to varying extents (Adams et al., 2019; Csányi et al., 2021; F. & Trabelsi, 2018; Francopoulo & Schaub, 2020; Garat & Wonsever, 2022; Glaser et al., 2021; Oksanen et al., 2019; Tamper et al., n.d.). Ideally, NER tools would also have been useful when processing FINDarC. However, examining the prediction quality on a manually annotated test section of the data set, suggested that the available tools suffered from a domain mismatch in addition to the inherent mismatch between personal data and named entity classes. This was not completely surprising since the text domain also caused problems for human annotators when creating the test section. Because the available tools tended to miss entities of interest (low recall) and be incorrect when detecting entities (low precision), we did not consider them efficient pre-processing tools for FINDarC in their current state.

4.1 Common Personal Identifiers

Instead of using NER tools, we decided to use full-text search, to find common personal identifiers with relatively rigid formats, such as social security numbers and phone numbers. We defined a target set of textual patterns (regular expressions) and searched for matches in message bodies. Specifically, we were interested in finding expressions matching

1. (Finnish) social security numbers
2. (Finnish) phone numbers
3. Email addresses
4. IBAN bank accounts
5. IP addresses

all of which have relatively rigid formats. We applied the search to all posts in the data and assigned the matches manually to personal data and non-personal data according to post context. We did not filter out noise from the data and instead applied the search to all 3,104,976 posts in the original corpus.

4.2 Regular Expressions

In what follows, we provide brief descriptions of the applied regular expressions.

Social Security Numbers. The Finnish social security number (SSN) is a sequence of 11 characters assigned to individuals by the Finnish government based on their date of birth and gender. The first 10 characters of the sequence are 6 numbers (date of birth) followed by a hyphen or A, followed by 3 numbers. The last character is alphanumeric, i.e., a number or a letter. Valid sequences likely have,

[4]This is in agreement with the recommendation of the site administrator.

therefore, format ”121212-1234” and ”121212-123A”. We detect the sequences using the regular expression `\d\d\d\d\d\d\-\d\d\d[a-zA-Z0-9]`. Persons born in the 2000s, who would have an ”A” instead of hyphen, were not found in the sample.

Phone Numbers. According to the specification of the Finnish telephone network numbering, Finnish mobile phone numbers begin with a routing number (04-, 050, or 059) and are followed by a subscriber number, such as, ”040 1234567”, ”059 1234567”, and so forth.[5] The first zero (”0”) of the number can optionally be replaced by the country code of Finland +358 (e.g. ”+358 40 1234123”, ”+358 59 4321432”, etc.). Based on a preliminary examination of the data set, we detect common phone number formats using two regular expressions: `[\+]?358[\-\s]?0[45][\-\s]?\d\d\d[\-\s]?\d[\-\s]?\d\d\d` which matches numbers starting with the country code and `0[45]\d[\s\-]?\d\d\d[\-\s]?\d[\-\s]?\d\d\d` which detects numbers with the country code omitted. Moreover, the expressions detect most commonly used grouping patterns using hyphens (e.g. ”+358-40-12345-567”) and whitespaces (e.g. ”059 123 4567”). While the subscriber part of the number can, in principle, vary in length, the patterns match the most common length of 7 digits. Landline numbers would be shorter but follow the same principles; none were however found in the data.

Email Addresses. According to the RFC 5322 standard[6], an email address as an identifier which contains a locally interpreted string followed by the at-character (”@”) followed by an internet domain, such as ”name@domain.com”, ”firstname.surname@subdomain.domain.com”, and ”underscore-hyphen-plus+sign@domain.com”. We detect the addresses using a regular expression `\S+\@\S+\.\S+` which successfully detects all the above examples from a running text.

IBAN Bank Accounts. We search for bank account numbers matching the International Bank Account Number (IBAN) structure specified by the ISO 13616-1:2020 standard [7]. The IBAN formatted numbers consist of the Finnish bank account number (14 digits) preceded by a two letter country code (”FI” for Finland) and two check digits (e.g. ”FI72 1234 5678 1234 12”). We detect the pattern using the regular expression `[Ff][Ii]\d\d[\s\-]?\d\d\d\d[\s\-]?\d\d\d\d[\s\-]?\d\d\d\d[\s\-]?\d\d` which takes into consideration the letter case of the country code and the commonly used grouping whitespaces.

IP Addresses. IP (internet protocol) addresses are unique addresses which identity devices on the internet and local networks. We search for IP addresses using the following regular expression `(25[0-5]\2[0-4][0-9]—[01]?[0-9][0-9]?)̇3(25[0-5]—2[0-4][0-9]—[01]?[0-9][0-9]?—` which matches patterns such as 88.777.66.555 and so forth.

4.3 Search Results

The frequencies of matched social security numbers, phone numbers, email addresses, bank account numbers, and IP addresses are presented in Table 2, which shows that the most and least frequent matched types were email addresses and bank account numbers with 1,840 and 12 regular expression matches, respectively. Due to the sufficiently low number of original matches, we were able to perform manual verification of all the cases.

The phone numbers and email addresses occurred in two contexts. First, similarly to the instant messaging usernames , 491 out of 858 and 1,622 out of 1,837 of the phone numbers and email addresses, respectively, were posted as contact information by the individuals themselves. The remaining cases were posted as a means of targeting people. In such cases, personal details (e.g., name, relationship information, area of residence) were shared in connection with one or more usernames, in order to paint the person as a potential target for violence. Bank account numbers occurred similarly in two contexts.

[5] Specification of numbers in the Finnish phone network is available at: https://www.finlex.fi/fi/viranomaiset/normi/480001/47180

[6] The RFC 5322 specification is available at: https://datatracker.ietf.org/doc/html/rfc5322

[7] //www.iso.org/standard/81090.html

Table 2: Matched regular expression frequencies. The columns titled *matches* and *verified* denote the number of found regular expression matches and the number of manually verified cases, respectively, The columns titled *posts* and *threads* denote the number of distinct posts and threads where the verified cases occurred.

	matches	verified	posts
phone	875	858	699
hetu	91	73	65
email	1,840	1,837	1,707
iban	12	12	12
ip_address	121	16	14
total	2,939	2,796	2,261

Out of the 16 IP addresses, 10 cases were included as a means of targeting, while the remaining 6 were provided as a type of contact information. Finally, all 73 and 12 found cases of social security numbers and bank account numbers were posted with a purpose of targeting. Thus, we identified and removed in total 667 cases of targeting by removing 295 posts using this method. Finally, we created a second regular expression list using words and prefixes related to the personal information contained in the identified 295 targeting posts. This list consisted of 77 keywords and parts of person names and addresses.[8] After performing a second search with these patterns and a subsequent manual inspection, we identified and removed the additional 166 posts submitted as a means of targeting. In conclusion, posts with personal information concerning the submitting individual were kept while 461 posts aimed at targeting others were removed.

5 Protective measures

In the following, we discuss the consequences of anonymizing or pseudonymizing the data, before we explore an alternative approach.

5.1 Consequences of Anonymization

Conventionally, the most direct approach to protecting data subjects from re-identification has been to anonymize the data by removing/obscuring the parts containing personal information (Ohm, 2009). This process aims at potentially being able to release the data to the public. However, it appears evident that, if implemented successfully, this type of processing would have a profound impact on the usefulness of FINDarC for research purposes. For example, subsequent to removing usernames from their post contexts or from the data altogether, one would not be able to replicate the study of Hämäläinen et al. (2021) who examined how sellers and buyers of illegal drugs represent themselves in their usernames. In turn, subsequent to removing location and/or timestamp data, one would no longer be able to replicate the study of Karjalainen et al. (2021) who studied the availability of drugs specifically in the city of Tampere during the COVID-19 epidemic in the spring of 2020. From a utility point of view, therefore, it could be argued that reducing personal information from the buy/sell post threads would quickly degrade, or destroy, the usefulness of the corpus as a data source for research. This problem is generally referred to as the privacy-utility trade-off within the data privacy literature (Alvim et al., 2011; Li & Li, 2009). However, in case anonymisation is not an option, there are other means to protect the data which minimize the risk of additional exposure for the data subjects and which can therefore justify the use of the data. The efficacy of the protective measures can be evaluated with the help of a data protection impact assessment.

Due to the problematic privacy-utility trade-off, we posit here that reducing the FINDarC extensively would not be appropriate even if sufficient resources could be allocated for domain-specific tool development and manual labour. Furthermore, we note that Torilauta and other drug trading sites have

[8] We do not present the list here due to obvious privacy issues.

also been under observation by other parties, including both criminals and law enforcement agencies. Therefore, it is our assessment that, if restricting access to academic research, leaving the sell/buy posts, which form the majority of the FINDarC, largely intact poses few additional risks to the studied populations as they had entered their data for public use. However, in addition to the sell/buy posts, the data also contains posts with the intention of doxxing/targeting individuals. Here, our position is that removing these submissions is warranted from an ethical point of view while not significantly decreasing the value of the corpus as a data source. This is because these posts are not directly related to the main functionality of the site as an online marketplace. Accordingly, we removed all 461 posts containing identified doxxing/targeting information from the corpus The reduced corpus, therefore, comprises 3,104,515 posts.

Finally, as per the Terms of Service of Torilauta, the site users gave consent to data collection for academic use by using the site. Consequently, site users could opt out of the data collection by not submitting new posts and/or contacting the site administration about previously submitted posts. However, it could be argued that by removing a previously submitted post, a user has withdrawn the permission to use the data. Unfortunately, the original data set received from the site administration did not include information about the reasons behind post deletions. Therefore, we were not able to exclude any posts from the corpus based on the deletion status.

5.2 Technical and Organisational Measures to Protect the Data

Due to the limited applicability of data reduction as a means of protecting data subjects from re-identification, we instead need to restrict access to the corpus. Since the FINDarC resource in its current form contains personal data, both copyright and personal data legislation apply and the corpus cannot be published with open access. Instead, FINDarC has protected access under the CLARIN RES licence which means that permission to download and use the corpus is only granted to researchers based on written applications reviewed by the data controller (principal investigator of the ENNCODE consortium) when including a data protection impact assessment of the intended research. The purpose of this limitation is to ensure that the material is accessed only by verified researchers for legitimate research purposes. It also lessens sharing-related risks to both the researchers and the subjects of study, as mandated by the consortium's data management policy.

Whenever researchers wish to deposit a resource in the Language Bank of Finland, the Language Bank (formally represented by the University of Helsinki) negotiates a deposition license agreement with the researcher and/or their home organization, unless the resource to be deposited already has an open license. The agreement defines, e.g., the end-user license in the CLARIN framework, including the resource-specific data protection terms and conditions for resources that contain personal data. In the case of the FINDarC corpus, the original right holder and data controller authorized the Language Bank as the data controller responsible for the redistribution of the resource. Thus, the Language Bank has the right to independently maintain the resource and to process the applications from users without consulting the original right holder. In order for the Language Bank to accept the resource, the original research project was required to assess the potential risks involved in processing the data according to the instructions from their home organisation, and to present documentation of the extent and the rationale behind their process for minimizing personal data, which influenced the need for additional protective measures at the Language Bank.

Extra Security Measures to Protect the Data. In addition to the security measures described in the Language Bank's Core Trust Seal Certificate (see Core Trust Seal, 2022, R16), extra measures to ensure the security of the data have been taken. The data is encrypted at-rest according to the Language Bank encryption guidelines[9]. Each authorized person needs to provide a self-generated public key, the corresponding private key is secured by a password only known to the staff member. At present, a minimum of five members of FIN-CLARIN staff have access to the corpus. A sixth emergency access key is stored in clear text without a password in a physical safe at CSC. A change in personnel requires re-encryption to reach the minimum of authorized staff to be able to access the data, so that a former member of staff

[9]Language Bank Encryption Guideline: https://urn.fi/urn:nbn:fi:lb-202401191

no longer can access the re-encrypted data set. The emergency key and a minimum of five authorized staff ensures that the data is accessible even in case of staff absence and personnel changes.

Security Measures to Ensure Authorized Access. Restricting access to the FinDARC corpus as described in this paper is in line with the current literature on personal data sharing (Elliot et al., 2018, 2020; Ohm, 2009; Rubinstein & Hartzog, 2016; Stalla-Bourdillon & Knight, 2016) which adheres to the FAIR principles, while acknowledging the limitations of data anonymization/reduction and encouraging the use of user group limitations.

At the Language Bank we have implemented restricted access to corpora using the Language Bank Rights service[10], which is based on CSC's REMS service[11]. For FinDARC, an applicant is required to provide an application according to our policy for corpora containing personal data, where she explains the need for the corpus and the associated research project. Since the resource contains personal data, the applicant must also submit a public link to the openly available privacy notice about the personal data processing regarding the research purpose in question. In addition, the applicant needs to supply a self-generated public key via the same Language Bank Rights application in order to be able to receive a copy of the corpus in encrypted form[12]. The application is then evaluated by the data owners who approve or deny access. If access is granted, an authorized member of staff decrypts a copy of the corpus using his private key and re-encrypts it using the key provided by the applicant. The encrypted data set is then sent to the applicant. This method ensures that only the holder of the corresponding private key, the applicant, can open the delivered copy.

Limitations of the Approach. While it is unlikely that all six access keys are lost, it is not impossible. In that case the data is not recoverable from within the Language Bank. Furthermore, we effectively prevent leaking of the data at-rest within the Language Bank and during transit to the researcher, but we cannot guarantee the secure handling of the data after the researcher has decrypted it. However, delivering the data set in encrypted form makes it easier for researchers to keep the data encrypted at-rest since they needed to set up the necessary tools and keys to receive the data in the first place.

Since most Universities use the same access credentials for university email which are used to access Language Bank Rights, identity theft is possible. In such a scenario, the attacker would need to monitor the applicants email account closely to remove automatic messages from the Language Bank Rights application and the Language Bank staff, as the messages might otherwise seem odd to the legitimate researcher. Two factor authentication would further mitigate the risk of identity theft.

6 Conclusions

We have discussed the archiving procedure of FINDarC, a Finnish dark web marketplace corpus, in the Language Bank of Finland. It was unlikely that the corpus could be fully anonymized to be shared publicly without also compromising its value for research, so instead other protective measures were taken to make it possible to share the data. The discussion included an overview of the data, assessment of the risk and impact of data subject re-identification, assessment and implementation of viable data reduction approaches using manual and automatic text processing, assessment of privacy and security measures implemented by the Language Bank of Finland, and a corpus management plan implemented and coordinated by the Language Bank of Finland outlining the protective measures applied in the Language Bank and the justifications a prospective researcher needs to produce to get access to the data to get access to the corpus under the CLARIN RES licence.

Acknowledgments

We acknowledge the funding for the Language Bank and FIN-CLARIN by the Research Council of Finland.

[10] See instruction for RES corpora in https://www.kielipankki.fi/support/access/

[11] For details on REMS see https://urn.fi/urn:nbn:fi:lb-2014120230

[12] See https://urn.fi/urn:nbn:fi:lb-2023051121

References

Adams, A., Aili, E., Aioanei, D., Jonsson, R., Mickelsson, L., Mikmekova, D., Roberts, F., Valencia, J. F., & Wechsler, R. (2019). AnonyMate: A toolkit for anonymizing unstructured chat data. Proceedings of the Workshop on NLP and Pseudonymisation, 1–7.

Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., Degano, P., & Palamidessi, C. (2011). Differential privacy: On the trade-off between utility and information leakage. In International workshop on formal aspects in security and trust (pp. 39–54).

Branwen, G., Christin, N., Décary-Hétu, D., Andersen, R. M., StExo, E. P., Anonymous, D. L., Sohhlz, D. K., Cakic, V., Buskirk, V., Whom, M. M., & Goode, S. (2015). *July* [Dark net market archives, 2011-2015.]. https://www.gwern.net/DNM-archives

City Digital Group. (2021). Suomi24 virkkeet -korpus 2001-2020, Korp-versio. http://urn.fi/urn:nbn:fi:lb-2021101525

Core Trust Seal. (2022). *Core trust seal certificate for the language bank of finland* (tech. rep.). Core Trust Seal. https://www.coretrustseal.org/wp-content/uploads/2022/05/20220530-the-language-bank-of-finland_final.pdf

Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., & Orosz, T. (2021). Challenges and Open Problems of Legal Document Anonymization. Symmetry 13*(8): 1490*.

Elliot, M., Mackey, E., & O'Hara, K. (2020). In *The anonymisation decision-making framework 2nd Edition: European practitioners' guide*.

Elliot, M., O'hara, K., Raab, C., O'Keefe, C. M., Mackey, E., Dibben, C., Gowans, H., Purdam, K., & McCullagh, K. (2018). Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review*, *34(2)*, 204–221.

F., D. C., & Trabelsi, S. (2018). Towards personal data identification and anonymization using machine learning techniques. *In European Conference on Advances in Databases and Information Systems*, 118–126.

Francopoulo, G., & Schaub, L. P. (2020). Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. *workshop on Legal and Ethical Issues (Legal2020)*, 9–14.

Garat, D., & Wonsever, D. (2022). Jan. Automatic Curation of Court Documents: Anonymizing Personal Data. In *Information 2022, vol. 13* (pp. 1–27). https://doi.org/10.3390/INFO13010027

Glaser, I., Schamberger, T., & Matthes, F. (2021). Anonymization of German legal court rulings. In *Proceedings of the 18th international conference on artificial intelligence and law, icail 2021* (pp. 205–209). https://doi.org/10.1145/3462757.3466087

Haasio, A., Harviainen, J. T., & Savolainen, R. (2020). Information needs of drug users on a local dark Web marketplace. *Information Processing and Management*, *57*(2), 1016. https://doi.org/10.1016/j.ipm.2019.102080

Hämäläinen, L., Haasio, A., & Harviainen, J. T. (2021). Usernames on a Finnish Online Marketplace for Illegal Drugs. *Names - A Journal of Onomastics*. https://doi.org/10.5195/NAMES.2021.2234

Hämäläinen, L., & Ruokolainen, T. (2021). Kukkaa, amfea, subua ja essoja: Huumausaineiden slanginimitykset Tor-verkon suomalaisella kauppapaikalla. *Sananjalka*, *63*, 130–153. https://doi.org/10.30673/sja.106615

Harviainen, J. T., Haasio, A., & Hämäläinen, L. (2020). Drug traders on a local dark web marketplace. *ACM International Conference Proceeding Series*, 20–26. https://doi.org/10.1145/3377290.3377293

Harviainen, J. T., Haasio, A., Ruokolainen, T., Hassan, L., Siuda, P., & Hamari, J. (2021). Information protection in dark web drug markets research. *Hawaii International Conference on System Sciences*.

Jin, Y., Jang, E., Lee, Y., Shin, S., & Chung, J. W. (2022). *Shedding new light on the language of the dark web* [*arXiv preprint (To appear in NAACL 2022)*].

Karjalainen, K., Nyrhinen, R., Gunnar, T., Ylöstalo, T., & Ståhl, T. (2021). Huumeiden saatavuus, käyttö ja huumausainerikollisuus Tampereella koronakeväänä 2020. *Yhteiskuntapolitiikka*, *86*(2), 80–90.

Laippala, V., & Ginter, F. (2014). Syntactic n-gram collection from a large-scale corpus of internet finnish. *Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT*.

Leedham, M., Lillis, T., & Twiner, A. (2021). Creating a corpus of sensitive and hard-to-access texts: Methodological challenges and ethical concerns in the building of the WiSP Corpus. *Applied Corpus Linguistics*, *1*(3). https://doi.org/10.1016/j.acorp.2021.100011

Li, T., & Li, N. (2009). On the tradeoff between privacy and utility in data publishing. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 517–526.

Lindén, K., Ruokolainen, T., Hämäläinen, L., & Harviainen, J. (2023a). Ethically archiving a hard-to-access massive research data set in the language bank of finland: The finnish dark web marketplace corpus (findarc) [Publisher Copyright: © 2023 Copyright for this paper by its authors.; Conference on Technology Ethics, Tethics ; Conference date: 18-10-2023 Through 19-10-2023]. In M. Rantanen, S. Westerstrand, O. Sahlgren, & J. Koskinen (Eds.), *Proceedings of the conference on technology ethics 2023 - tethics 2023* (pp. 114–131). CEUR-WS.org.

Lindén, K., Ruokolainen, T., Hämäläinen, L., & Harviainen, J. (2023b, December). Sharing the finnish dark web marketplace corpus (findarc) [CLARIN Annual Conference ; Conference date: 16-10-2023 Through 18-10-2023]. In K. Lindén, J. Niemi, & T. Kontino (Eds.), *Clarin annual conference proceedings* (pp. 134–139). CLARIN ERIC. https://www.clarin.eu/event/2023/clarin-annual-conference-2023

Nabki, A., M. W., E. F., Alegre, E., & Paz, I. D. (2017). Classifying illegal activities on tor network based on web textual contents. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 35–43.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*, 1701.

Oksanen, A., Tamper, M., Tuominen, J., Hietanen, A., & Hyvönen, E. (2019). AnoPpi: A pseudonymization service for Finnish court documents. In *Jurix 2019* (pp. 251–254). IOS Press.

Rubinstein, I. S., & Hartzog, W. (2016). Anonymization and risk. *Wash. L. Rev*, *91*, 703.

Stalla-Bourdillon, S., & Knight, A. (2016). Anonymous data v. personal data-false debate: an EU perspective on anonymization, pseudonymization and personal data. *Wis. Int'l LJ*, *34*, 284.

Tamper, M., Oksanen, A., Tuominen, J., Hyvönen, E., & et. al., A. H. (n.d.). Anonymization Service for Finnish Case Law: Opening Data without Sacrificing Data Protection and Privacy of Citizens. *International Conference on Law via the Internet, LVI*.

Ylilauta. (2016). *Ylilauta corpus downloadable version [text corpus]* (tech. rep.). Kielipankki. http://urn.fi/urn:nbn:fi:lb-2016101210

Constructing SABeD: A Spoken Academic Belgian Dutch Corpus

Jolien Mathysen
KU Leuven, Belgium
jolien.mathysen@kuleuven.be

Vincent Vandeghinste
Instituut voor de Nederlandse Taal, The Netherlands
vincent.vandeghinste@ivdnt.org
KU Leuven, Belgium

Elke Peters
KU Leuven, Belgium
elke.peters@kuleuven.be

Patrick Wambacq
KU Leuven, Belgium
patrick.wambacq@kuleuven.be

Abstract

We present the Spoken Academic Belgian Dutch (SABeD) corpus and a description of its construction. It was compiled from selected first bachelor academic lectures in higher education institutions in Flanders, as students indicate that the language used in such lectures is one of the hurdles for comprehension and academic success. We first applied speech recognition on these lectures and then applied manual utterance segmentation and manual correction of the automated transcription. A filtered version of the resulting transcriptions was automatically punctuated and linguistically annotated with CLARIN tools and is currently available for search in the Autosearch online corpus query environment. The manual transcriptions and the ELAN files with the final annotation will soon be made available to the research community for download in the CLARIN infrastructure at http://hdl.handle.net/10032/tm-a2-w4.

1 Introduction

In higher education, students are confronted with academic language use, with which they are often not familiar. Since academic language skills are a necessary condition for study success, higher education institutions in Flanders and the Netherlands focus on language support for students. In many institutions, these efforts evolved into formal, embedded language policies, but research into their implementation is limited (Bonne & Casteleyn, 2022). The number of international students pursuing higher education in Flanders is estimated at around 2,500 per year (Deygers & Malone, 2019). Research (Deygers, 2017; Deygers et al., 2017) shows that Dutch language learners struggle with academic spoken Dutch, even when they passed the university entrance language tests, ITNA (*Interuniversitaire Taaltest Nederlands voor Anderstaligen*; Interuniversity Test of Dutch for Speakers of Other Languages)[1] or CNaVT (*Certificaat Nederlands als Vreemde Taal*; Certificate of Dutch as a Foreign Language)[2]. These tests are required for international students wishing to follow programmes taught in Dutch in higher education in Flanders. Although academic listening can be part of such tests, learners have indicated that the listening tasks used in these tests are easier than actual lectures (Deygers et al., 2018). One reason for this discrepancy is that the linguistic features of these listening tasks have not been empirically validated because of the lack of a corpus of spoken academic Dutch.

One of the key predictors of success in university entrance language tests (Heeren et al., 2020; Trenkic & Warmington, 2019) is learners' vocabulary knowledge. At the same time, academic vocabulary

[1] https://www.itna.be/
[2] https://cnavt.org/

Jolien Mathysen, Vincent Vandeghinste, Elke Peters and Patrick Wambacq 2023. Constructing SABeD: A Spoken Academic Belgian Dutch Corpus. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 153–163. DOI: https://doi.org/10.3384/ecp210

has also been put forward as one of the main challenges that university students face when listening to academic lectures. Research into English has shown that listening comprehension increases with lexical coverage, or the percentage of known words in a text (Durbahn et al., 2020; Van Zeeland & Schmitt, 2013), and that vocabulary knowledge is a strong predictor of academic listening comprehension (Wallace, 2020). More concretely, learners should be familiar with 95% to 98% of the running words in spoken academic English to reach reasonable to very high comprehension. This means that to reach detailed comprehension of lectures, learners need to know high frequency vocabulary, as well as the 3,000-9,000 most frequent word families (plus proper nouns and marginal words) in English respectively (Dang, 2022; Dang & Webb, 2014; Stæhr, 2009).

This was one of the main reasons for starting the SABeD (Spoken Academic Belgian Dutch) project. The principal aim of this project was building a multimodal corpus of spoken academic Belgian Dutch, which consists of (transcripts and audio/video recordings of) academic lectures[3]. Besides, lectures can be said to constitute the dominant form of instruction for students entering higher education in Flanders. This means that they must process new subject matter in a register they may not yet fully master. This is particularly problematic, considering that research has shown a link between academic language skills and success in higher education (Heeren et al., 2020, 2021; Milton & Treffers-Daller, 2013; Trenkic & Warmington, 2019). The compilation of a Spoken Academic Belgian Dutch corpus is, thus, warranted. Especially since earlier research (Dang et al., 2017) has already demonstrated the usefulness of such corpora in analysing the lexical demands of academic lectures and in developing/validating vocabulary learning materials that can help students reach the required lexical coverage levels for detailed comprehension of those lectures. Moreover, such corpus-based learning materials may also raise awareness among both teachers and students about their academic language use and struggles (Dang, 2022; Dang et al., 2021; Nation, 2016; Uchihara & Webb, 2022). Due to the covid pandemic, recorded video lectures are available in abundance. The SABeD corpus contains a mix of written-to-be-read and spontaneous spoken language.[4] This paper presents how the corpus was created.

Section 2 describes related work, Section 3 describes how the corpus was compiled, and Section 4 describes the semi-automatic transcription process, consisting of a phase of automated speech recognition (Section 4.1) and a phase of manual post-editing (Section 4.2). Section 5 describes how the manual transcriptions were filtered and postprocessed using CLARIN tools and how the corpus is made available to specific researchers, again using CLARIN tools. Section 6 draws conclusions and sketches the plans of using the corpus for the creation of an academic vocabulary list and empirical validation of a new university entrance vocabulary test.

2 Related Work

Some of the most notable corpora featuring spoken academic language currently in existence include the MICASE (*Michigan Corpus of Academic Spoken English*) corpus (Simpson et al., 2002), T2K-SWAL (*TOEFL 2000 Spoken and Written Academic Language*) corpus (Biber et al., 2002), BASE (*British Academic Spoken English*) corpus (Thompson & Nesi, 2001) and the CGN (*Spoken Dutch Corpus*) (Oostdijk et al., 2002). We will briefly discuss these four corpora because of their prominence within the field, and the standards they provided for designing the SABeD corpus in terms of balance and representativeness decisions.

[3]The other aims of the project are investigating the effectiveness of speech technology for automatic transcription of spoken texts, developing a word frequency list of spoken academic Belgian Dutch, and creating a vocabulary test of spoken academic Belgian Dutch, but these fall outside of the main focus of this paper.

[4]In analogy with CGN (Oostdijk et al., 2002), we consider our corpus a *spoken corpus*. The initial collection of 972 recordings did include a number of pre-recorded lectures and knowledge clips in which the lecturers had prepared and/or read out their text. However, in the selection of the 200 recordings for the final corpus, live lectures taught on campus were given priority. Additionally, other spoken corpora, such as the CGN also feature a variety of (semi-)structured instances of speech (e.g. interviews, news bulletins, masses, formal speeches, and even recited texts) (cf. https://ivdnt.org/images/stories/producten/documentatie/cgn_website/doc_Dutch/topics/overview.htm#inleiding). Apart from that, other spoken academic corpora, among other the MICASE corpus (Simpson et al., 2002), also conceptualise academic speech in a broad sense as any speech occurring in an academic setting. Consequently, they also tend to contain (semi-)structured instances of speech, such as presentations or dissertation defenses.

The MICASE corpus comprises 152 different academic speech events (i.e. small and large lectures, discussion and lab sections, seminars, student presentations, advising sessions, colloquia, dissertation defenses, interviews, meetings, office hours, service encounters, study groups, tours and tutorials) from the University of Michigan, amounting to nearly 200 hours or approximately 1.7 million words worth of transcriptions. It spans four major academic divisions (i.e. Humanities and Arts, Social Sciences and Education, Physical Sciences and Engineering, and Biological and Health Sciences) and also features discourse and text-linguistic annotation (Simpson et al., 2002). The spoken component of the T2K-SWAL corpus includes 1.7 million words recorded at four different American universities. The largest part of this (1.2 million words) was taken from 176 class sessions, while the remaining 50,000 came from office hours (Biber et al., 2002). BASE consists of 160 lectures and 40 seminars recorded at the University of Warwick and the University of Reading between 2000 and 2005. This 1,186,290 token corpus was compiled from four disciplinary sub-corpora: Arts and Humanities, Life and Medical Sciences, Physical Sciences, and Social Sciences. Except for Physical Sciences, each sub-corpus contains 40 lectures and 10 seminars (Thompson & Nesi, 2001). CGN is a balanced corpus with several variants of spoken Dutch (from read-aloud text to spontaneous conversations, from Belgium and the Netherlands), and which contains 30,917 words from university lectures (Oostdijk et al., 2002). Apart from these four corpora, clearly, other less-known and/or smaller spoken academic corpora have been created for English, as well as for other (non-Indo-European) languages. However, a domain-specific spoken corpus like the one we introduce in this paper was until recently not available for (Belgian) Dutch.

3 Corpus Compilation

In the corpus compilation stage, we selected academic lectures from both KU Leuven and Ghent University, because these constitute the predominant form of instruction in higher education institutions in Flanders, especially in the first bachelor year. Our corpus does not include university college lectures, because it was not possible to collect as many of them as of the university lectures. In addition, the university college lectures that we did manage to collect were predominantly associated with the social sciences. As such, including these recordings in the corpus without obtaining any more university college lectures first would have resulted in an unbalanced corpus.

For the purpose of our corpus, lectures are defined as instructional discourse presented before an audience of at least 40 students, in which the lecturer is the dominant speaker and the level of interactivity is modest to low. This is in line with the distinction made by the MICASE (Simpson et al., 2002) corpus between small and large lectures, for which the cut-off point was also 40 students. Furthermore, there are several reasons validating a corpus specifically oriented toward this kind of prototypical lecture. First, one should consider if and to what degree the language of lectures is influenced by the size of their audiences (e.g. frequency and nature of the interactions between lecturer and students). In this regard, it should also be taken into account that both native speakers and foreign learners of Dutch specifically indicated the language used in Flemish lectures as one of the hurdles for comprehension and academic success (Deygers, 2017; Deygers et al., 2017). In addition, research for English has also found that lectures are more (lexically) demanding as opposed to e.g. labs and tutorials (Dang et al., 2021).

Academic Division	No. of Videos	No. of Speakers	Full Video length	Transcribed length	Tokens (raw)	Tokens (clean)
Biological and Health Sciences	50	5	56:20:04	25 hrs	201,122	190,943
Physical Sciences and Engineering	50	11	75:16:55	25 hrs	207,398	200,883
Humanities and Arts	50	9	61:54:18	25 hrs	227,360	215,374
Social Sciences and Education	50	6	99:55:52	25 hrs	267,279	254,487
Total	200	31	293:24:09	100 hrs	903,159	861,787

Table 1: Corpus size over the different academic divisions. Video length is shown in `HH:MM:SS` (Hours:Minutes:Seconds format) and represents the lengths of entire lectures.

We specifically chose first year bachelor lectures, because they also constitute the first encounter of the target group (i.e. Flemish first year bachelor students and international students commencing university education in Belgian-Dutch) of our corpus with spoken academic Dutch. In this regard, it is particularly

important to note the primary pedagogical goal of the SABeD project and corpus, i.e. developing learning materials for students entering Flemish higher education. As such, these lectures make up a solid base for our corpus compilation, especially considering that we cannot be certain if and to what degree the language of lectures in later bachelor and master years differs from that in the first bachelor year.

To ensure that the corpus is both representative and has sufficient power to make statistical inferences, lectures from a considerable number of lecturers needed to be included (Biber, 1993). However, the selection of lectures for the transcription stage was impeded by the fact that, due to technical reasons, it was not possible to automatically download lectures from the video platforms used by Flemish universities. All lecturers/professors had to be contacted individually and all lectures had to be downloaded and processed one-by-one before they could be added to the corpus. Informed consent was obtained and metadata was collected (e.g. speaker data such as age, gender, teaching experience, place of birth). Eventually, we obtained 972 recordings, out of which we selected 200 lectures to undergo manual post-editing. This is still more lectures than in other existing corpora of spoken academic English such as the MICASE corpus (62 lectures; Simpson et al., 2002), T2K-SWAL corpus (176 lectures; Biber et al., 2002) or BASE (160 lectures; Thompson and Nesi, 2001). In terms of tokens (cf. Table 1), our corpus is smaller than the corpora mentioned in Section 2, but this is mainly due to the fact that we chose to only manually transcribe 30 minutes per lecture instead of transcribing a considerably smaller number of entire lectures. Only the 30-minute parts of lectures that had undergone manual post-editing were included in our base corpus to ensure a well-balanced corpus. However, the parts of the lectures that were transcribed strictly automatically will also be included at a later stage. Additionally, the remaining 772 videos that were not selected to be part of the initial base corpus will also be processed and added later, thus ensuring that the corpus will still grow significantly in size.

	Biological and Health Sciences	**Physical Sciences and Engineering**	**Humanities and Arts**	**Social Sciences and Education**
Anatomy	23	0	0	0
Archaeology, Art, History, Philosophy	0	0	10	0
Biochemistry	14	0	0	0
Chemistry	0	17	0	0
Dutch linguistics	0	0	18	0
Economics and Law	0	0	0	16
Electronics and Programming	0	5	0	0
General Linguistics	0	0	22	0
Genetics	13	0	0	0
Maths	0	13	0	0
Physics	0	15	0	0
Psychology	0	0	0	18
Research in Social Sciences	0	0	0	16
Total	50	50	50	50

Table 2: No. of lectures per discipline and academic division

Due to us only manually transcribing 30 minutes per lecture, the 200 selected lectures had to be at least 30 minutes in length. However, this criterium was also, to a lesser extent, dictated by the existing data and circumstances. First, the prevalence and use of shorter lecture recordings (between 30 minutes and 1 hour) seems to have increased since the Covid-19 pandemic and thus in the data that we collected as well. Secondly, including these shorter lectures also led to greater variation in terms of speakers in the corpus. Following the example of the MICASE (Simpson et al., 2002) and BASE (Thompson & Nesi, 2001) corpora, the selection of lectures was further informed by academic division (Biological and Health Sciences, Humanities and Arts, Physical Sciences and Engineering, Social Sciences and Education). Table 1 presents the size of each of the academic divisions in the corpus. Generally, the academic division was attributed to a lecture by checking the study programme in which the lecture had been taught. In ambiguous cases, for instance, a science course being taught as part of the archaeology programme, the research unit and department of the lecturer determined the allocation of the academic division. Within each academic division, a broad range of disciplines is covered. We aimed to compose a corpus that is sufficiently representative for the purpose of composing a word list of spoken academic Dutch. At

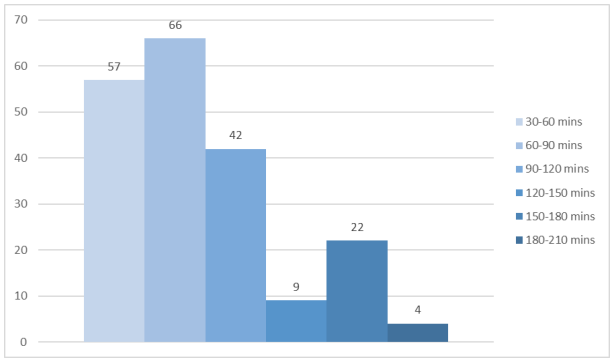


Figure 1: No. of lectures in the corpus according to their length in minutes

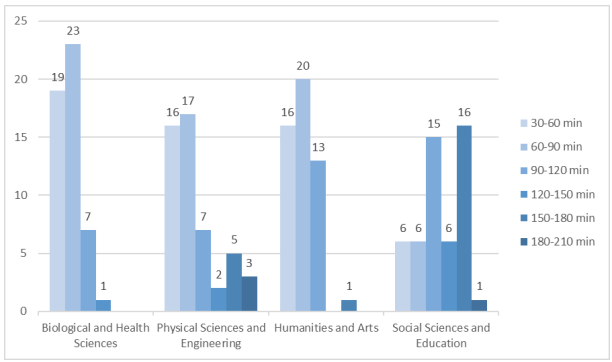


Figure 2: No. of lectures in the corpus according to their length in minutes per academic division

the same time, the corpus should contain sufficient data to obtain standards for more specific academic divisions. Consequently, each of the aforementioned academic divisions was equally represented, thus amounting to 50 lectures per division. A distribution of the lengths of the videos is shown in Figure 1. A more detailed overview of the range of disciplines represented by the videos in the corpus per scientific division is given in Table 2, and a distribution of the length of the videos per academic division is shown in Figure 2. Table 2 was obtained by grouping together the videos of courses in the corpus that covered identical, similar or related subjects within each scientific division. This was done to be able to provide additional insight into the disciplines that are incorporated, while also considering the privacy of lecturers and GDPR (General Data Protection Regulation) by not sharing the actual course names. Finally, we also collected about 800,000 tokens of written course materials, to increase the accuracy and performance of the speech recognition system (cf. section 4.1).

4 Transcription

As mentioned before, we only manually transcribed 30 minutes per lecture, instead of transcribing a considerably smaller number of entire lectures. More specifically, we transcribed only the first 25 and last 5 minutes of the lectures. The main benefit of this practice is that any differences between institutes and disciplines in length of lectures are eliminated, ensuring a well-balanced corpus. Moreover, transcribing the introductory parts and conclusions of the lectures ensured that the corpus is sufficiently representative of the more general spoken academic language spanning across different divisions. In these parts, one could expect a higher occurrence of general academic language and vocabulary than in the middle parts which could contain relatively more subject-specific technical language and vocabulary. However, by also transcribing portions of the lectures that stretch beyond their introductions, the corpus also still captures data on more specific academic divisions and disciplines. This distinction between general aca-

demic and technical language/vocabulary is based on the definition of these terms by Nation, 2016.

4.1 Automated Speech Recognition

Speech recognition was performed with an ASR system tuned for Belgian Dutch (Van Dyck et al., 2021), which is Kaldi-based (Povey et al., 2011). The Kaldi toolkit makes use of state-of-the-art deep neural networks. The new acoustic model was trained on data from the Spoken Dutch Corpus (Oostdijk et al., 2002) and tested using the N-best benchmark (Kessens & van Leeuwen, 2007). The output of the ASR system consists of `ctm` files,[5] which contain time stamps for each recognised word and a confidence level for each word. An alternative for this system would be to use the recently emerging end-to-end systems, but these do not allow independent training of language models, which is one of the additional goals of the SABeD project. The additional text material from textbooks and course materials will be used to improve the lexicon and language model independently from the acoustic model. This latter effort is still ongoing. The *raw ASR* output was inserted into an ELAN file as a separate tier. ELAN (Wittenburg et al., 2006) is well-known software for linguistic annotation of audio-visual material.

4.2 Manual Post-Editing

Manual post-editing was performed in two stages. Section 4.2.1 describes utterance segmentation and Section 4.2.2 describes manual correction of the speech recognition results.

4.2.1 Utterance Segmentation

The first step in manual post-editing consisted of applying utterance segmentation. This was done to make post-editing easier and faster. If we would not have applied segmentation, ASR correction would necessarily have needed to be performed at the word level. This would have been cumbersome as ASR errors can occur across word boundaries, requiring annotators to not only correct the transcript, but also to manually manipulate the time stamps. This segmentation annotation was registered in a separate ELAN tier. It entailed the placement of boundaries that indicate the start and end of a piece of transcription in the audio signal. The unit that demarcates these boundaries is the chunk, which is defined as a speech fragment that lasts about 2-3 seconds and is delimited on both sides by a (short) audible and visible pause. Chunks can, but do not need to, correspond to sentences or phrases.

Shorter pauses (nearly always shorter than 1 second) that occurred within a chunk were ignored. As such, chunks that lasted less than 2-3 seconds were only possible if we came across a speech fragment of less than 2-3 seconds that was, on both sides, clearly separated from any other fragments by a long pause (i.e. of more than 3 seconds). Chunks longer than 3 seconds were allowed (e.g. multiple co- or subordinated clauses, long enumerations). However, if chunks lasted longer than 6 seconds, they were split up in front of a conjunction or where a comma would appear in written language. All in all, the length of chunks was highly dependent on the pace of the speaker under scrutiny, with slower speakers generally yielding fewer and longer chunks and faster speakers usually requiring more and shorter ones. An absolute criterion for the segmentation consisted of keeping compounds within the same chunk.

Any student interactions or substantial background noises (e.g. an opening door) were also isolated in segments and tagged in a separate ELAN tier. This way students' voices could be easily removed later to respect their privacy and GDPR, as we could not identify them nor obtain their informed consent. Noises could be taken into consideration when these influenced the flow of speech. If some noise persisted in the background throughout all or most of the recording, no segment was created and this was instead signalled by adding a comment to the ELAN file. The *raw ASR* tier was then combined with the *manual segmentation* tier into a *segmented ASR* tier[6].

4.2.2 Correcting the Automated Transcriptions

The second step in manual post-editing consisted of correcting the automated transcription at the segment level. Annotators manually corrected the *segmented ASR* tier using a transcription protocol that was based

[5] `ctm` stands for time-marked conversation file.

[6] Privacy and GDPR inhibit us from sharing the contents of both the *raw ASR* tier and the *segmented ASR* tier, since the automatic transcriptions may contain personal data (e.g. names of lecturers, courses, student names) or speech of students from whom we did not get informed consent. All this information was removed in a later stage of the transcription process.

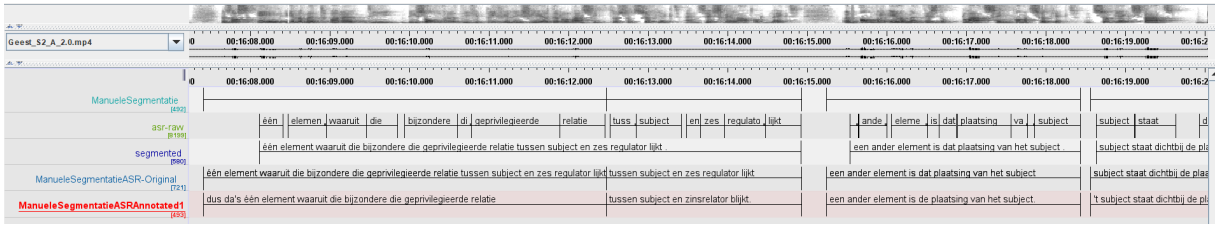


Figure 3: ELAN screenshot showing the different tiers for manual post-editing

on that of the CGN (Oostdijk et al., 2002) and put the correction into the *manual transcription* tier. More specifically, there were two phases involved in the manual correction of the transcriptions. During the first phase, transcriptions were corrected at the orthographic level. This entailed that the spelling was standardised and punctuation was added. Interjections that did not constitute words themselves (e.g. *ah, hè, uhm*) and words from languages other than Dutch (e.g. Latin medical terms or English book titles, but not loanwords that have been accepted into Dutch) were annotated using designated codes. Students' speech was cut out because it was too difficult to track them down and get their consent. Personal data (e.g. names of lecturers, students, courses) were also anonymised to uphold privacy and GDPR. During the second phase, transcriptions were checked at the acoustic level. This included the annotation of reductions (e.g. *'k moet hier weg*; I got to get out of here or *da's niet waar*; that's not true), dialect words (e.g. *talloor*; dialect term for plate), slips of the tongue, aborted words and sentences, unintelligible pieces of speech, and noises made by the speakers (e.g. coughing or sneezing). The number of files processed and the token count (using the linux `wc` command) on the extracted *manual transcription* tier are presented in Table 1.

5 Processing with CLARIN Tools

The ELAN files containing the manual transcriptions and other annotation layers are available for download at https://hdl.handle.net/10032/tm-a2-w4. When extracting the transcription for further processing and compilation of frequency lists, we removed the special codes that were annotated in the manual transcription, such as tokens indicating hesitation/non-understandable audio pieces and special indicators marking dialect/foreign words. The resulting token counts after such cleaning are presented in Table 1. Note that these codes were included in the manual corpus annotation as such annotations may be relevant for other types of research related to spoken language or as training data for speech recognition models.

The corpus was then processed with the Full Stop Punctuation Inserter for Dutch (Vandeghinste & Guhr, 2023), which was built specifically for projects like SABeD. The resulting corpus was then processed using the CLARIN tool Frog (van den Bosch et al., 2007),[7], an NLP processing toolkit for Dutch. This resulted in the following analyses: tokenization, part-of-speech tagging, lemmatization, morphological segmentation, dependency parsing, and named entity labeling. Frog outputs FoLiA format (van Gompel & Reynaert, 2013), an XML format made for linguistic annotations and CoNLL type tab separated files, containing the same information.

[7]https://webservices.cls.ru.nl/frog

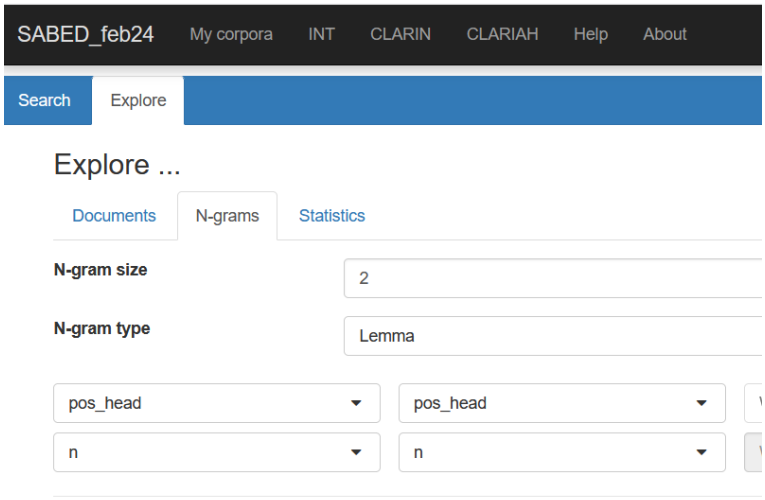


Figure 4: Query for generating bigrams of lemmas of nouns in Autosearch

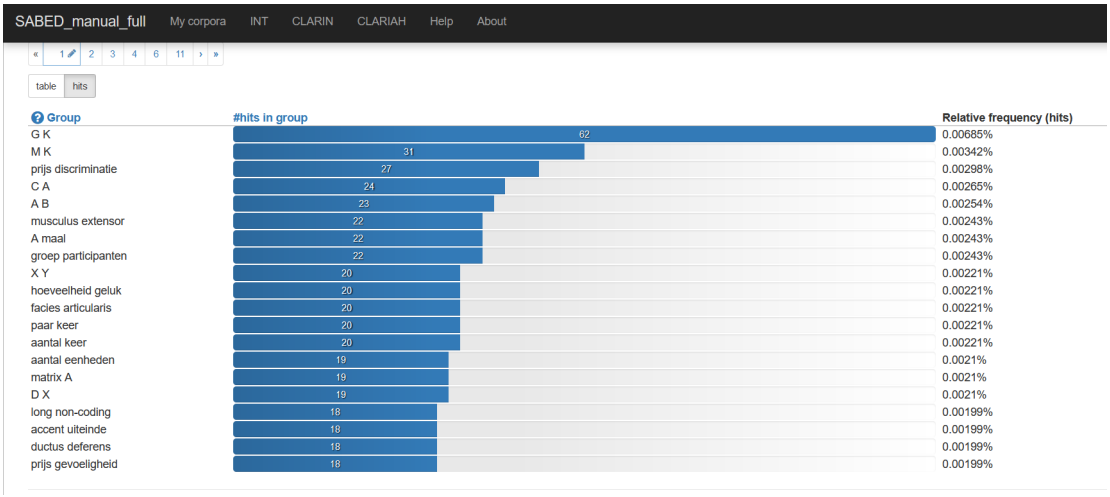


Figure 5: First page of results from querying lemma bigrams for nouns.

These FoLiA data have been uploaded to the CLARIN Autosearch engine[8] for annotated corpora, making the corpus now searchable with Corpus Query Language (CQL) and sharable with other CLARIN users. Figure 4 shows the query to generate lemma bigrams of nouns in the corpus, and Figure 5 shows the results page. Results are also downloadable as `csv` files. While it is clear that the results are not always informative, the Autosearch interface to the corpus provides easy ways of exploring the corpus, and the queries can easily be made more refined. Autosearch can also be used as a concordancer.

6 Conclusions and Future Work

We have presented the corpus compilation efforts for a corpus of spoken academic Belgian Dutch. Because we prioritised manual post-editing of parts of a limited selection of lectures in favour of balancing, our corpus is currently limited in token size (cf. Table 1) compared to other similar corpora. However, the strictly automatically transcribed parts of this selection of lectures, as well as 772 videos that were not part of the base corpus, will still be added at a later stage. This will significantly expand the size of the

[8] Autosearch is a version of Blacklab (de Does et al., 2017), which works with customizable corpora. http://hdl.handle.net/10032/tm-a2-c5

corpus in the future. Once all data is processed, combining the metadata with the linguistic annotations in TEI format will allow even more fine-grained querying of the corpus, not only on linguistic criteria but also on metadata criteria. The corpus will also be made available in a Blacklab (de Does et al., 2017) corpus query engine for all CLARIN users. Even so, we acknowledge that variation in terms of speakers and disciplines was one of the main challenges and concerns in the construction of this corpus. Additionally, we would like to point out that the matter of anonymising multimedia materials (with regard to personal data such as names of lectures, course names and students' speech), specifically audio and video, should not be taken lightly. Nevertheless, we aim to make the video and audio recordings of the lectures available to the research community as well once this process has been completed.

Concerning automatic speech recognition we can assert that while ASR speeds up the manual transcription, it is clear that a general domain ASR system does not contain a specialised vocabulary like the one that is being used in academic lectures, and therefore tuning the vocabulary and language model of the ASR system towards the specific domains is expected to greatly improve ASR accuracy and reduce post-editing effort, which should result in a speedier post-editing process.

Even though validation of a new word list and vocabulary test for spoken academic Belgian Dutch was one of the main reasons for collecting the corpus, this is still future work. The academic spoken word list will be developed based on the frequency and range of the words in the corpus (Dang et al., 2017; Szudarski, 2017) with the lemma as counting unit. This functionality is included in the Blacklab environment. To determine which words can be considered academic words, the frequency list will be compared to the general Dutch word list of Tiberius and Schoonheim, 2013. Words not occurring in that list are potential candidates for the spoken academic vocabulary list (depending on their frequency and distribution in the corpus). We will distinguish proper names, general academic words, and domain-specific words. After the example of the English Academic Spoken Word list (Dang et al., 2017), the Belgian Dutch list will be divided into sublists of 50 words, based on frequency. Additionally, we will process the corpus with term extraction tools such as D-terminer (Rigouts Terryn et al., 2022), TermTreffer[9] or their successors.

We will also develop a frequency-based spoken academic vocabulary test targeting students' aural recognition of academic word forms and their meanings. The test will be divided into test sections, corresponding to the sublists of the frequency list. It will have an online multiple-choice format. Students will be provided with the spoken form of the word and have to tick off the correct option that corresponds to the word's meaning. The first test version will be piloted with a small group of Dutch-speaking students (n=30) before the start of the actual larger-scale validation process.

For the moment, exploring and implementing further practical applications, as well as evaluation of the corpus, are still areas of future research. Of course, once the corpus has been made available to researchers, a multitude of new uses and applications can be envisaged, such as comparisons at lexical, syntactic, and other levels with other (spoken and/or written) Dutch corpora.

Acknowledgments

The SABeD project is funded by KU Leuven Internal Funding, Research Project 3H200610, with additional support from the Instituut voor de Nederlandse Taal.

Use of CLARIN-NL tools on the SABeD data was work done in ClaSABeD (Clariah-NL tools in SABeD), funded by CLARIAH-PLUS Work package 3, project no. CP-WP3-22-001.

We would like to explicitly thank the transcribers of the corpus for their dedicated and tedious work: Israe Ahale, Ellen Boucké, Mirne De Beule, Gijs Heynickx, Emma Janssens, Yitse Murillo, Kim van den Berg, and Joke Van Mol.

References

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*, 243–257.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly*, *36*, 9–48.

[9] https://termtreffer.org/

Bonne, P., & Casteleyn, J. (2022). Taalbeleid en taalondersteuning: Op zoek naar een gedeelde basis en strategie voor implementatie. *Tijdschrift voor Onderwijsrecht en Onderwijsbeleid*, *4*, 279–293.

Dang, T. N. Y. (2022). Vocabulary in academic lectures. *Journal of English for Academic Purposes*, *58*, 101–123.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning*, *67(4)*, 959–997.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2021). Vocabulary in academic spoken English. *New Zealand Studies in Applied Linguistics*, *26(2)*.

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66–76.

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating Research Environments with BlackLab. In *CLARIN in the Low Countries*. Ubiquity Press.

Deygers, B. (2017). Validating university entrance policy assumptions. Some inconvenient facts. In E. Gutiérrez Eugenio (Ed.), *Learning and Assessment: Making the Connections – Proceedings of the ALTE 6th International Conference* (pp. 46–50). Cambridge: ALTE.

Deygers, B., & Malone, M. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*, *36(3)*, 347–368.

Deygers, B., Van den Branden, K., & Peters, E. (2017). Checking assumed proficiency: comparing L1 and L2 performance on a university entrance test. *Assessing Writing*, *32*, 43–56.

Deygers, B., Van den Branden, K., & Van Gorp, K. (2018). University entrance language tests: A matter of justice. *Language Testing*, *35*, 449–476.

Durbahn, M., Rodgers, M., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System*, *88*.

Heeren, J., Speelman, D., & De Wachter, L. (2020). A practical academic reading and vocabulary screening test as a predictor of achievement in first-year university students: Implications for test purpose and use. *International Journal of Bilingual Education and Bilingualism*, *0(0)*, 1–16.

Heeren, J., Speelman, D., & De Wachter, L. (2021). Bepaalt taal wie het haalt? de samenhang tussen een academische taalvaardigheidscreening en het behalen van een bachelordiploma aan de universiteit. *Tijdschrift voor Hoger Onderwijs*, *39(1)*, 39–54.

Kessens, J. M., & van Leeuwen, D. A. (2007). N-best: the northern- and southern-Dutch benchmark evaluation of speech recognition technology. *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, 1354–1357. http://www.isca-speech.org/archive/interspeech%5C_2007/i07%5C_1354.html

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: The link between vocabulary size and academic achievement. *Applied Linguistic Review*, *4(1)*, 151–172.

Nation, I. S. P. (2016). Making and using word lists for language learning and testing.

Oostdijk, N., Goedertier, W., van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., & Baayen, H. (2002, May). Experiences from the spoken Dutch corpus project. In M. González Rodríguez & C. P. Suarez Araujo (Eds.), *Proceedings of the third international conference on language resources and evaluation (LREC'02)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2002/pdf/98.pdf

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit [IEEE Catalog No.: CFP11SRW-USB]. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2022). D-terminer : online demo for monolingual and bilingual automatic term extraction. In R. Costa, S. Carvalho, A. Ostroski Anic, & A. F. Khan (Eds.), *Proceedings of the Workshop on Terminology in the 21st century : many faces, many places* (pp. 33–40). European Language Resources Association (ELRA). %7Bhttps://lt3.ugent.be/dterminer/%7D

Simpson, R. C., Briggs, S. L., J., O., & Swales, J. M. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31(4)*, 577–607.

Szudarski, P. (2017). *Corpus linguistics for vocabulary. A guide for research*. Routledge.

Thompson, P., & Nesi, H. (2001). The British Academic Spoken English (BASE) Corpus Project. *Language Teaching Research*, *5*, 263–264.

Tiberius, C., & Schoonheim, T. (2013). *A frequency dictionary of Dutch: Core vocabulary for learners*. Routledge.

Trenkic, D., & Warmington, M. (2019). Language and literacy skills of home and international university students: How different are they, and does it matter? *Bilingualism: Language and Cognition*, *22*, 349–365.

Uchihara, T., & Webb, S. (2022). Materials for teaching vocabulary. In *The routledge handbook of materials development for language teaching* (pp. 202–217). Taylor; Francis.

van den Bosch, A., Busser, G., Daelemans, W., & Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. van Eynde, P. Dirix, I. Schuurman, & V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting* (pp. 99–114). Centre for Computational Linguistics.

Van Dyck, B., BabaAli, B., & Van Compernolle, D. (2021). A Hybrid ASR System for Southern Dutch. *Computational Linguistics in the Netherlands Journal*, *11*, 27–34. https://clinjournal.org/clinj/article/view/119

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied linguistics*, *34(4)*, 457–479.

Vandeghinste, V., & Guhr, O. (2023). FullStop: Punctuation and Segmentation Prediction for Dutch with Transformers. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-023-09676-x

van Gompel, M., & Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation – a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, *3*, 63–81. https://clinjournal.org/clinj/article/view/26

Wallace, M. P. (2020). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*, *72(1)*, 5–44.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 1556–1559.

Adding political orientation metadata to ParlaMint corpora

Katja Meden
Dept. of Knowledge Technologies,
Jožef Stefan International Postgraduate School,
Jožef Stefan Institute, Slovenia
katja.meden@ijs.si

Jure Skubic
Institute of Contemporary History,
Ljubljana, Slovenia
jure.skubic@inz.si

Tomaž Erjavec
Department of Knowledge Technologies,
Jožef Stefan Institute, Slovenia
tomaz.erjavec@ijs.si

Abstract

Parliamentary debates are an important source for political discourse research as well as research in other disciplines. The ParlaMint project aims to create comparable corpora of parliamentary debates which, through unified encoding, provide a comprehensible resource to support such research. Within these corpora, speeches are attributed to speakers, and speaker metadata, including temporal affiliations with different organizations such as parliamentary groups and political parties. This paper discusses the addition of metadata on the political orientation of parties and parliamentary groups to the ParlaMint corpora. The paper explains our two sources for this information, namely the Chapel Hill Expert Survey Dataset and Wikipedia, the process of data collection and its subsequent encoding in the corpora. Furthermore, the paper presents an analysis of the extent of the added metadata, along with an example of exploratory data analysis. It also outlines the distribution of utterances across political orientation categories within ParlaMint, offering a comprehensive overview of the diverse perspectives and ideologies within the corpora. The inclusion of this supplementary metadata could prove valuable for parliamentary data research, while the methodology developed could be used to add further metadata to the ParlaMint corpora.

1 Introduction

Parliaments are of interest to the humanities and social sciences as they shape legislation that affect people's daily lives and are a source of power for MPs and other politicians (Bischof & Ilie, 2018). Parliamentary speeches and parliamentary data are of great importance for analysis at the (inter)national level and are an interesting topic for various research projects. In addition to transcripts of parliamentary debates, metadata (such as age, gender, party affiliation, political orientation, political role, etc.) are crucial for the study of parliamentary discourse, as they provide useful additional information that can be used in parliamentary discourse research and provide even more relevant and reliable research results.

The ParlaMint[1] projects, funded by CLARIN, aimed to create comparable and uniformly encoded corpora of speeches in European parliaments and make them openly accessible. In ParlaMint I (2020-2021), corpora for 17 European parliaments were created, made available, and used in research and education (Erjavec et al., 2022). The project continued as ParlaMint II (2022-2023), providing 12 new corpora, adding newer transcripts, improving the annotation schema and validation, machine-translating the corpora into English, and expanding the corpus metadata.

The additional metadata added to the corpora consists of information on whether and when a speaker is or was a minister and the political orientation of the parliamentary group or political party to which the speaker belongs. Both of these additions have been suggested by researchers (cf. Fišer and Pahor

[1] https://www.clarin.eu/parlamint

Katja Meden, Jure Skubic and Tomaž Erjavec 2023. Adding political orientation metadata to ParlaMint corpora. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 164–175. DOI: https://doi.org/10.3384/ecp210

De Maiti, 2021) who had experience in using ParlaMint corpora so that analyses could take these further variables into account. But while the information on who is a minister is an objective and verifiable fact that can be easily found, political orientation is a much more controversial piece of information.

2 Political orientation

Political orientations (or political positions) are an interesting research concept in the social sciences, understood as a set of ethical ideals, principles, or doctrines of a social movement, institution, class, or large group that explain how society should function and provide a political and cultural blueprint for a particular social order (Blattberg, 2001). They are concerned with the allocation and use of (political) power and are usually pursued by political parties.

Political orientation can refer to any number of dimensions but is most often characterized and classified on a political left-to-right spectrum, usually represented with geometric axes corresponding to independent political dimensions (Heywood, 2021). The left-to-right (LR) dimension is one of the most common dimensions used as a measure of social, political, and economic stance. Originally, the terms ”left” and ”right” were used to describe the nature and ideological beliefs of political parties: ”left” as the ”parties of movement,” which are radical, progressive, and liberal, and ”right” as the ”parties of order,” which are conservative, traditional, and authoritarian (Knapp & Wright, 2006), and such classification has, although in various forms, been retained until today. Left-right conceptualization is often considered controversial not only in terms of being defined as too simplistic and insufficiently representative to describe variations in political beliefs but also in terms of dimensionality. Most commonly LR divide is understood as unidimensional (structured by socio-economic issues) whereas some authors opt for multidimensionality where despite the importance of the socio-economic content, the left-right divide also correlates with other, non-economic issues (such as religious or ”new politics” issues) (Freire, 2015). Despite said controversies, left-right conceptualization is still the most common way to describe the ideological position of political parties and their members. The division into ”left” and ”right” has formed a categorization of ideologies, a tool for classifying political orientation, a communication code, and an instrument for guiding voters in interpreting decisions and political phenomena (Freire, 2015).

The left-right characterization of political parties plays a crucial role in theorizing about many different aspects of democratic processes (Gabel & Huber, 2000), and sociology and political science have adopted and used it despite various scholarly reservations. Some disciplines, such as history, however, often refrain from using the left-right political spectrum to characterize the ideological beliefs of political parties.

Data on political positions are often collected by conducting expert surveys, analyzing the positions of party supporters in mass surveys, or analyzing party manifestos. In expert surveys, experts provide estimates of the left-right position of parties by ranking them on a predetermined political position scale. According to Gabel and Huber, 2000, such surveys are useful but have several limitations, the most common being irregular implementation. Analysis of partisan orientation in mass surveys (Eurobarometer, World Values Survey, etc.) is more common but often provides incomplete data because they are available only for a limited number of countries. In recent years, scholars have attempted to overcome these problems by extracting party positions from party manifestos. Several studies (Gabel & Huber, 2000; Heywood, 2021) conclude that such data are very useful because they provide comparable means of assessing party positions over a long period of time in different countries. Data from party manifestos are also consistent with parties' self-positioning on the left-right spectrum and provide useful insights into how parties view themselves in terms of their political ideology.

Most work in NLP attempts to determine political orientation directly from texts (whether from political tweets (Cohen & Ruths, 2021) or parliamentary debates (Yan et al., 2017)) and thus focuses on individual speeches. Unlike related work, we have instead focused on providing information about the political orientation of a political party rather than speeches and thus took the political orientation of a speech to follow from membership in a particular party to which the speaker belonged at the time of their speech. However, as mentioned earlier, the addition of metadata labels with information about the political orientation of individual political parties collected from a combination of sources can add value

to the already extensive corpora and facilitate future research.

3 Data sources

The information on the political orientation of political parties contained in the ParlaMint corpora was gathered from three sources:

1. the Chapel Hill Expert Survey Europe (CHES Europe) (Jolly et al., 2022)[2];
2. Wikipedia entries on political parties; and
3. the corpus compilers' knowledge of political parties and their orientations.

We discuss each one in turn.

Chapel Hill Expert Survey: The CHES datasets contain expert data with built-in contextual and domain knowledge. They contain data on parliamentary political parties from countries, primarily from the EU, their attitudes toward European integration and specific EU policies, and on more specific topics such as corruption and anti-Islam rhetoric. We used two CSV files provided by CHES, namely the 1999-2019 trend file[3], which gives the values of the variables according to the covered years, and CHES 2019[4], which adds data for Norway, Iceland, and Turkey, as these were not covered in the CHES 1999-2019 trend file. This also means that these three corpora do not contain diachronic data.

The union of both CHES files provides 85 distinct variables on a given (political) position for each party and year covered, with most having a real value on the scale from 0 to 10, e.g. the variable `lrgen` measures the party's position in relation to its overall ideological stance on a scale from 0 (extreme left) to 10 (extreme right), with 5 representing the centre position. This wealth of data could be of great value to political scientists basing their research on the ParlaMint corpora. However, the CHES information also has drawbacks which can be seen especially in its coverage:

- CHES does not cover all ParlaMint corpora, in particular Bosnia, Serbia and Ukraine, as they are not part of the EU (candidate countries), nor Catalonia and Galicia, as they are not countries but autonomous regions;
- Many political parties included in ParlaMint could not be identified in the CHES dataset: of the 576 political parties belonging to the countries covered by CHES and that are included in ParlaMint, only 237 (41%) could be matched with a CHES party identifier;
- Even for the parties that are identified, CHES only covers the period to 2019, while ParlaMint extends to 2022; furthermore, not all variables are covered for all years, nor do the two input files share all the variables.

Wikipedia: The second source and type of data included is Wikipedia, in particular the data on the left-right spectrum of political orientation. This data was gathered by manually searching for the political parties' Wikipedia pages, which typically list their political orientation in the infobox of the Wikipedia article, although, for some, a more detailed examination of the Wikipedia article was required. We based our research on the English versions of the Wikipedia pages. When we could not find relevant information on the English page, we searched and translated the Wikipedia pages in the native language of the party's country. However, if there was no Wikipedia article for a particular political party or the political orientation information was not available there (in English or native language), we checked other sources (e.g. the websites of national parliaments) and extracted the information from there, also preserving the URL. It should be noted, however, that such cases were rare. Wikipedia uses values ranging from far-left to far-right, where in total, we identified 13 different values within the left-right scope, as well as 5 additional values which refer to specific political orientations outside the left-right scope, which are shown in Table 1.

[2]https://www.chesdata.eu/ches-europe
[3]https://www.chesdata.eu/s/1999-2019_CHES_dataset_meansv3.csv
[4]https://www.chesdata.eu/s/CHES2019V3.csv

Abbreviation	Value
FL	Far-left
LLF	Left to far-left
L	**Left**
CLL	Centre-left to left
CL	Centre-left
CCL	Centre to centre-left
C	**Centre**
CCR	Centre to centre-right
CR	Centre-right
CRR	Centre-right to right
R	**Right**
RRF	Right to far-right
FR	Far-right
BT	**Big tent**[5]
PP	**Pirate Party**[6]
SY	**Syncretic politics**[7]
SI	**Single-issue politics**[8]
NP	**Nonpartisanism**[9]

Table 1: Political orientation values, identified in the Wikipedia data.

The information from Wikipedia covers the ParlaMint political parties and parliamentary groups quite well: out of 932 such entities currently defined in ParlaMint, only 20 (2.2%) could not be assigned a left-right orientation.

Encoder classification: The third source of data were the encoders (i.e. compilers of the corpus), who, if they so decided, entered their classification on the left-right orientation, which was mainly so as to be able to mark the political parties that were not covered by Wikipedia. Currently, only three of the partners made use of this option.

The combination of sources proved useful in several aspects: The CHES datasets provided us with expert data on many dimensions associated with the political orientation of parties on a numerical scale and also offered the possibility of tracing changes in the political orientation of a particular party over the years, provided that the party had been a member of parliament for several years. However, as mentioned, its coverage is limited. Therefore, the second source, Wikipedia, has much greater coverage, even if the data is not as reliable as that of the CHES expert dataset, and also gives us only one dimension or political orientation, i.e. its category on the left-to-right scale.

[5] A big tent party, or catch-all party, is a term used in reference to a political party's policy of permitting or encouraging a broad spectrum of views among its members. https://en.wikipedia.org/wiki/Big_tent.

[6] Pirate Party refers to political parties that support civil rights, direct democracy, encourage innovation and creativity, free sharing of knowledge, information privacy, free speech, anti-corruption, net neutrality and oppose mass surveillance, censorship and Big Tech. https://en.wikipedia.org/wiki/Pirate_Party.

[7] Syncretic politics refers to politics that combine elements from across the conventional left–right political spectrum. https://en.wikipedia.org/wiki/Syncretic_politics.

[8] Single-issue politics refers to a political stance that is based on one essential policy area or idea.https://en.wikipedia.org/wiki/Single-issue_politics.

[9] Nonpartisanism refers to a political stance that does not agree with the current political party system. https://en.wikipedia.org/wiki/Nonpartisanism.

4 Data encoding

The task of encoding the data was divided into two parts: The first part consisted of the automatic extraction of the values from the CHES dataset for each political party included in the ParlaMint corpora. After the initial extraction of the CHES data, the identifiers of the parties in the dataset (CHES_ID) were automatically matched with the abbreviations from ParlaMint (PM_ID) using the following heuristics:

- Exact match: if the ParlaMint abbreviation was an exact match to the CHES identifier, the matching values were given in the corresponding fields (PM_ID and CHES_ID);
- Fuzzy match: an attempt was made to match the ParlaMint abbreviation without punctuation; if a fuzzy match was found, the matching values were given in the corresponding fields;
- Multiple matches: if multiple matches were found, all ParlaMint party abbreviations were output in separate rows with identical CHES-related columns;
- No match found: if no match was found for a CHES_ID, the PM_ID in the corresponding row was given a value of ”-” for ”unknown”.

For all ParlaMint parties for which no match was found, additional rows were added to the TSV; these contain the PM_ID, with all other CHES-related columns having the value ”0”. The second part consisted of manually editing the automatically generated TSV files to match the ParlaMint parties with the CHES parties in cases where no automatic match was found, but one was present. Special attention had to be paid to parliamentary groups that did not correspond to a single party but included several parties with possibly different political orientations - we handled such cases by inserting the value of the closest political party (if such a party existed) or we did not insert the value at all if no party corresponded well to the parliamentary group.

Since the ParlaMint corpora are encoded in XML according to the Text Encoding Initiative (TEI) Guidelines, the structures encoding the added metadata can be quite complex. Therefore, to simplify the process of adding metadata and make it less error-prone, we did not require the orientation data to be entered directly into XML but prepared tabular TSV files for each country that were pre-populated with the abbreviations of all political parties.

The Wikipedia URLs and the orientation data as well as the encoder orientation data were then added in Excel, possibly with comments, and the files were saved as TSV[10]. An XSLT script then takes the TSV files and the XML corpus file with the organisational data and inserts the new data into the XML file. A similar procedure was applied to the CHES data: Here, too, the CHES CSV files were converted to TSV, the party abbreviations from CHES were mapped semi-automatically in Excel to the ParlaMint party identifiers, the results were saved as TSV and again inserted into the XML files.

Figure 1 gives an example of the political orientation encoding. It should be noted that the CHES variables as well as the Wikipedia and encoder left-right orientations are pointers to taxonomy categories, which give the name and explanation of the reference, e.g. similarly to the categories and explanations presented in Table 1.

5 Metadata analysis

This section presents statistics of the added political orientation metadata, first examining the coverage of the CHES and Wikipedia TSV files separately to determine the coverage of both datasets, particularly with regard to missing values. With regard to the completeness of the CHES dataset, we first examined the percentage of available data for each CHES value (85 values in total) per ParlaMint country, the results are shown in Figure 2.

Austria (AT) stands out as the country with the most comprehensive variable coverage, with certain variables reaching up to 97% of the values (e.g., `lrgen`, `lrecon`, `eumember`, or `galtan`). Following closely behind are Estonia (EE) and Lithuania (LT), where the best-covered variables range between

[10]The TSV files are available on the ParlaMint GitHub page at the following link.

```
<org role="parliamentaryGroup" xml:id="MR">
  <orgName full="abb">MR</orgName>
  <orgName full="yes">Mouvement Réformateur</orgName>
  <idno type="URI"
  subtype="wikimedia">https://en.wikipedia.org/wiki/Reformist_Movement</idno>
  <state type="politicalOrientation">
    <state type="encoder" source="#GrietDepoorter" ana="#orientation.CRR">
       <note xml:lang="en">Orientation determined by encoder, using own
       knowledge of the parliamentary group.</note>
    </state>
    <state type="Wikipedia"
     source="https://en.wikipedia.org/wiki/Reformist_Movement"
     ana="#orientation.CR">
      <note xml:lang="en">From 1992 the Reformist Movement (MR) consisted of:
      FDF, MCC, PRL and PFF.
      In September 2001, FDF decides to leave the alliance and chooses a
      new name, becoming DeFI.</note>
    </state>
  </state>
  <state type="CHES" key="106" n="MR" from="2002" to="2018"
    source="https://www.chesdata.eu/s/1999-2019_CHES_dataset_meansv3.csv">
    <state type="variable" ana="#ches.lrgen">
      <state type="value" from="2002" to="2005" n="6.35"/>
      <state type="value" from="2006" to="2009" n="6.67"/>
      <state type="value" from="2010" to="2013" n="7.0"/>
      <state type="value" from="2014" to="2018" n="7.0"/>
    </state>
    ...
    <state type="variable" ana="#ches.vote">
      <state type="value" from="2002" to="2005" n="10.1"/>
      <state type="value" from="2006" to="2009" n="11.4"/>
      <state type="value" from="2010" to="2013" n="9.28"/>
      <state type="value" from="2014" to="2018" n="9.6"/>
    </state>
  </state>
</org>
```

Figure 1: Encoding of political orientation in ParlaMint.

92% and 94%. Conversely, Turkey (TR) and France (FR) exhibit the lowest variable coverage. For FR, the variables with the most data only encompass 27%, while for TR the coverage does not exceed 12%. This data scarcity for TR may not be surprising considering that the country was only included in the 2019 edition of the CHES surveys alongside Norway (NO) and Iceland (IS). In comparison, however, NO and IS have some variables that are still relatively well covered (between 75% and 80%).

In general, the variables with the most comprehensive coverage in the dataset are `year`, `lrgen`, `lrecon`, `galtan` (party's position in relation to its views on social and cultural values), `eu_position` (overall orientation of the party leadership towards European integration), `eu_dissent` (degree of dissent on European integration) and `eu_salience` (relative importance of European integration in the party's public stance), (which all account for 68.03% when calculating the percentage of available data per variable), while the variables with the least available data `galtan_sd` (standard deviation of expert placement of the party in 2019 concerning its views on democratic freedoms and rights), `lrecon_sd` (standard deviation of the party's expert ranking in 2019 in relation to its ideological stance on economic issues) with 1.92% of available data and `eu_econ_require` (party's position on fulfilling the economic requirements of EU membership), `eu_googov_require` (party's position on fulfilling the good governance requirements of EU membership) and `eu_political_require` (party's position on fulfilling the political requirements of EU membership) with only 0.44% of available data[11]. One of the reasons for the low coverage of the

[11]Expanded definitions for the variables can be found in the 1999-2019 Chapel Hill Expert Survey (CHES) trend file

Figure 2: Percentage of data available for each variable in the CHES dataset per an individual ParlaMint country, variable definitions are available in the 1999-2019 CHES Codebook.

aforementioned variables with the lowest coverage is the fact that these variables are only included in the 2019 CHES dataset and were not measured in any other year/survey. In contrast, the percentage of missing data for the Wikipedia values on political orientation is only 15.25%, which provides good coverage for further analysis.

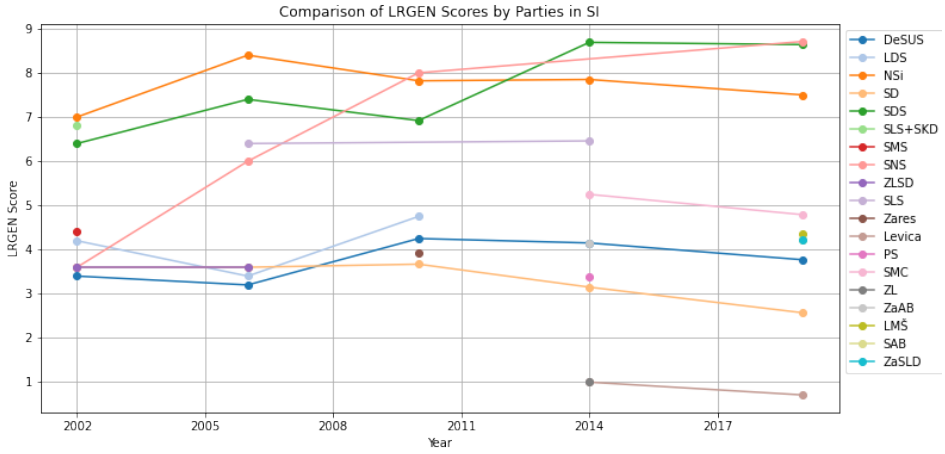
In analyzing the content of the CHES variables, we examined the aforementioned `lrgen` variable (indicating the party's position in a given year concerning its overall ideological stance) and the `lrecon` variable (indicating the party's position in a given year in terms of its ideological stance on economic issues) for the CHES dataset, visualizing some data trends for country comparisons as well as examining individual countries (and their political parties), in particular creating line charts for each country (provided that they were included in CHES dataset) to visualize changes in `lrgen` and `lrecon` values to examine trends in parties' political orientations over time. An example of such an analysis of the variables `lrgen` (Figure 3a) and `lrecon` (Figure 3b) per year for ParlaMint-SI is shown in Figure 3, which allows a comparison of the values in the case of Slovenian political parties for several years in the period from 2002 to 2019. In figure 3a, a distinction between (centre-) left and (centre-) right can be seen, with some of the notable examples, such as the political parties SDS (Slovenian Democratic Party) and NSi (New Slovenia – Christian Democrats) on the far right of the spectrum, DeSUS (Democratic Party of Pensioners of Slovenia) and SD (Social Democrats) on the left and ZL (United Left) and Levica (The Left, successor to United Left) on the far-left, with one exception - the political party SNS (Slovenian National Party) starts with a value of 3.6, a relatively (centre-)left value in 2002, which rises sharply to a value of 8.7 by 2019, surpassing the SDS (value 8.64) as one of the most right-wing political parties in Slovenia.

Similar distributions can also be seen in Figure 3b, where the distribution of `lrecon` values is relatively similar to that of `lrgen` values, which could indicate a possible correlation between the parties' general ideological position and their economic policies. This type of analysis could be extended further by comparing the scores with, for example, the variable `family`, which indicates the ideology of a single political party (where, for example, the SDS is noted as a conservative party, while the SNS is labelled as a radical right-wing (Rad right) political party, despite having a very similar, almost identical `lrgen` score). However, as shown, the data set is very limited for specific variables or countries, so any analysis should be carried out carefully and the coverage of the selected countries and/or variables should be checked. In the case of several countries (e.g. NO, IS, HR and TR), the data points are very limited and often cover only one or two years, making accurate analysis impossible.

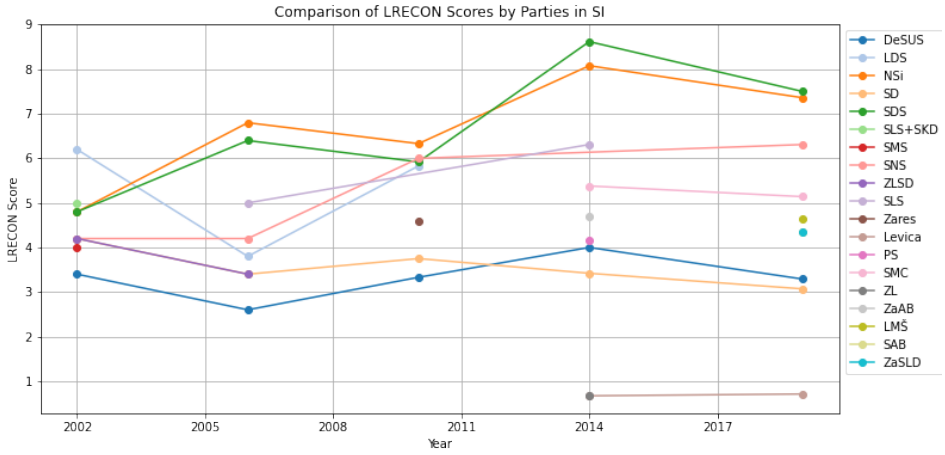
In addition to the analysis of the CHES dataset, we also performed a more in-depth analysis of the Wikipedia dataset, for which we aggregated the per-speech metadata of all corpora, (utterance ID, speaker name, their political party at the time of speaking, L-R orientation from Wikipedia etc.) into one large TSV file comprising almost 8 million lines. The entire dataset consists of 7,995,766 utterances, 22,641 unique speakers, 774 unique political parties, and 54 identified political orientation categories. However, the dataset had to be filtered due to the problem of multiple affiliations of speakers, where in some cases a small number of utterances belong to speakers who are noted as members of multiple political parties, due to the fact that the information was either not processed correctly when the dataset was created or the data was not coded correctly, due to a 1-day overlap when a person changed their political party[12]. The problem manifested itself in problematic values for political orientation such as "`Big tent;Centre`" or even "`Centre-leftCentre-leftCentre-left-centre-left`". However, such values account for only 1.63% of the whole dataset (or, more specifically, 1,63% of all utterances 2.48% of total unique speakers and 7.75% of total unique parties) and these were not considered in the statistical analysis.

The filtered dataset, where the problematic values were omitted, consists of 22,078 unique speakers, 714 unique political parties, 18 orientations (preserving the Wikipedia-extracted values in Table 1) and 7,865,408 unique utterances in total. In particular, we examined the number of utterances and the number of speakers, the parties from which the utterances were spoken, and which were linked to specific political orientations for the entire ParlaMint dataset, as shown in Table 2.

[12]This problem will be corrected with the next maintenance version of the ParlaMint dataset.



(a) Plot of the lrgen values for SI per individual year.



(b) Plot of the lrecon values for SI per individual year.

Figure 3: Example of an analysis of the variables *lrgen* (a) and *lrecon* (b) for the political parties of ParlaMint-SI. The diagrams show the changes in the general and economic political orientation of the political parties for a period between 2002 and 2019.

The table shows the number of utterances, speakers and parties per individual political orientation as given in the Wikipedia values we extracted. One of the pieces of information included is also the number of utterances that do not contain any information on political alignment (in the table labelled as Missing data). This is either because this information was not available for a particular party, or in cases where the speaker does not belong to a political party (e.g. a guest speaker). Of the other categories with available data, `Centre-right` is the orientation with the largest number of spoken utterances, followed by `Centre-left` (which also contains the largest number of political parties and spoken utterances, making centre-left speakers the most vocal), `Right` and `Centre`. This is generally not too surprising, as these are relatively common categories when it comes to the political spectrum between left and right. Of the more nuanced political orientations (which tend to be less present in the left-right spectrum), `Centre-left to left`, closely followed by `Centre-right to right` and `Right to far-right` seem to predominate in terms of spoken utterances (and the large proportion of active speakers). Finally, looking at the distribution of spoken utterances for the orienta-

Table 2: Summary of political orientation statistics - an overview number of speakers, political parties, and utterances that belong to individual political orientation categories. Political orientations are based on Wikipedia-extracted values and range from far-left to far-right, with additional categories for other political alignments outside the left-right scope (Big tent, Pirate party, Single Issue Politics, Syncretic politics).

Political Orientation	Utterances	Speakers	Parties
Missing data	692,341	7121	147
Far-left	49,293	106	9
Left to far-left	176,621	245	16
Left	198,034	534	58
Centre-left to left	406,568	815	40
Centre to centre-left	269,422	730	62
Centre-left	1,517,916	3623	107
Centre	644,572	2076	196
Centre to centre-right	331,666	831	40
Centre-right	1,743,189	3621	256
Centre-right to right	401,967	1225	41
Right	759,368	1616	52
Right to far-right	385,432	1058	47
Far-right	76,019	322	21
Big tent	175,955	980	24
Pirate Party	10,950	29	1
Single Issue Politics	19	2	2
Syncretic Politics	26,076	53	5
Total	7,865,408	24987	1124

tions outside the left-right range, it can be seen that relatively many spoken utterances (and a large number of speakers) come from the political parties of the `Big Tent`, while for the parties of the `Single Issue Politics` only two speakers from two different parties can be found in the data set. Furthermore, the political orientation "non-partisanship" (NP) does not appear in the ParlaMint corpus, or rather, no utterance was produced by a speaker belonging to a non-partisan political party.

6 Conclusions

We presented ongoing work to add political orientation metadata to the ParlaMint II parliamentary corpora. We have captured the political orientation of more than 350 European political parties by relying on two highly informative data sources, the Chapel Hill Expert Survey dataset and the Wikipedia pages of the respective parties, facilitating manual annotation of the political orientation on individual speeches from the corpora.

We faced several challenges and conceptual constraints, such as dealing with the political orientation of parties that were derived from others or were renamed. Regarding the CHES dataset, it could be argued that the dataset is somewhat sparse and "outdated" as it was last updated in 2019 and therefore does not provide information on the political orientation of parties formed after 2019[13]. As we initially only collected data for political orientation (i.e., only the `lrgen` variable, before deciding to integrate the entire CHES dataset) we identified this as a potential problem, which was addressed by using Wikipedia as a secondary source.

[13]However, a series of new CHES surveys have just been published, which will provide new data for the period up to 2022.

Contrary to our initial assumptions when comparing numerical values, we found that the Wikipedia data was highly consistent with the CHES variable `lrgen` and no major discrepancies were found between the comparison of the two sources. We attribute this to the fact that we had originally chosen to label the parties in more detail (e.g. left to centre-left) rather than simply left/centre/right. This allowed us to bridge minor differences between the two data sets. Example: When CHES indicated political orientation as centre-left and Wikipedia indicated left to centre-left, we understood this not as a contradiction, but as two alternative ways of labelling party orientation. If CHES labelled a particular party as left and Wikipedia as centre, this was understood as an inconsistency and we had to adjust our workflow accordingly. However, this was done during the initial compilation of the dataset and would require further testing to fully confirm.

We are aware that the political orientation of parties does not necessarily coincide with the personal orientation of the speaker belonging to the respective party and also recognize that people's ideological beliefs, as well as what they say, are often fluid and therefore difficult to capture. Nevertheless, the method that we have employed does give each speech its implied political orientation. The analysis carried out so far first gave us an insight into the composition of the metadata sets, both for the CHES and for the values extracted from Wikipedia, particularly with regard to data availability (especially for the CHES dataset). While the Wikipedia values have a much better coverage compared to the CHES data, there are still a large number of utterances (Table 2) that do not contain information on political orientation (the reason for this could be that the speakers do not belong to a political party or the information was not available for that particular party). On the other hand, even if its coverage is problematic, the CHES dataset still contains enough data on some country and/or variables so that the analysis can be performed without concerns about the balance of the dataset. One such example we presented in the analysis is tracking changes in general political orientation and economic policy orientation (`lrgen` and `lrecon` variables, respectively) for the period between 2002 and 2019 for Slovenian political parties. Finally, the analysis of the distribution of utterances between political orientation categories (from Wikipedia) for the entire metadata of the ParlaMint corpus provided a more comprehensive picture of the political landscape within the ParlaMint corpora, as it shows the distribution of political orientations among speakers and parties, indicating the diversity of perspectives and ideologies within political discourse.

In the future, we would like to gain further insights into the data by extending the current analysis to include the analysis of individual corpora using the CHES variables. At the time of writing, a new set of CHES datasets has just been released, alleviating some of the limitations in data availability and providing new variables for new types of analysis. In addition, we would like to expand our current analysis to focus more on the exploration of the content present in the corpora. Specifically with regard to political orientation, we would also like to enable a comparison of the speeches of left/right or centre-leaning speakers (or political parties) with each other to see whether they speak according to their political alignment or rather according to the political orientation of the political party to which the speaker belongs – instead of relying solely on the speaker's metadata, we could use various NLP-based techniques to analyse the speeches, statements or topics discussed and infer the speaker's current political stance, which may differ from the political orientation of the party the speaker belongs to. This type of analysis could then also be done for specific topics (e.g., attitudes toward European integration) that are included in the CHES metadata. In addition, the metadata will be used as part of the shared task on ideology and power identification in parliamentary debates[14], which will be part of the Touché lab[15] at the CLEF 2024[16] conference[17]. Lastly, we hope to include additional metadata useful to humanities and social scientists using ParlaMint corpora for their research, such as V-Dem[18] (Coppedge et al., 2021) and Party Facts[19] (Döring & Regel, 2019) datasets.

[14] https://touche.webis.de/clef24/touche24-web/ideology-and-power-identification-in-parliamentary-debates.html
[15] https://touche.webis.de/clef24/touche24-web/index.html
[16] http://clef2024.clef-initiative.eu/
[17] For simplicity, only the left-to-right labels will be used, flattening the fine-grained annotations but still making use of it.
[18] https://v-dem.net/
[19] https://partyfacts.herokuapp.com/documentation/about/

References

Bischof, K., & Ilie, C. (2018). Democracy and discriminatory strategies in parliamentary discourse. *Journal of Language and Politics*, *17*(5), 585–593. https://doi.org/https://doi.org/10.1075/jlp.00017.edi

Blattberg, C. (2001). Political philosophies and political ideologies. *Public Affairs Quarterly*, *15*(3), 193–217.

Cohen, R., & Ruths, D. (2021). Classifying Political Orientation on Twitter: It's Not Easy! *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 91–99. https://doi.org/10.1609/icwsm.v7i1.14434

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Alizada, N., Altman, D., Bernhard, M., Cornell, A., Fish, M. S., et al. (2021). V-dem dataset v11. 1.

Döring, H., & Regel, S. (2019). Party facts: A database of political parties worldwide. *Party Politics*, *25*(2), 97–109. https://doi.org/10.1177/1354068818820671

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2022). The ParlaMint corpora of parliamentary proceedings [https://doi.org/10.1007/s10579-021-09574-0]. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0

Fišer, D., & Pahor De Maiti, K. (2021). "First, I'm a Female Politician, Not a Male One, and Second …": A Corpus Approach to Parliamentary Discourse Research. *Contributions of contemporary history*, *61*(1), 144–179. https://doi.org/10.51663/pnz.61.1.07

Freire, A. (2015). Left–right ideology as a dimension of identification and of competition. *Journal of Political Ideologies*, *20*(1), 43–68.

Gabel, M. J., & Huber, J. D. (2000). Putting parties in their place: Inferring party left-right ideological positions from party manifestos data. *American Journal of Political Science*, 94–103.

Heywood, A. (2021). *Political ideologies: An introduction*. Bloomsbury Publishing.

Jolly, S., Bakker, R., Hooghe, L., Marks, G., Polk, J., Rovny, J., Steenbergen, M., & Vachudova, M. A. (2022). Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies*, *75*, 102420. https://doi.org/https://doi.org/10.1016/j.electstud.2021.102420

Knapp, A., & Wright, V. (2006). *The government and politics of France*. Routledge.

Yan, H., Lavoie, A., & Das, S. (2017). The perils of classifying political orientation from text. *Linked Democracy: Artificial Intelligence for Democratic Innovation*, *858*, 8.

Re-Reading Lists in Historical Newspapers: Digital Insights into an Overlooked Text Type

Nina C. Rastinger
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences
ninaclaudia.rastinger@oeaw.ac.at

Abstract

The paper presents an ongoing doctoral project dedicated to periodically published lists in historical newspapers between 1600 and 1850. By employing approaches from Corpus Linguistics and Digital Humanities, the project aims to locate the studied 'small' texts within existing digital resources, analyse them with regard to their textual characteristics and evaluate their potentials and challenges for automated information extraction. The article primarily focuses on two key aspects: firstly, on search strategies for locating lists in digital newspaper corpora and collections, and secondly, on a case study into lists of arriving persons published in the *Wien[n]erisches Diarium* between 1703 and 1725. These empirical investigations reveal that periodically published lists form a central and frequent component of early modern newspapers and offer numerous potentials for Digital Humanities research due to their textual features, such as periodicity, repetitiveness or inherent (semi-)structuredness. In this regard, the paper identifies the overlooked newspaper text type as a data treasure awaiting discovery and underscores the need to investigate 'small' newspaper texts on a large scale.

1 Introduction

The creation and use of lists constitute central components of everyday life and function as fundamental cultural techniques (cf., e.g., Vismann, 2000: 20; Adelmann, 2021: 26). Subsequently, we are highly proficient in dealing with lists, as Esposito (2017: 356) summarises: "Lists are easy to write and easy to read." According to Goody (1977) and Waldispühl (2019: 197–198), this self-evidence of lists is rooted in a long-standing tradition that began with the emergence of writing systems and has since continued through different eras and writing cultures, constituting a historical and cultural continuum of the text form. Paradoxically, it seems to be precisely the omnipresence of lists that has led to them being neglected in research (cf. Waldispühl, 2019: 198; Schaffrick & Werber, 2017: 303). In addition, lists closely align with Hausendorf's (2009) concept of 'small' texts and are, in prototypical cases, perceived as 'small' in multiple senses:

1) size: lists as short and/or small-scale texts;
2) complexity: lists as simple, non-complex texts;
3) functionality: lists as praxis-oriented texts;
4) design: lists as stereotypical, template-like texts;
5) ambition: lists as unambitious, unelaborate texts.

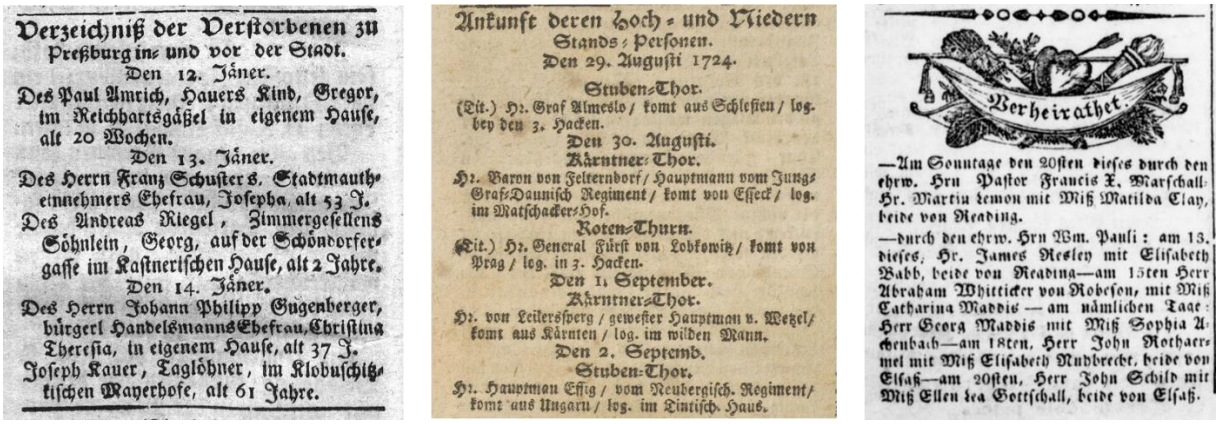
These five textual features, which are tightly intertwined with each other, might have further contributed to the oversight of lists: as 'small' and ubiquitous texts, lists tend to have been underestimated or deemed irrelevant for scholarly research.

A systematic exception is currently only formed by literary lists, i.e., lists in and around literature, as discussed, among others, by Belknap (2004), Jullien (2004), Mainberger (2003, 2017, 2018), Fludernik (2016), von Contzen (2018, 2017a, 2017b, 2016), Rüggemeier (2020), Barton et al.

Nina C. Rastinger 2023. Re-Reading Lists in Historical Newspapers: Digital Insights into an Overlooked Text Type. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 176–190. DOI: https://doi.org/10.3384/ecp210

(2022) and Chinca et al. (2022). Also, lists that originate from modern, digital contexts can be identified as a growing research subject. Esposito (2017), Adelmann (2021), and Esposito and Stark (2019), for instance, deal with ratings and rankings in popular culture and the World Wide Web, Diederichsen (2021) provides an insight into playlists, Temmerman and Vandenabelee (2018) analyse the new journalistic genre of 'listicles', and Bubenhofer (2020) discusses the use of lists as a form of visualisation in digital linguistics.

In the overall perspective, however, the gap prevails: contrary to their frequency, historical non-literary or 'practical' lists have only been considered as a text type to be investigated in isolated cases (e.g., Doležalová, 2009; Ledin, 2015; Waldispühl, 2019). This 'oversight' particularly concerns studies on early modern press products, as a high number of historical newspapers contain periodically published lists, e.g. arrival lists (i.e. lists of persons who have arrived in a certain city), death lists, marriage lists, birth lists or price lists (cf. Figure 1). Such lists were published recurrently over a certain period, appearing monthly, weekly or, in many cases, in every issue of a periodical. Hence, unlike lists printed as parts of news articles, advertisements, or other sections, they constitute a distinct component and specific text type of historical newspapers.


Verzeichniß der Verstorbenen zu Preßburg in- und vor der Stadt.
Den 12. Jäner.
Des Paul Amrich, Hauers Kind, Gregor, im Reichhartsgäßel in eigenem Hause, alt 20 Wochen.
Den 13. Jäner.
Des Herrn Franz Schuster s. Stadtmauthemehmers Ehefrau, Josepha, alt 53 J.
Des Andreas Riegel, Zimmergesellens Söhnlein, Georg, auf der Schöndorfergasse im Kastnerischen Hause, alt 2 Jahre.
Den 14. Jäner.
Des Herrn Johann Philipp Gugenberger, bürgerl. Handelsmanns Ehefrau, Christina Theresia, in eigenem Hause, alt 37 J.
Joseph Kauer, Taglöhner, im Klobuschitzkischen Mayerhofe, alt 61 Jahre.

Ankunft deren Hoch- und Niedern Stands-Personen.
Den 29. Augusti 1724.
Stuben-Thor.
(Lit.) Hr. Graf Almeslo / komt aus Schlesien / log. bey den 3. Hacken.
Den 30. Augusti.
Kärntner-Thor.
Hr. Baron von Felternborf / Hauptmann vom Jungs-Graf-Daunisch Regiment / komt von Essek / log. im Matschacker-Hof.
Roten-Thurn.
(Lit.) Hr. General Fürst von Lobkowitz / komt von Prag / log. in 3. Hacken.
Den 1. September.
Kärntner-Thor.
Hr. von Leilersperg / gewester Hauptmon v. Wetzel / komt aus Kärnten / log. im wilden Mann.
Den 2. Septemb.
Stuben-Thor.
Hr. Hauptmon Essig / vom Neubergisch. Regiment / komt aus Ungarn / log. im Tintisch. Haus.

Verheirathet.
—Am Sonntage den 20sten dieses durch den ehrw. Hrn Pastor Francis X. Marschall-Hr. Martin Lemon mit Miß Matilda Clay, beide von Reading.
—durch den ehrw. Hrn Wm. Pauli: am 13. dieses, Hr. James Nesley mit Elisabeth Babb, beide von Reading—am 15ten Herr Abraham Whitticker von Robeson, mit Miß Catharina Maddis — am nämlichen Tage: Herr Georg Maddis mit Miß Sophia Ahrenbach—am 18ten, Herr John Rothacrmel mit Miß Elisabeth Rudbrecht, beide von Elsaß—am 20sten, Herr John Schild mit Miß Ellen Lea Gottschall, beide von Elsaß.

Figure 1. Death list from the *Preßburger Zeitung* (16.01.1765: 8), arrival list from the *Wien[n]erisches Diarium* (02.09.1724: 8), and marriage list from *Der liberale beobachter und Berks, Montgomery und Schuylkill Caunties allgemeine anzeiger* (29.09.1840: 3)

Nevertheless, in contrast to other newspaper text types, like news articles (e.g., Haß-Zumkehr, 1998; Pfefferkorn et al., 2017), advertisements (e.g., Bendel, 1998; Kurzmann, 1999; Farø, 2005; Ehrenheim, 2011) or reader's letters (e.g., Fix, 2008: 273–333; Fix, 2011), periodically published lists have so far neither been discussed theoretically nor examined empirically for their textual properties. Furthermore, they are missing from general overviews of (historical) newspaper text types (cf. e.g., Dovifat, 1976; Brand & Schulze, 1993; Straßner, 2000; Lenk & Chesterman, 2005) and are partly already left out during the (full text) digitisation of periodicals, which makes them all the more invisible to the scientific and public eye.

Given the lack of research, the dissertation project discussed here is dedicated to the identification and analysis of periodically published lists in German-language newspapers published between 1600 and 1850. Specifically, approaches from Corpus Linguistics and Digital Humanities are employed to pursue three key research interests: first, the text type should be identified in existing digital corpora and text collections (cf. Chapter 2), second, it should be analysed in regard to its textual characteristics (cf. Chapter 3), and third, it should be explored with respect to its challenges and potentials for the field of Digital Humanities, especially focusing on (semi-)automatic approaches to text analysis and information extraction (cf. Chapter 4). This three-step process of the dissertation, both in terms of content and practical procedure, will be detailed in the following chapters.

2 Looking for lists: identification of 'small' texts in historical newspapers

In the last decades, newspapers and journals have undergone mass digitisation (cf., e.g., Nicholson, 2013; Blome, 2018; Ehrmann et al., 2023). As a result, users can nowadays access facsimiles and/or full texts of (German) historical newspapers via various online resources, such as *AustriaN Newspapers Online* (= ANNO, Österreichische Nationalbibliothek), *Chronicling America* (Library of Congress),

digiPress (Bayerische Staatsbibliothek), *Digitales Forum für Mittel- und Osteuropa* (= DIFMOE, Digitales Forum Mittel- und Osteuropa e.V.), *impresso* (Clematide et al.), *Teßmann digital* (Landesbibliothek Dr. Friedrich Teßmann), *zeit.punkt NRW* (Universitäts- und Landesbibliothek Bonn et al.), and many other corpora and text collections. Within the CLARIN infrastructure, there are also multiple resources that offer (often already structurally or semantically annotated) full texts of historical newspapers, for example, *Deutsches Textarchiv* (= DTA, Berlin-Brandenburgische Akademie der Wissenschaften), *HIST – Archive of the Historical Corpora* (Institut für Deutsche Sprache) and *Wien[n]erisches DIGITARIUM* (Resch & Kampkaspar). Concentrating on printed German-language newspapers from 1600–1850, the dissertation project uses (a selection of) the named resources as a starting point to locate lists in digitised press texts.

One challenge in this endeavour clearly lies in the amount of unknown factors when searching for a specific text type rather than a concrete linguistic unit, entity or topic. The fact that both the lexical content and the (typographical) structure of periodically published lists may have varied over time and between different newspapers (cf. Figure 1) prevents the use of a universal approach that encompasses all cases. Instead, to cover and uncover a wide range of lists, the doctoral project employs a mixed methods approach that combines multiple search strategies. To be precise, four different strategies are presently being used in parallel to look for lists in digital newspaper corpora and collections, namely (1) the reuse of existing annotations, (2) full text search in combination with the calculation of a word/phrase reuse ratio, (3) the training of layout recognition models and (4) the close reading of random samples.

2.1 Reuse of existing annotations

Some of the available newspaper corpora and collections do not only offer facsimiles and plain full texts, but already include structural and/or semantic-markup. A small subset of these resources also account for lists in their annotation systems. The W*ienerisches DIGITARIUM* (Resch & Kampkaspar), created at the Austrian Centre for Digital Humanities and Cultural Heritage, for instance, contains over 300 full text issues of the historical *Wiener Zeitung* in XML/TEI-P5 encoding. Within this data set, which is downloadable through the CLARIN-B centre ARCHE (ACDH-CH, https://hdl.handle.net/21.11115/0000-000F-746A-8), both singular and periodically published lists were annotated with the <list> element and further distinguished into separate entries through the <item> element (cf. Resch et al., 2023). As a result of this mark-up, users are able to systematically retrieve lists by querying for certain structural elements, with help of the Oxygen XML Editor and XPath, Beautiful Soup or other tools able to parse XML.

Although such a reuse of pre-existing annotations is highly efficient, it is rarely applicable in practice, since, at present, resources providing structural and/or semantically annotated full texts of historical German-language newspapers form an exception within the digital landscape. Instead, the majority of resources publicly available offers transcripts which were created automatically through Optical Character Recognition (OCR) and not corrected or edited further. This pragmatic decision in favor of quantity over quality makes sense considering the vast amounts of newspapers that have been preserved from the past, but, at the same time, requires users to adapt their approach to a resource accordingly. In the end, multiple factors may hinder the successful text recognition through OCR and lead to a high error rate, such as low image quality, (changes between) certain print types (cf. Kampkaspar, 2019, about OCR and Fraktur; Rastinger, in print, about OCR and print type changes between Fraktur and Antiqua), or the incorrect recognition of text regions during layout analysis (e.g., due to complex list layouts). OCR errors, in turn, not only complicate a newspaper's enrichment through annotations, but may also distort search processes and skew analyses, as shown, among others, by Torget (2023). Hence, even when deploying separate search strategies besides the reuse of existing annotations, one essentially always needs to take into account potential limitations due to restricted full text quality.

2.2 Full text search and reuse ratio

It is especially important to address the quality of full texts when using the full text search at hand in most digital corpora and resources to date. One strategy adopted here to deal with this limitation is the calculation of a reuse ratio of certain words and/or phrases (potentially) tied to periodically published lists. This approach makes use of the observation that periodically published lists tend to involve word and/or phrase reuse in multiple ways, namely (1) within a certain issue, (2) over multiple issues and (3)

between different newspapers. To give a concrete example, the death list published in the *Tiroler Zeitung* under the title *Allhier Verstorbene in und vor der Stadt* 'deceased persons here in and outside the city' repeatedly includes the term *J.* [Jahr] 'year' as an age measure, uses similar list titles and subtitles over several issues and shares certain key words (e.g., *Verstorbene* 'deceased') with periodically published lists in other newspapers, for instance the death lists of the *Bozner Zeitung* (list title: *Verstorbene in Bozen*), the *Klagenfurter Zeitung* (list title: *Verzeichniß der hier Verstorbenen*), and the *Tiroler Zeitung* (list title: *Verstorbene*).[2] In the doctoral project, such inter- and intratextual overlaps are utilised by querying for both accumulations of certain linguistic elements within a single newspaper page and patterns of word/phrase reuse across entire newspapers and newspaper collections. Both strategies should – when carried out on large datasets – help with spotting periodical text reuse and mitigating potential OCR errors.

On a practical level, this search strategy can be demonstrated using an example: a lemma that repeatedly appears within the list titles gathered so far is the noun *Civilstand* 'civil status'. In the context of early modern newspapers, this term seems to serve as an overarching concept to group together birth, marriage, and death records, and can be linked to the French *état civil*, the 'management of population lists' (*Führung der Bevölkerungslisten*), according to Herders Conversationslexikon (1854–1857). The noun was queried for in a selected collection, namely the newspaper portal *zeit.punktNRW*, funded by the state of North Rhine-Westphalia. This digital resource provides access to over 400 local German newspapers, among which 112 periodicals fall within the study period from 1600 to 1850. The approach started with compiling a list of these 112 historical newspapers. Subsequently, a web crawler, created with the python library Beautiful Soup, was employed to perform automated full text searches within the periodicals and capture the total number of pages per newspaper that include the exact term *Civilstand*. Additionally, a second search was carried out for the highly frequent conjunction *und* 'and', as to estimate the total number of (searchable) pages per newspaper. The collected counts were then used to calculate an approximate percentage of pages within one newspaper that contained the word *Civilstand*. Also, to get a better overview of the results and easily compare them with each other, every queried newspaper was given a number and was visualised as part of a heatmap:

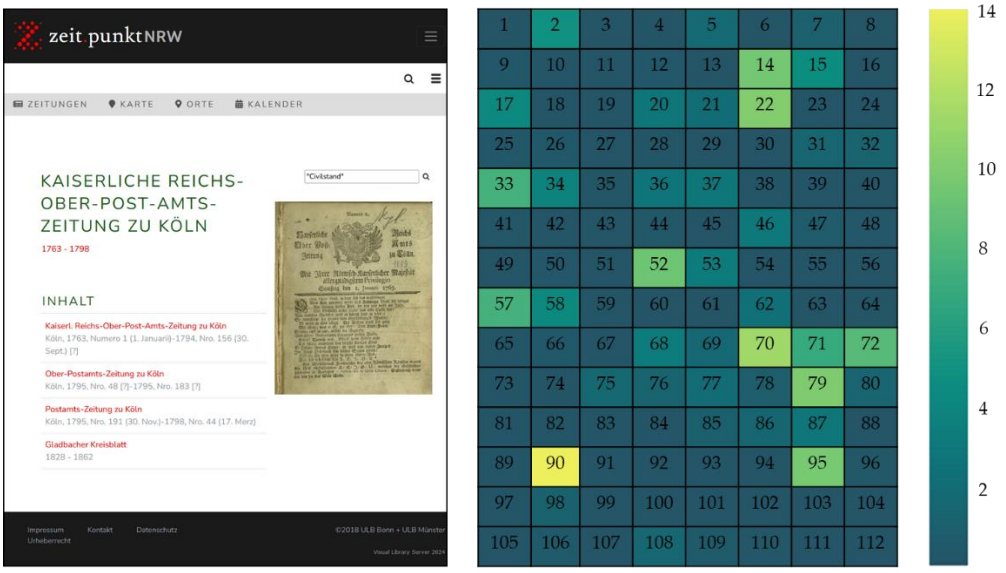


Figure 2. Search interface of an exemplary newspaper in *zeit.punktNRW* (left) and heatmap of the reuse ratio of *Civilstand* within 112 German newspapers from 1600–1850 (right)

In theory, a higher percentage – visualised in the heatmap in Figure 2 as a lighter color – suggests a greater likelihood of a certain newspaper to contain periodically published texts including the term *Civilstand*. As can be seen in Figure 2, multiple newspapers (e.g., 90 = *Täglicher Anzeiger für Berg und Mark*, 70 = *Neue Rheinische Zeitung*) available through *zeit.punktNRW* showcase such a high

[2] This should not lead to the erroneous assumption that specific types of lists, such as death lists, are homogenous in their nomenclature. Various alternative titles exist, e.g., *Starb:* 'Died:', *Beerdigte* 'Buried persons', *Sterbefälle* 'Cases of death', or *Todesfälle* 'Cases of death'.

probability. When taking a closer look into all newspapers with a ratio of at least 4 % (= 16 cases), it was found that each of them actually involved periodical lists on locals birth, marriages and deaths. Among the identified text series are, for example, *Civilstand der Stadt Luxemburg* published in the *Luxemburger Wort*, *Civilstand der Stadt Aachen* as part of the *Echo der Gegenwart*, *Civilstand der Bürgermeisterei Kempen* printed in the *Kempener Kreisblatt*, and *Civilstand der Stadt Neuwied* featured in the *Neuwiedische Nachrichten*. From a methodological view point, these and further successful findings speak for adopting a word reuse ratio as another strategy to systematically identify lists in historical newspapers and simultaneously investigate the spatial and temporal developments in their titling.

2.3 Layout recognition

At the same time, two approaches are still unlikely to be suited for all types of reoccuring lists – which is why it makes sense to expand one's perspective from the linguistic-content level to the visual-typographic level. Looking at their text surface, lists published in historical newspapers tendentially involve more 'white space' than other text types, like news or official announcements, which is usually due to more line breaks, indentations, subheadings, etc. This visual nature of lists is also addressed in research literature, for example when Belknap (2004: 15) defines a list as "a formally organized block of information that is composed of a set of members" [emphasis: N.C.R] or when Bubenhofer (2020: 31) locates lists in the transitional area between diagram and text because they relate linguistic signs to an arrangement in space.

Approaching the topic from the field of Digital Humanities, this characteristic of lists can be leveraged by training and using list-specific layout recognition models. Currently, this approach has been explored through training a layout recognition model on the layout of the historical *Wiener Zeitung* of the late 18th century. Using the transcription platform Transkribus (READ-COOP), 24 issues and over 300 pages of the Austrian periodical were structurally annotated, meaning all text regions were identified and categorised. The latter was done according to a tag set differentiating, among other elements, between paragraphs, headings (cf. pink marking in Figure 3), figures, seperators, catch words, and – of course – lists (cf. lilac marking in Figure 3). Based on these structural tags, a field model was trained, reaching an accuracy (mAP) of 88,06 %.

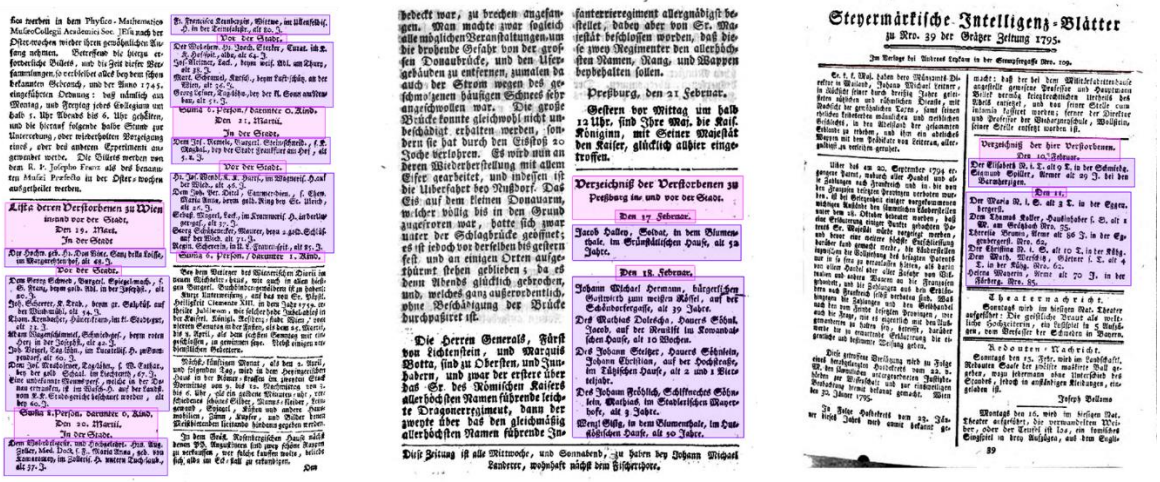


Figure 3. Exemplary results of using layout analysis for list detection, in the *Wiener Zeitung* (left), *Preßburger Zeitung* (middle) and *Grätzer Zeitung* (right)

As illustrated in Figure 3, the resulting layout recognition model cannot only support the automatic identification and annotation of lists within the *Wiener Zeitung* (left), but may also be applied to search through further newspapers with a similar layout, such as the *Preßburger Zeitung* (middle), available through *ANNO* and *DIFMOE*, or the *Grätzer Zeitung* (right), accessible in *ANNO*. By focusing on the visual level of periodically published lists, the approach enables the efficient search of large image datasets and contributes further to the systematic discovery of the 'small' texts in historical newspapers.

2.4 Towards an open-ended 'list of lists'

The three search strategies presented so far can be viewed as forms of distant reading (cf. Moretti 2016) and are complemented by the more traditional approach of close reading. Due to the amount of data available, it makes sense to select certain resources, especially prioritising those that are less accessible with distant methods, and draw random, time-distributed samples per newspaper. For assembling these samples, it has proven beneficial that multiple newspaper collections (e.g. *ANNO*, *digiPress*, *zeit.punktNRW*) offer access to an IIIF Image API with which facsimiles can systematically be exported into various viewers (e.g., *Mirador*, *Universal Viewer*) and tools (e.g., *Transkribus*). Based on the premise that the investigated lists are in particular characterised by their periodicity, systematic close reading should assist in identifying additional representatives of the newspaper text type not yet (un)covered through other approaches. In combination, the four search strategies presented both complement and inform each other, as each new (type of) list found with one method might again provide input for another approach.

The ultimate goal of this iterative and ongoing process of looking for lists is to create an open-ended 'list of lists', i.e. a database on lists which were periodically circulated through historical newspapers. Especially in the absence of research literature (cf. Chapter 1), such a compilation allows for a first-time overview of the examined 'small' texts, provides insights into their publication through time and space, and may possibly even shed light on larger scale trends and shifts within the early modern print market. Moreover, a systematic identification of periodically published lists serves as a foundational step for corpus-based analysis, as it supplies the necessary knowledge and texts for the assembly of a representative analysis corpus.

While the envisioned 'list of lists' is currently still being compiled, a preliminary interim status can already be provided: with the help of the methodic repertoire discussed in this chapter, eleven digital corpora and collections that contain German printed newspapers from 1600 to 1850 have already been (partly) queried so far. Concretely, the search has been completed for six resources (*Chronicling America, Darmstädter Tagblatt, DTA, impresso, Teßmann digital, Wienerisches DIGITARIUM*) and started for five resources (*ANNO, DIFMOE, digiPress, ZEFYS, zeit.punktNRW*).Within these corpora and collections, 192 periodically published lists were identified so far, which originate from 62 different newspapers and cover almost the whole 18th and early 19th century, with the earliest finding dating back to 1703.[3] In parallel and even though only German-language newspapers were being taken into account, the publishing locations of the found texts are scattered all over Central Europe and the text type can even be observed in periodicals published in 19th century North America. To sum it up, these initial findings already indicate the high frequency of periodically published lists in historical newspapers – and make the need for research in this area all the more clear.

3 Looking at lists: digital analysis of textual characteristics

Systematically collecting periodic lists might reveal diachronic patterns within the early modern newspaper landscape. The temporal and spatial distribution of different list types or 'paradigms', for instance, could be highly informative and forms one of the doctoral projects' core interests. In this context, the concept of the 'list paradigm' refers to a unifying aspect to which all elements of a list are related and through which they become part of a whole – a notion that is, even if not mentioned explicitly, evident in many definitions of lists, for instance by Tankard (2006) and Mainberger (2003):

(1) "A list is a written or printed series of names, dates, numbers, or items, organized according to some need or principle." (Tankard, 2006: 339)
(2) "Enumerations name distinct elements and equalize them under a thematic or formal aspect." (Mainberger, 2003: 7)

It is also true for the lists identified so far that multiple connecting themes reoccur. Currently, the most frequent list types are death lists, found in 26 newspapers (e.g., *Wiener Zeitung*, *Tiroler Zeitung*), and arrival lists, present in 17 newspapers (e.g., *Luxemburger Wort, Hamburgischer Unpartheyischer*

[3] As of now, no lists from the 17th century have been found, which may however be attributed to the choice of corpora and collections examined so far as they mainly contain newspapers from the 18th century onwards.

Correspondent). These are closely followed by price lists (16 newspapers, e.g., *Der Vaterlandsfreund*, *Innsbrucker Wochenblatt*), and marriage lists (13 newspapers, e.g., *Der Morgenstern*, *Gülich and bergische wöchentliche Nachrichten*). Moreover, one repeatedly encounters theatre performance lists (12 newspapers, e.g, *Wiener Zeitung*, *K. K. priv. Prager Zeitung*), birth lists (10 newspapers, e.g., *Bozner Zeitung, Rheinischer Merkur*) and various other types (e.g., departure lists, newspaper agent lists). Although this enumeration is of preliminary nature, it already showcases that periodically published lists in historical newspapers constitute not only a common, but also a manifold phenomenon, which could potentially be relevant to multiple disciplines.

Besides differentiating between different types of lists, the doctoral project addresses the text type's diversity by quantitatively and qualitatively inspecting its textual characteristics. Based on the 'list of lists' currently in compilation (cf. Chapter 2.4), an analysis corpus as representable as possible is going to be created, that should cover different list types, newspapers, and time periods, making them comparable to each other. The focus of the corpus linguistic analysis will then be put on textual patterns on various levels of periodically published lists, namely on the level of (1) their selection and organisation of entries, (2) their typographic structure, (3) their linguistic features and (4) their pragmatic context. On each level, selected aspects are empirically examined, e.g., the inclusion of 'empty' items (1, e.g. *Niemand* 'no one' in arrival lists), (the number of) applied ordering principles (1, e.g., hierarchical, alphabetical, chronological, geographical), typographic resources used for text structure (2, e.g., indentations, blockings, bullet points), abbreviation density (3), morpho-syntactic completeness or 'grammaticality' (3, e.g., use of verbs, interpunctuation), and text purpose (4, e.g., information, legitimation).

On a practical level, the analysis will again be carried out by integrating close and distant reading approaches, resulting in a form of 'blended reading' (Lemke & Wiedemann, 2015) or 'scalable reading' (Mueller, 2020). Analogous to Google Earth, this approach should allow for different perspectives on a singular research subject, as Mueller (2020) notes: "you can zoom in and out of things and discover that different properties of phenomena are revealed by looking at them from different distances". In this way, the doctoral project aims to obtain a first, multi-faceted picture of lists published in historical newspapers and a better understanding of their textual characteristics and diachronic development.

4 Case study: digitisation, annotation and mapping of arrival lists

The third and final aspect of the doctoral project focuses on the potentials and limitations of the newspaper text type within the field of Digital Humanities. This part of the investigation is grounded in the idea that the studied lists might be especially well-suited for semi-automatic information extraction approaches due to their periodicity, repetitive nature, semi-structured format, and high density of named entities. This theory is put into practice through a concrete use case, namely the City of Vienna funded project "Visiting Vienna – digital approaches to the (semi-)automatic analysis of the arrival lists found in the *Wien[n]erisches Diarium*" (PI: Nina C. Rastinger, 2022–2023).

As the project title reveals, the focus is put on arrival lists, namely the ones printed in the *Wien[n]erisches Diarium* (since 1780: *Wiener Zeitung*) from August 1703 to March 1725. Published under the title *Ankunft deren Hoch= und Niedern Stands=Personen*[4] 'arrival of high- and low-ranking persons of social standing', these texts were disseminated with every issue of the semi-weekly appearing Austrian newspaper and document the arrival of upper class persons, such as aristocrats, secretaries, clerics or couriers. Both Johann Baptist Schönwetter and Johann Peter van Ghelen who consecutively served as publishers of the *Diarium* at the time had access to this information through an imperial privilege (Mader-Kratky et al., 2019: 99). The preserved arrival lists are distinguished by their rich detail, providing not only personal information about the individuals arriving, such as names, professions, titles, and affiliations, but also detailed spatial and temporal data about their taken travel routes. Typically, this encompasses the place of departure, the date of arrival, the city gate used to enter Vienna, the accommodation inside or around the city, and/or the final destination. An exemplary item containing (almost) all of these components is depicted in Figure 4:

[4] Here, one exemplary spelling is displayed; overall, the list title is represented in over 60 different writing variants.

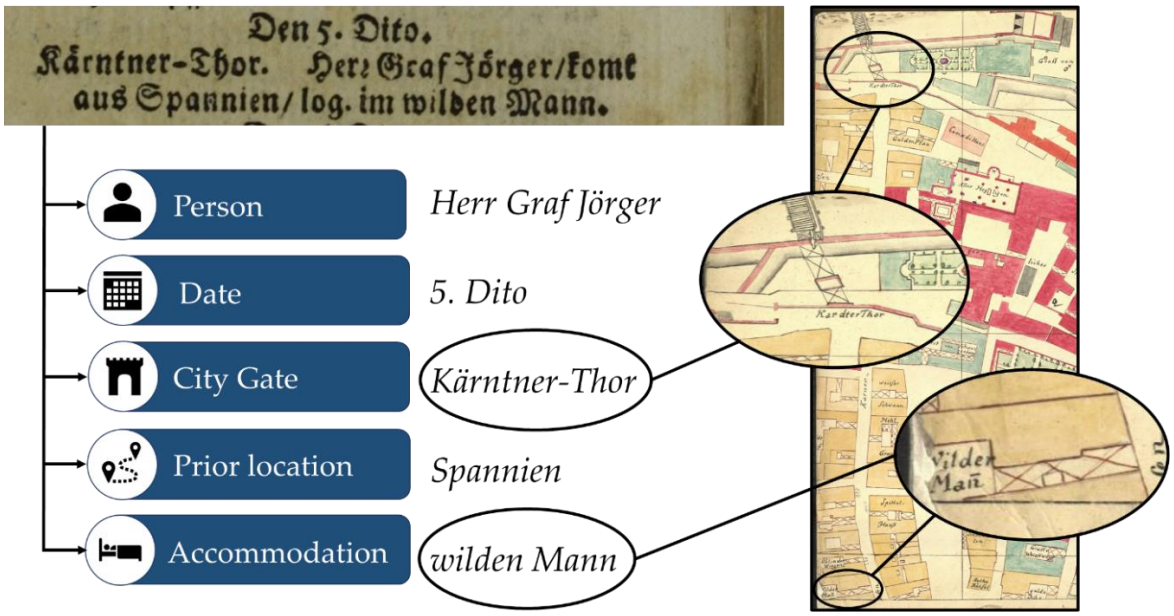


Figure 4. Graphical representation of NER and mapping of an exemplary list entry (*Wienerisches Diarium*, 07.01.1711: 7) on a replication of the Steinhausenplan (1710) by Schimmer (1847)

The illustration further portrays part of the steps taken over the course of the case study: First, reliable full texts were created (cf. Chapter 4.1), secondly, these texts were semantically enriched through named entity recognition (NER) (cf. Chapter 4.2), thirdly, the identified geographical entities were mapped on historical city plans of Vienna (cf. Chapter 4.3), and fourthly, the resulting data will be published online and archived in a long-time repository, enabling its safe storage and reuse in future research (cf. Chapter 5). With the help of this four-step workflow, the doctoral project leverages the information density of the arrival lists to trace past movements in and around an early modern city, while also exploring the potentials and limitations of the newspaper text type for (semi-)automatic information extraction processes.

4.1 Full text digitisation

To effectively analyse the Viennese arrival lists with the help of digital methods, it is essential to first establish a comprehensive and high-quality full text dataset. The newspaper *Wienerisches Diarium*, in which the periodical texts were published, is presently accessible via two digital platforms, namely *AustriaN Newspapers Online* (*ANNO*), hosted at the Austrian National Library, and the afore-mentioned *Wienerisches DIGITARIUM*, developed at the Austrian Centre for Digital Humanities and Cultural Heritage. As detailed in Chapter 2.1, the *DIGITARIUM* provides a selection of 18th century newspaper issues in a reliable, annotated full text format plus facsimiles, among which 76 issues include arrival lists. In comparison, *ANNO*, which also provides both image and text, hosts a more extensive, though also not exhaustive,[5] collection of 18th century arrival lists. The limitation is that these full texts were produced automatically using OCR without further error handling and thus partially exhibit a high Character Error Rate (CER). This issue seems to be particularly significant for the arrival lists, which are prone to have more complex layouts than standard news texts. An analysis of a balanced, random sample of ten lists from *ANNO*, encompassing around 1,500 tokens, revealed a CER of 50 %, indicating that, on average, every second character of the arrival lists was incorrectly recognised. Consequently, the full texts offered in *ANNO* are not suitable for automated NLP processes that rely on token-level accuracy and necessitate high precision, such as named entity recognition (cf. also Müller, 2016).

[5] Currently missing are, for instance, the volumes for the years 1717 and 1718, for which the arrival lists thus have to be collected and processed from other sources, e.g., the microfilms available in the Wienbibliothek im Rathaus (https://www.wienbibliothek.at). This is presently being done, with future plans to add these lists to the corpus as well.

Against this backdrop, all arrival lists were first collected as facsimiles from *ANNO*, grouped by year, pre-processed and uploaded to the transcription platform *Transkribus* (READ-COOP), where they underwent (semi-)automatic layout and text recognition. In the latter step, the HTR model "German Fraktur 18th Century – WrDiarium_M9" (Resch & Kampkaspar, 2020), that was developed during the creation of the *Wienerisches DIGITARIUM* (Resch & Kampkaspar), was applied. Since this model was trained on the 18th century *Wiener Zeitung*, achieves a low CER of 0.8 % and is publicly available, it is ideally suited for the text recognition of the arrival lists. Additionally, to ensure the high accuracy of the transcriptions, all texts were manually inspected and corrected where necessary. This digitisation process has presently been completed for almost 1,500 arrival lists from the *Wien[n]erisches Diarium*, resulting in a dataset of around 176,500 tokens which spans from August 1703 to March 1725.

4.2 Named entity recognition (NER)

High quality full texts pose a fundamental basis for reliable results from NLP tasks (cf. Ehrmann et al., 2023; Torget, 2023). However, even when given without OCR errors and other artificial noise, historical newspaper texts tend to present difficulties for NER due to their high graphematical variation and language dynamics, in addition to a scarcity of resources and other obstacles (cf. Ehrmann et al., 2023; González-Gallardo et al., 2023). In light of these challenges, multiple approaches were explored and compared throughout the case study, namely (1) the reuse of existing NER models for historical German, (2) a rule-based pipeline using spaCy's EntityRuler, and (3) the use of Large Language Models (LLMs), concretely GPT-3.5 (OpenAI). Based on a small (!), time-distributed gold standard (10 lists, 83 items, 1474 tokens),[6] it has been shown that the out-of-the-box use of existing NER models for historic German[7] yielded the worst results for the arrival lists of the *Wien[n]erisches Diarium*, achieving only low F1-scores ranging from 0.17 to 0.42. These poor results might be attributed to the fact that NER models are typically trained on running text, such as news articles, and thus struggle with the minimal or absent syntactic context of lists.

In comparison, slightly better results were obtained by creating project-specific RegEx-based rules for NER, as a relatively small set of rules was able to identify the majority of all entities in the gold standard. However, the creation of individual rules has proven to be both time-consuming and inflexible, especially considering the diachronic nature of the dataset and the planned future reuse of the workflow in other datasets. The NER strategy that ultimately achieved the best results was the use of the LLM GPT-3.5, accessed via the OpenAI API. The input and output of the model were optimised using Promptify, an open-source Python library dedicated to prompt engineering. With its help, a one-shot approach was taken in which the LLM was provided with one annotated example, i.e., one arrival list with nine entries. While all instructional prompts on the NER task were presented in English, the labels were given in German and set in close accordance to the textual structure of the studied lists. Besides refering to PERSON (person) and DATUM (date) as common NER categories, places were differentiated into STADTTOR (city gate), HERKUNFTSORT (place of departure), UNTERKUNFTSORT (place of accommodation) and ZIELORT (destination). Lastly, a postprocessing and feedback loop was integrated into the workflow to determine the correct positions of the entities and detect possible hallucinations, i.e. named entities that differ in orthographical or grammatical form in input versus output. As can be seen in the exemplary output in Figure 5, this approach yielded very good results:

[6] The gold standard was annotated with CATMA 6 (Gius et al., 2022) and exported with GitMA (Vauth et al., 2022).
[7] All models used are publically available on HuggingFace. Among the models tested were the flair models *flair-de-ner*, *flair-historic-ner-onb*, and *flair-historic-ner-lft* as well as the spaCy models *de_RTA_NER* and *easy_fnhd*.

Figure 5. Annotations of an exemplary arrival list (*Wienerisches Diarium*, 28.07.1708), visualised with displaCy

A first systematic evaluation with the previously mentioned gold standard (F1-score: 0.99) as well as a cursory inspection of the data indicate a high potential of LLM-based NER for historical, semi-structured texts. In particular, the comparatively easy adaptation to the text material (e.g. via free choice of labels, annotated examples) has shown to be of significant advantage in context of the use case. While pre-trained models come with a fixed set of categories, general-purpose LLMs, like GPT-3.5, offer a free choice of labels and thus allow for more fine-grained entity types and a certain degree of specialisation without having to create extensive training material. The same is true for the use of one- or few-shot approaches, which seem to work especially well in combination with inherently structured input, such as lists.

These factors might also explain why the results presented in this paper differ from what González-Gallardo et al. (2023) have observed in their investigation on ChatGPT-based NER in historical newspapers texts: The authors report F1-scores from 0.278 to 0.794, depending on the evaluated dataset and evaluation mode, and primarily emphasise the limitations posed by using the LLM. This differing outcome might be due to one or more of the following differences between the NER approach employed by González-Gallardo et al. (2023) and the one taken here: (1) ChatGPT web interface vs. API, (2) zero-shot vs. one-shot approach, (3) coarse-grained NER with high-level entity types vs. fine-grained NER with highly text-specific entity types, and (4) multiple, mainly unstructured and heterogenous newspaper text types vs. one semi-structured and homogenous newspaper text type. As González-Gallardo et al. (2023) state themselves, "the capacity of ChatGPT of identifying named entities is really dependent on the dataset and the type of entities."

In case of the arrival lists published in the *Wien[n]erisches Diarium*, a GPT-based approach has proven highly fruitful: In total, over 39,500 named entities could be identified and extracted from 1,447 texts. Besides including over 10,100 mentions of persons and over 4,200 cases of dates, this data set is especially rich on geographical entities and contains around 5,200 instances of city gates, 10,000 instances of points of departure, 9,400 instances of accommodations and 500 instances of destinations.

4.3 Spatio-temporal mapping on historical city maps

Having such an extensive data set on early modern named entities allows for various quantitative analyses. What seemed particularly promising due to the multitude of geographic entities, was mapping the data on historical city maps of Vienna. To do so, it was first neccessary to deduplicate (cf. 3), disambiguate (cf. 4) and geocode (cf. 3, 4) the identified entities, as illustrated here for places of accommodations:

(3) *7. Saullen*, *7. Säulen*, *7. Saulen. auf dem Neuenmarck*, *sieben Saüllen*, … > ‚Sieben Säulen' (16,37047294; 48,20650766)
(4) *Weissen Wolffen, weissen Wolff* (16,3771622; 48,21009885) ≠ *weißen Wolfis. Hauß, Weissenwolffischen Hauß* (16,36914529; 48,21192704)

To this end, multiple external knowledge resources were consulted, both automatically through distance measures (e.g., Levenshtein distance) and manually: (1) the platform *Wien Geschichte Wiki* (= WGW,

City of Vienna), (2) the prosopographical portal *ViecPro* (Romberg et al.), (3) a house directory published by Joseph Anton von Trattner (1773),[8] and (4) a house sign register from the year 1795.[9]

In addition to reusing existing digital resources, a new gazetteer was created on basis of the so-called Steinhausenplan, a city map of Vienna created by Werner Arnold Steinhausen in 1710. The plan, which has previously been georeferenced by the City of Vienna, is characterised by its high accuracy and a large number of labels. To make this data automatically processable, all labels found on the historic map were transcribed and categorised in regard to their readability. Moreover, as far as possible, each location was supplemented with information from the book *Schatz / Schutz / und Schantz Deß Ertz=Herzogthumbs Oesterreich / Das ist Ein sehr genaue / und ordentliche Beschreibung […] der berümten Haubt= und Kayserl. Residentz=Statt Wienn […]*[10] (1701), written by the imperial mail carrier Johann Jordan to allow for better navigation inside early modern Vienna. Combined, the resulting gazetteer consists of 1,469 data points, with 957 locations (65 %) including information from both the Steinhausenplan and Johann Jordan's house register.

Utilising both the external and internal resources mentioned, it was so far possible to precisely locate 71 % of all mentions of accommodations and 55 % of all distinct accommodation places.[11] Furthermore, 2 % respectively 13 % of all buildings used for lodging could be located approximately, i.e. linked to a certain street, square or neighboring house. These merged success rates of 73 % and 68 % result in almost 7.000 spatio-temporal data points, which were visualised and analysed with the open-source software QGIS, as demonstrated in Figure 6:

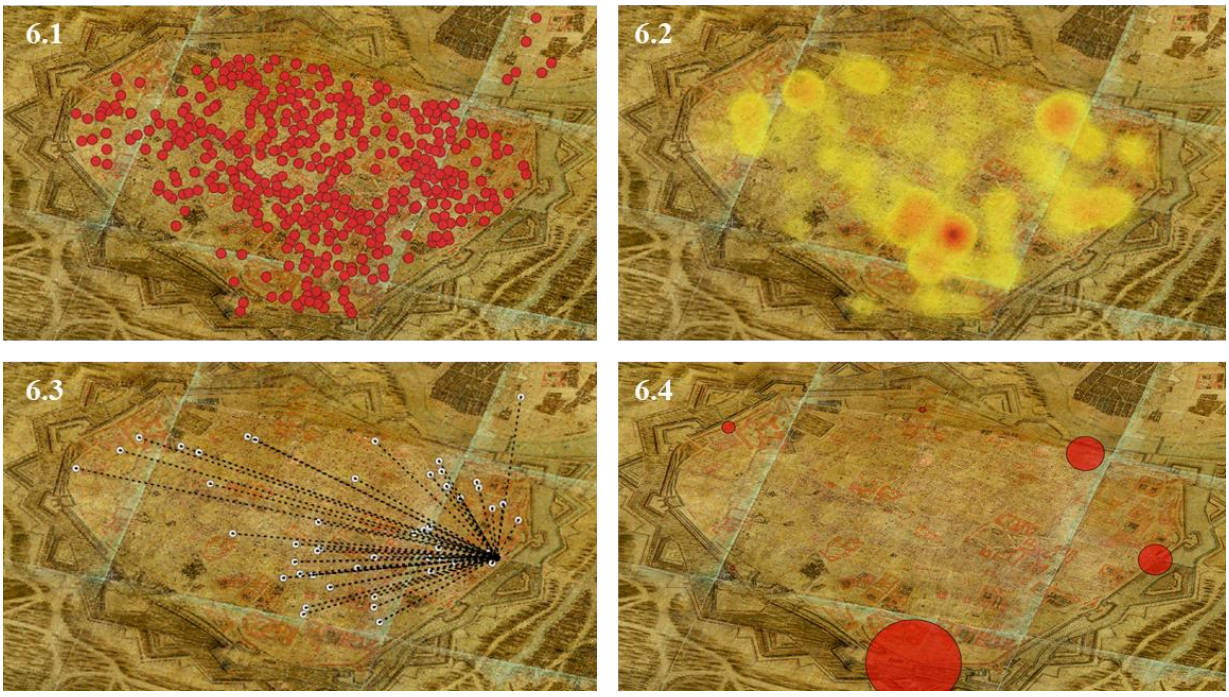


Figure 6. Four exemplary visualisations of arrival list data on the Steinhausenplan (1710) with QGIS

Like the exemplary visualisations in Figure 6 illustrate, mapping early modern arrival lists enables various insights into the urban history of a city, in this case of Vienna. It becomes possible, for example, to understand in which parts of the city (high-ranking) individuals found accommodations (cf. 6.1),

[8] *Verzeichniß der in der k.k. Haupt= und Residenz=Stadt Wien befindlichen Gassen Hausinnhabern deren Schildern und numerirten Häusern* 'Directory of the Alleys, House Owners, Their Signs, and Numbered Houses in the Imperial and Royal Capital and Residence City of Vienna'

[9] *Wiener Schildregister, oder Anweisung, wie man sich auf der Stelle helfen kann, wenn man in Wien den Schild eines Hauses oder eines Kaufmannsgewölbes in und vor der Stadt suchen, und ihn finden will* 'Viennese Sign Register, or Instructions on How to Immediately Assist Oneself in Locating and Finding the Sign of a House or a Merchant's Vault in and Around the City of Vienna'

[10] 'Treasure / Protection / and Bulwark of the Archduchy of Austria / that is a very precise and orderly Description […] of the famous Capital and Imperial Residence City Vienna […]'

[11] In the second percentage, cases requiring precise identification of an arriving individual to determine their lodging are not included. This applies, for example, when someone stays with a relative such as father, mother, or brother-in-law.

which inns or private houses were particularly popular (cf. 'hot spots' on the heatmap of 6.2), where individuals went after entering through a specific gate (cf. visitors entering through Stubentor from August 1719 to August 1720 in 6.3), and how city gates differed in their frequency of use (cf. 6.4). These findings exemplify the broader utility of periodically published lists in historical research, as the presented workflow could potentially also be applied to other cities or other types of lists. Moreover, they demonstrate that including a 'small', previously undervalued text type can offer new perspectives and foster the development of innovative research questions and workflows.

5 Conclusion: lists in historical newspapers as 'small' texts with great potential

The paper has showcased a variety of potentials that lists found in historical newspapers have to offer for (Digital) Humanities research: they constitute a frequent, both spatially and temporally distributed phenomenon, enable diachronic analyses due to their periodicity, and are relevant for various disciplines due to their manifold forms and contents. Adding to this, they showcase a high density of named entities, a repetitiveness that promises to be useful for the study of language change, and an inherent semi-structuredness that forms an optimal prerequisite for automatic annotation approaches. In summary, periodically published lists prove to be 'small' texts with great potential and a valuable data treasure for the (Digital) Humanities.

To lift this treasure trove to some extent, the doctoral project aims, parallel to its threefold approach (cf. Chapters 2–4) and in alignment with CLARIN's mission of accessability, to make the developed resources as publicly available as possible. Specifically, the created datasets (e.g., facsimiles, full texts, NER data of arrival lists; gazetteer for early modern Vienna; analysis corpus) should be (re-)integrated into the CLARIN infrastructure by archiving them in the long-term repository and CLARIN-B centre ARCHE (ACDH-CH). Additionally, it is planned to publish parts of the developed workflow, such as the scripts created for NER, to facilitate future adaptation and reuse for other (semi-structured) historical texts.

Moreover, the open-ended 'list of lists', which is currently being compiled and should, among other things, include information on a list's title, its publishing medium and its digitisation status, may also serve as a valuable resource. In the end, periodically published lists in historical newspapers are not limited to text linguistic purposes, but may, depending on their content, also be of great interest for researchers in other fields, such as onomastics (e.g., toponyms, family names in lists), lexicology and orthography (e.g., repetition of the same lemmas over time), social and medical history (e.g., arrival lists, death lists), prosopography (lists of persons in general), economics (e.g., price lists), or theatre studies (e.g., lists of theatre performances). A publicly accessible 'list of lists' could thus serve as a central reference point for interested scholars and would certainly further contribute to the overall vision of the doctoral project: by empirically investigating periodically published lists in historical newspapers on a large scale, it aims to deepen our understanding of the underrepresented 'small' texts as well as to make them both (more) visible and (re-)usable in Digital Humanities research.

Acknowledgements

The author would like to thank the three anonymous reviewers for their valuable comments and suggestions, PD Mag. Dr. Claudia Resch for her continued support as supervisor of the doctoral project, and Thomas Kirchmair, Stefanie Salzburger, Oliver Jakovljevic, Sarah Indries, and Kerstin Manninger for their helpful contributions to the project "Visiting Vienna" as student assistants or interns. The research presented was made possible through funding by the City of Vienna (MA 7 – Kultur, Wissenschafts- und Forschungsförderung, 2022–2023, *Visiting Vienna – digital approaches to the (semi-)automatic analysis of the arrival lists found in the Wien[n]erisches Diarium*) and the Austrian Academy of Sciences (ÖAW DOC Fellowship Programme, 2023–2026, *Lists in historical newspapers: Challenges and potentials of digitally analysing an overlooked text type*).

References

Adelmann, R. (2021). *Listen und Rankings*. Transcript.

ARCHE – A Resource Centre for Humanities. Austrian Centre for Digital Humanities and Cultural Heritage. https://arche.acdh.oeaw.ac.at (04.04.2024).

AustriaN Newspapers Online (ANNO). Austrian National Library. https://anno.onb.ac.at (04.04.2024).
Barton, R. A., Böckling, J., Link, S., & Rüggemeier, A. (2022). *Forms of List-Making: Epistemic, Literary, and Visual Enumeration.* Palgrave Macmillan.
Beautiful Soup. https://pypi.org/project/beautifulsoup4 (04.04.2024).
Belknap, R. E. (2004). *The List. The Uses and Pleasures of Cataloguing*. Yale University Press.
Bendel, S. (1998). *Werbeanzeigen von 1622–1798: Entstehung und Entwicklung einer Textsorte*. Niemeyer.
Blome, A. (2018). Zeitungen. In: Busse, L., Enderle, W., Hohls, R., Meyer, T., Prellwitz, J., & Schuhmann, A. (eds.), *Clio Guide – Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften*. 2nd edition. DOI: 10.18452/19244
Brand, P., & Schulze, V. (eds.) (1993). *Die Zeitung: ein Handbuch*. Institut für Medienentwicklung und Kommunikation.
Bubenhofer, N. (2020). *Visuelle Linguistik: Zur Genese, Funktion und Kategorisierung von Diagrammen in der Sprachwissenschaft*. De Gruyter.
Chinca, M., Eikelmann, M., Stolz, M., & Young, C. (eds.) (2022). *Sammeln als literarische Praxis im Mittelalter und in der Frühen Neuzeit. XXVI. Anglo-German Colloquium, Ascona 2019*. Narr Francke Attempto.
von Contzen, E. (2018). Experience, Affect, and Literary Lists. *Partial Answers* 16(2), 315–327.
von Contzen, E. (2017a). Die Affordanzen der Liste. *Zeitschrift für Literaturwissenschaft und Linguistik* 47, 317–326.
von Contzen, E. (2017b). Grenzfälle des Erzählens: Die Liste als einfache Form. In: Koschorke, A. (ed.), *Komplexität und Einfachheit. DFG-Symposion 2015*. J. B. Metzler, 221–239.
von Contzen, E. (2016). The Limits of Narration: Lists and Literary History. *Style* 50(3), 241–260.
Deutsches Textarchiv (DTA). Berlin-Brandenburgische Akademie der Wissenschaften. http://www.deutschestextarchiv.de (04.04.2024).
Diederichsen, D. (2021) (ed.). *Listen to lists*. Spector Books.
digiPress. Bayerische Staatsbibliothek. https://digipress.digitale-sammlungen.de (04.04.2024).
Digitales Forum Mittel- und Osteuropa (DiFMOE). Digitales Forum Mittel- und Osteuropa e.V. https://www.difmoe.eu (04.04.2024).
Doležalová, L. (ed.) (2009). *The charm of a list: From the Sumerians to computerised data processing*. Cambridge Scholars Publishing.
Dovifat, E. (1976). *Zeitungslehre*. 2 volumes. 6th edition, ed. by J. Wilke. De Gruyter.
Ehrenheim, A. (2011). *Das Textdesign der Stellenanzeige. Linguistisch und interdisziplinär*. Peter Lang.
Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys* 56(2). https://doi.org/10.1145/3604931
Esposito, E. (2017). Organizing without Understanding: Lists in Ancient and in Digital Cultures. *Zeitschrift für Literaturwissenschaft und Linguistik* 47(3), 351–359.
Esposito, E., & Stark, D. (2019). What's Observed in a Rating? Rankings as Orientation in the Face of Uncertainty. *Theory, Culture & Society* 36(4), 3–26.
Farø, K. (2005). „Parkettfähige Deutsche sucht jantefreien Dänen". Eine kontrastive Untersuchung zu Lexik, Phraseologie und Interkultur in der Textsorte ‚Kontaktanzeige'. In: Lenk, H. E. H., & Chesterman, A. (eds.), *Pressetextsorten im Vergleich*. Olms, 261–282.
Fix, U. (2011). Leserbriefe. Die mediale Konstruktion von Diskursgemeinschaften. In: Held, G., & Stöckl, H. (eds.), *Pressetextsorten jenseits der "News": medienlinguistische Perspektiven auf journalistische Kreativität*. Peter Lang, 139–156.
Fix, U. (2008). *Texte und Textsorten – sprachliche, kommunikative und kulturelle Phänomene*. Frank & Timme.
Flair. https://github.com/flairNLP/flair (04.04.2024).
Fludernik, M. (2016). Descriptive Lists and List Descriptions. *Style* 50(3), 309–326.
Gius, E., Meister, J. C., Meister, M., Petris, M., Bruck, C., Jacke, J., Schumacher, M., Gerstorfer, D., Flüh, M., & Horstmann, J. (2022). *CATMA (6.5.0).* https://doi.org/10.5281/zenodo.6419805
González-Gallardo, C.-E., Boros, E., Girdhar, N., Hamdi, A., Moreno, J. G., & Doucet, A. (2023). Yes but.. Can ChatGPT Identify Entities in Historical Documents? https://doi.org/10.48550/arXiv.2303.17322
Goody, J. (1977). *The domestication of the savage mind*. Cambridge University Press.
Haß-Zumkehr, U. (1998). *„Wie glaubwürdige Nachrichten versichert haben". Formulierungstraditionen in Zeitungsnachrichten des 17. bis 20. Jahrhunderts*. Narr.

Hausendorf, H. (2009). Kleine Texte. Über Randerscheinungen von Textualität. *Germanistik in der Schweiz. Onlinezeitschrift der Schweizerischen Akademischen Gesellschaft für Germanistik* 6, 5–19.
Herders Conversations-Lexikon (1854–1857). Digitised version (01/23). Wörterbuchnetz des Trier Center for Digital Humanities. https://www.woerterbuchnetz.de/Herder (04.04.2024).
HIST – Archive of the Historical Corpora. Leibniz-Institut für Deutsche Sprache. https://www2.ids-mannheim.de/cosmas2 (04.04.2024).
HuggingFace. https://huggingface.co (04.04.2024).
Impresso – Media Monitoring of the Past. https://impresso-project.ch/app (04.04.2024).
Jullien, François (ed.) (2004). *Die Kunst, Listen zu erstellen*. Merve.
Kampkaspar, D. (2019). Das DIGITARIUM – Volltexterstellung und Nutzungsmöglichkeiten. *Wiener Geschichtsblätter* 74(2), 131–135.
Kurzmann, K. (1999). *Entwicklungstendenzen in der Syntax der Anzeigenwerbung, untersucht an der Zeitung "Die Presse" (1850–1998).* Master Thesis. University of Vienna.
Ledin, Per (2015). Listans och tabellens semiotik. *Sakprosa* 7(1), 1–25.
Lemke, M., & Wiedemann, G. (2015) (eds.). *Text Mining in den Sozialwissenschaften. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse*. Springer.
Lenk, H. E. H., & Chesterman, A. (eds.) (2005). *Pressetextsorten im Vergleich*. Olms.
Chronicling America – Historic American Newspapers. Library of Congress. https://www.loc.gov/collections/chronicling-america (04.04.2024).
Mader-Kratky, A., Resch, C., & Scheutz, M. (2019). Das Wien[n]erische Diarium im 18. Jahrhundert. Neue Sichtweisen auf ein Periodikum im Zeitalter der Digitalisierung. *Wiener Geschichtsblätter* 74(2), 93–113.
Mainberger, S. (2003). *Die Kunst des Aufzählens: Elemente zu einer Poetik des Enumerativen*. De Gruyter.
Mainberger, S. (2017). Exotisch – endotisch oder Georges Perec lernt von Sei Shonagon: Überlegungen zu Listen, Literatur und Ethnologie. *Zeitschrift für Literaturwissenschaft und Linguistik* 47(3), 327–350.
Mainberger, S. (2018). A List Is (Not) a List: Some (Non-Conclusive) Remarks on Lists and Literature. *Literatur in Wissenschaft und Unterricht* 1, 8–21.
Mirador. https://projectmirador.org/ (04.04.2024).
Moretti, F. (2013). *Distant Reading*. Verso.
Mueller, M. (2020). *Scalable Reading*. https://sites.northwestern.edu/scalablereading/2020/04/26/scalable-reading (04.04.2024).
Müller, C. (2016). ANNO – Der digitale Zeitungslesesaal der Österreichischen Nationalbibliothek. *Bibliothek: Forschung und Praxis* 40(1), 83–89.
Nicholson, B. (2013). The Digital Turn. Exploring the methodological possibilities of digital newspaper archives. *Media History* 19(1), 59–73.
OpenAI API. OpenAI. https://openai.com/blog/openai-api (04.04.2024).
Oxygen XML editor. Syncro Soft. https://www.oxygenxml.com (04.04.2024).
Pfefferkorn, O., Riecke, J., & Schuster, B.-M. (2017). *Die Zeitung als Medium in der neueren Sprachgeschichte. Korpora – Analyse – Wirkung*. De Gruyter.
Promptify. Promptslab. https://github.com/promptslab/Promptify (04.04.2024).
QGIS. https://www.qgis.org (04.04.2024).
Rastinger, N. C. (in print). An der Schnittstelle von Sprache und Bild: Drucktypenwechsel zwischen Antiqua und Fraktur in historischen Periodika. In: Rammerstorfer, L., & Stiemer, H. (eds.): *Re-Writing the Past in the Digital Age: Perspektiven und Methoden der Journalforschung*. De Gruyter.
Resch, C., & Kampkaspar, D. (eds.). *Wien[n]erisches DIGITARIUM*. https://digitarium.acdh.oeaw.ac.at (04.04.2024).
Resch, C., & Kampkaspar, D. (2020). *German Fraktur 18th Century – WrDiarium_M9*. https://readcoop.eu/de/modelle/german-fraktur-18th-century/
Romberg, M., Kaiser, M., & Schlögl, M. (eds.). *ViecPro*. https://viecpro.oeaw.ac.at (04.04.2024).
Rüggemeier, A. (2020). Life Writing and the Poetics of List-Making: On the Manifestations, Effects, and Possible Uses of Lists in Life Writing. *Auto/Biography Studies* 36, 183–194.
Schaffrick, M., & Werber, N. (2017). Die Liste, paradigmatisch. *Zeitschrift für Literaturwissenschaft und Linguistik* 47, 303–316.
SpaCy. https://spacy.io (04.04.2024).

Stäcker, T., Müller, M., & Kampkaspar, D. (eds.). *Eine Darmstädter Zeitung in drei Jahrhunderten – Digitalisierung des Darmstädter Tagblatts (1740–1986)*. https://www.ulb.tu-darmstadt.de/forschen_publizieren/forschen/darmstaedter_tagblatt.de.jsp (04.04.2024).
Straßner, E. (2000). *Journalistische Texte*. Niemeyer.
Tankard, P. (2006). Reading lists. *Prose Studies* 28(3), 337–360.
Temmerman, M., & Vandenabelee, D. (2018). Listicles: A journalistic genre for digital media. *Language and the New (Instant) Media*, 149–158.
Teßmann digital. Landesbibliothek Dr. Friedrich Teßmann. https://digital.tessmann.it (04.04.2024).
Torget, A. J. (2023). Mapping Texts: Examining the Effects of OCR Noise on Historical Newspaper Collections. In: Bunout, E., Ehrmann, M., & Clavert, Frédéric (eds.), *Digitised Newspapers – A New Eldorado for Historians? Reflections on Tools, Methods and Epistemology*. De Gruyter, 47–66.
Transkribus. READ-COOP. https://www.transkribus.org (04.04.2024).
von Trattner, J. A. (ed.) (1773). *Verzeichniß der in der k.k. Haupt= und Residenz=Stadt Wien befindlichen Gassen Hausinnhabern deren Schildern und numerirten Häusern*. Wienbibliothek im Rathaus. https://resolver.obvsg.at/urn:nbn:at:AT-WBR-8402
Universal Viewer. https://universalviewer.io (04.04.2024).
Vauth, M., Meister, M., Hatzel, H. O., Gerstorfer, D., & Gius, E. (2022). *GitMA (1.4.9)*. https://doi.org/10.5281/zenodo.6330464
Verlag der Expedition des Wienerjournals (ed.) (est. 1795). *Wiener Schildregister, oder Anweisung, wie man sich auf der Stelle helfen kann, wenn man in Wien den Schild eines Hauses oder eines Kaufmannsgewölbes in und vor der Stadt suchen, und ihn finden will*. Wienbibliothek im Rathaus. https://resolver.obvsg.at/urn:nbn:at:AT-WBR-8761
Vismann, C. (2000). *Akten. Medientechnik und Recht*. Fischer.
Waldispühl, M. (2019). Die Liste als Ordnungsmedium im mittelalterlichen Libri vitae. *Zeitschrift für Literaturwissenschaft und Linguistik* 49(2), 197–218.
Wien Geschichte Wiki. City of Vienna. https://www.geschichtewiki.wien.gv.at (04.04.2024).
Zeit.punktNRW. Universitäts- und Landesbibliothek Bonn, Universitäts- und Landesbibliothek Münster, Universitäts- und Landesbibliothek Düsseldorf, Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen. https://zeitpunkt.nrw (04.04.2024).
Zeitungsinformationssystem (ZEFYS). Staatsbibliothek zu Berlin. https://zefys.staatsbibliothek-berlin.de (04.04.2024).

Domain-Specific Languages for Epigraphy: the Case of ItAnt

Federico Boschetti
CNR-ILC
URT Venezia, Italy
federico.boschetti@ilc.cnr.it

Luca Rigobianco
Dipartimento di Studi Umanistici
Università Ca' Foscari Venezia
Venezia, Italy
luca.rigobianco@unive.it

Valeria Quochi
CNR-ILC
Pisa, Italy
valeria.quochi@ilc.cnr.it

Abstract

This contribution illustrates how the definition of a Domain-Specific Language can support the activities of epigraphists and historical linguists. It presents and discusses a method and technological solution, based on Domain-Specific Languages, for facilitating scholars in digitally representing the available knowledge of archaic languages and cultures. This is achieved by increasing the human readability of the encoded data without sacrificing compliance with standard models and formats. The work is framed within the context of an Italian National collaborative research project devoted to the study of the languages and cultures of ancient Italy. The platform developed within this project offers an interesting use case and motivation for experimenting with Domain-Specific Languages for the creation of necessary digital critical editions of the inscriptions relevant for these languages. After explaining the definition process of the DSL grammar, we finally test the applicability of the DSL grammar to five example inscriptions in the Faliscan language.

1 Introduction

The recovery, digitisation, and sharing of knowledge relating to ancient fragmentary languages and their cultures are primary concerns within the fields of historical linguistics and digital humanities, posing significant challenges. Fragmentary languages are dead languages attested through a highly restricted corpus of surviving texts. Such a corpus is limited due to socio-cultural choices on what to write as well as the randomness of the documentary findings. Due to these restrictions, the knowledge that can be derived is necessarily partial and sometimes uncertain, both in terms of grammar and lexicon, and with regard to language variation over time and space, along the social ladder, and according to the communicative situation. Such an incomplete, uncertain, and quantitatively scarce written evidence hampers the use of state-of-the-art AI or machine learning techniques and requires an adaptation of existing language technology. This can only be achieved through the collaboration between historical linguists and language technologists.

The very first and fundamental stage in this direction is the creation of robust, machine-actionable digital scholarly editions of the inscriptions and their linguistic content, a task what is by all means non-trivial. Recently, the ILA project (Sarullo, 2016) has taken a first step towards adapting the XML-TEI/EpiDoc standard model to an epigraphically attested fragmentary language such as archaic Latin (7th-5th century BC). Additionally, the i.Sicily digital corpus (Prag & Chartrand, 2019) deserves to be mentioned, as it collects texts from ancient Sicily dating from the 7th century BC to the 7th century AD, including fragmentary languages such as Sikel and Elymian.

In general, despite the considerable effort required, the challenge of adequately treating these languages digitally must be faced. This is necessary in order to preserve their documentation and knowledge and make them widely accessible. However, digitising scholarly editions proves to be a time-consuming and unfriendly task for many scholars. This contribution tackles this issue and introduces a method and technological solution based on Domain-Specific Languages (DSLs hereafter) to facilitate such tasks.

The paper is organised as follows: Section 2 describes the project that motivated this work and the online platform that will consume the produced critical editions, making them available to scholars for

Federico Boschetti, Luca Rigobianco and Valeria Quochi 2023. Domain-Specific Languages for Epigraphy: the Case of ItAnt. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 191–202. DOI: https://doi.org/10.3384/ecp210

creating interlinked lexical resources. Here the connection to the CLARIN infrastructure is also made explicit. Section 3 gives a brief introduction to DSLs and their advantages in Digital Humanities (DH) contexts before delving into the specific grammar designed for the languages of ancient Italy. Section 4 then demonstrates the applicability of the DSL grammar to five example inscriptions in the Faliscan language. Finally, Section 5, wraps up and suggests possible future directions.

2 The context: the ItAnt project

The project *Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models* (ItAnt hereafter) is a collaborative initiative funded by the Italian Ministry of University and Research. It aims to investigate the languages of ancient Italy by combining the methods of historical linguistics with digital technologies specifically designed to create a set of interconnected resources, particularly critical digital editions of inscriptions, lexica, and bibliographies[1].

With the sole exception of Roman Latin, the languages of ancient Italy (8th century BC-1st century AD) are fragmentary languages. Their evidence consists almost exclusively of epigraphic texts, which often present problems relating to the reading, segmentation into words, linguistic analysis, and interpretation. Therefore, one of the key challenges of the ItAnt project is to adapt existing digital methods and tools, practices, and methodologies of digital epigraphy and computational lexicography to the highly fragmentary nature of such documentation. Among the languages of ancient Italy, the project focuses on Oscan, Faliscan, Venetic, and Cisalpine Celtic, chosen as representative due to the quantitative and qualitative differences in their documentation and to their belonging to linguistic (sub)groups which are diverse as regards their genetic classification (Poccetti, 2017).

The main objectives of the project are thus to create and interlink a digital archive of critical editions of inscriptions, a multilingual computational lexicon, and a bibliographic dataset of relevant cited works in FRBRoo[2]/LRMoo[3]. Within the digital archive, the inscriptions are encoded in XML following the XML-TEI/EpiDoc model and schema[4]. Furthermore, these new editions of the inscriptions will be enriched with metadata defined in shared common vocabularies, enabling accurate semantic description of them as both linguistic and material objects. The DSL solution described in this paper is responsible for producing these enriched EpiDoc editions of the inscriptions, which will then be ingested by the DigItAnt platform for linking to the other resources mentioned above (see section 2.1 below for details on the platform).

2.1 The DigItAnt platform

Together with the production and publication of datasets for the four languages in focus, one of the main outcomes of ItAnt is a web platform for creating and then exploring the interlinked ecosystem of resources mentioned above: LOD-compliant lexica, critical editions of inscriptions, citations and bibliographic references, as well as other external available salient vocabularies and lexicons.

Assuming that more intuitive disciplinary editing tools can simplify the work of philologists and historical linguists in managing lexical and linguistic knowledge about ancient languages, the DigItAnt platform is designed to assist scholars in encoding lexical information of ancient languages and linking it to other relevant (re-)sources according to semantic web principles. Lexicon creation lies at the heart of the editing platform, which further enables scholars to enrich lexica with actionable links to related inscription, cited bibliographic items, and to other external datasets. Particularly central to the platform is the linking of lexical and morphological forms to their attestations in the texts encoded in XML-TEI/EpiDoc digital scholarly editions of relevant inscriptions[5].

Digital critical editions of inscriptions, while vital for scholars to consult, play a supporting role in the current version of the editing platform. They are considered instrumental because the platform, in

[1] For more information see the project website https://www.prin-italia-antica.unifi.it/

[2] https://cidoc-crm.org/frbroo/

[3] LRMoo is the new ontological bibliographic model developed from FRBRoo. A stable version was released only in October 2023.

[4] http://www.stoa.org/epidoc/gl/latest/

[5] See Quochi, Bellandi, Mallia, et al. (2022) and Quochi, Bellandi, Khan, et al. (2022) for additional details on the platform.

its current form, offers tools allowing only for the creation and updating of lexicons and the linking of lexical items to external resources, particularly for linking to inscriptions to describe their attestations. As a result, the platform expects inscriptions to be encoded independently in XML according to the XML-TEI/EpiDoc format, the de facto standard for digital epigraphic projects. Consequently, digital editions of inscriptions are considered external datasets, i.e., prepared separately, which the platform can ingest.

Within the ItAnt project, however, we have experimented with the use of a DSL, described in this article, as an alternative to the commonly used Oxygen XML editor. This DSL system offers scholars a lighter and more intuitive way to produce their digital editions. Thus, editions encoded with ItAntDSL, and then converted to EpiDoc XML as described in section 3, can later be ingested by the platform for linking and exploration purposes. In detail, the basic workflow involves historical linguists uploading one or more EpiDoc XML documents to the platform and operate to link specific text segments to either existing lexical items or newly created ones (as exemplified in Fig. 1, see also Quochi, Bellandi, Khan, et al. (2022)).

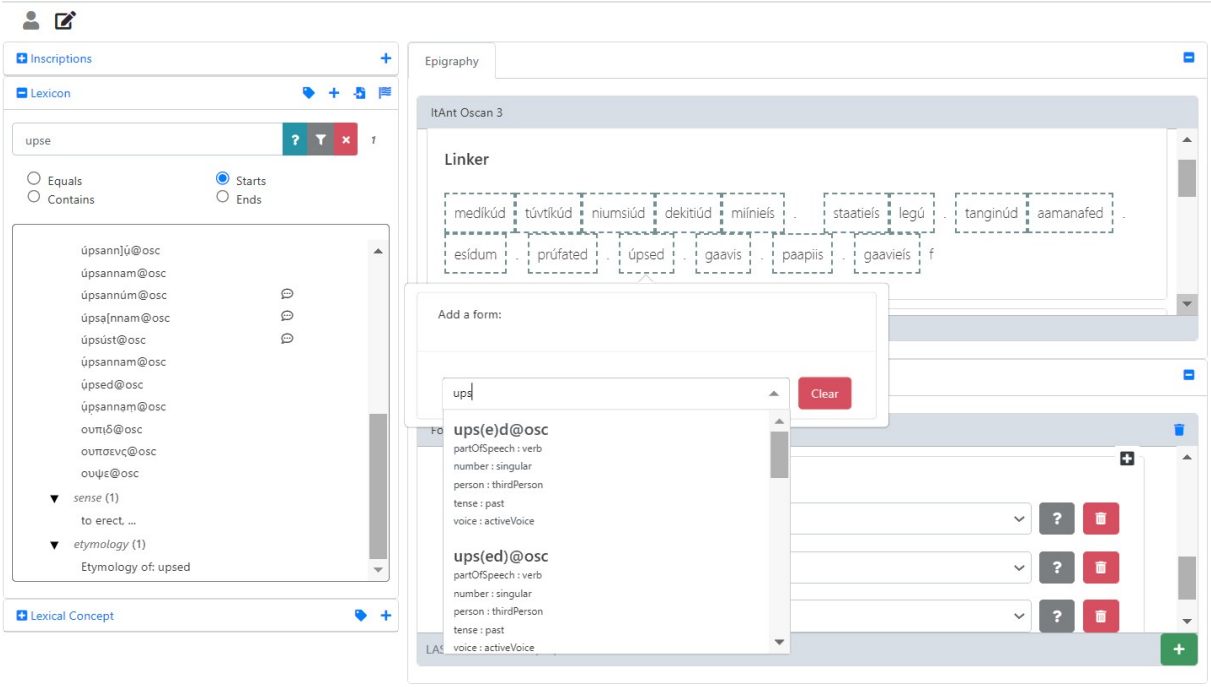


Figure 1: The DigItAnt editor: text - lexicon linking

Thanks to the EpiDoc XML encoding, a visualisation of contextual information as well as of the text according to the Leiden conventions[6] is also possible, both in the editor (EpiLexO) and in the exploration platform (DigItAnt-search), as shown in Fig. 2[7].

2.2 The Digital Edition of the Inscriptions

As mentioned above, the project envisages the inscriptions being encoded according to the XML-TEI/EpiDoc schema. Such a schema is the result of an international effort aimed at customising the Text Encoding Initiative's standard for the representation of epigraphic documents according to the Leiden Conventions. In particular, XML-TEI/EpiDoc provides markup for the text (edition, apparatus, translation, commentary, bibliography) as well as the materiality and history of the object on which the text appears (repository, support, layout, hand, place and date of origin, provenance).

[6]The Leiden conventions refer to the system of diacritics used in the scholarly edition of ancient texts, including epigraphic texts; in this regard, refer to Krummrey and Panciera (1980). Since the actual use of these conventions shows a certain variability, for the DSL we have specifically relied on the Leiden+ system, see https://papyri.info/docs/leiden_plus.

[7]The DigItAnt platform prototype is ready and already in use within ItAnt. It's an open-source code available at https://github.com/DigItAnt. It is currently maintained and continues to be improved with new functionalities.

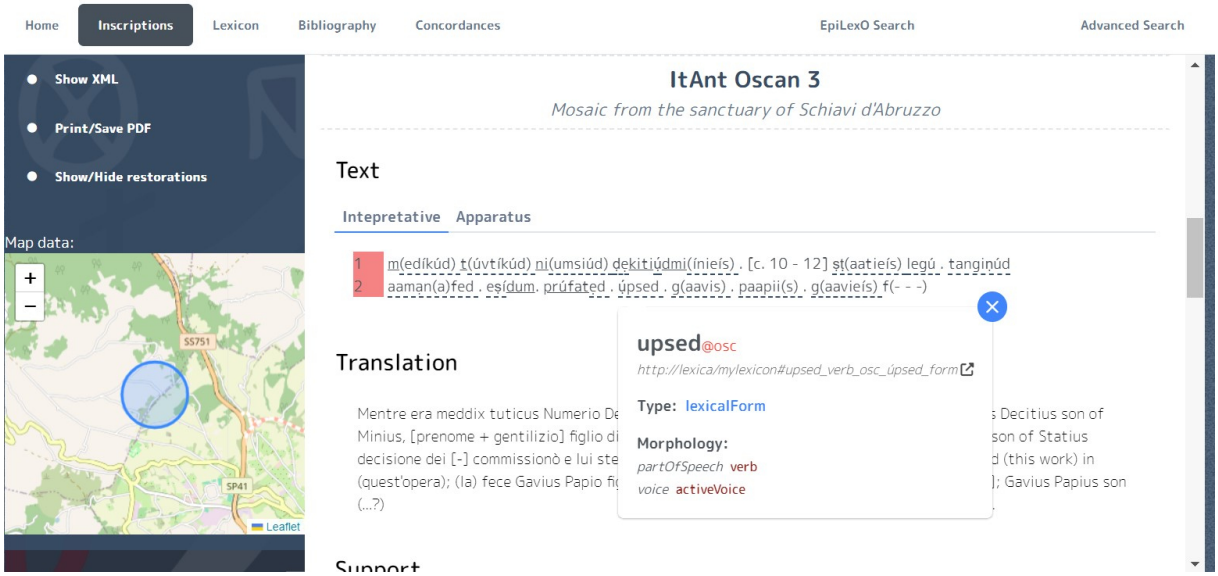


Figure 2: Text-lexicon data mesh-up in the Explorationplatform

Furthermore, thanks to the extensibility of XML and the versatility of XML-TEI/EpiDoc, ItAnt has proposed solutions for managing specific issues arising from the nature of the languages of ancient Italy as fragmentary languages and their specific epigraphic features (Murano et al., 2023). The customisation has mainly consisted of adding tags to the standard XML-TEI/EpiDoc set. Specifically, within the `<scriptNote>` element, we have opted to specify through `@type` attributes of a `<rs>` element the word division, assuming '*scriptio continua*', '*punctuation*', '*blank spaces*', and '*mixed*' as possible values, as well as the application and simplification of syllabic punctuation for the Venetic inscriptions [8]. Moreover, `<tei:rs>` elements have been added within the `<tei:support>` element to specify the object shape and the possible reuse of the support. A `<rs>` element have also been added within the `<layout>` element to specify if the inscription is opisthographic.

Additionally, a major problem with the XML-TEI/EpiDoc guidelines concerns the theoretical need for a clear distinction between a language and the system(s) used for writing it, since languages and scripts should be indicated together through a single `@ident` attribute of the `<tei:language>` element within the TEI header. For both overcoming such an issue and ensuring interoperability with other digital corpora, we have still followed the guidelines, but we have specified the script(s) regardless of the language(s) through a `<tei:rs>` element within the `<tei:scriptNote>` element, as well as the language(s) regardless of the script(s) through a further `<tei:language>` element. From a linguistic point of view, we have chosen to explicitly mark up the words through `<tei:w>` and `<tei:name>` elements, in order to make it possible to link them to the entries of the computational lexicon. In particular, each word is uniquely identified through an `@xml:id` attribute, whereas the provisioning of lexical information is dealt with in the companion lexicon. The identifier is built using information about the language, the line number, and the position of the word in the line, so as to be transparent and easily readable also by scholars. For example, the value `Fal_6_l_1_w_2` stands for 'second word of the first line of the sixth Faliscan inscription of the ItAnt collection'. The `<tei:name>` elements are further specified through a `@type` attribute as '*praenomen*', '*gentilicium*', '*patronymic*', etc. Furthermore, the use of a `@ref` attribute makes it possible to clearly identify onomastic formulas even in the case of a syntactic break between their components or of a component shared by two or more formulas. The onomastic formulas are then resumed in the commentary (`<tei:div type="commentary">`) through the `<listPerson>` element.

[8] On the peculiar Venetic syllabic punctuation see Marinetti (2020)

With the goal of data integration, ItAnt makes use of widely used vocabularies and gazetteers, in particular *The Art and Architecture Thesaurus* provided by *The Getty Research Institute* is used for specifying object type, material, and writing technique[9], the EAGLE vocabulary for the type of inscriptions (dedicatory, funerary, etc.)[10], and Pleiades and GeoNames for ancient and modern names respectively[11]. In addition, Trismegistos IDs are used, when available, to identify the texts[12] and bibliographical records are also linked through a specific library built up by using Zotero[13].

2.3 Relation with CLARIN

Part of the mission of the ItAnt project is to contribute and integrate data and tools into European Research Infrastructures for the Humanities and Social Sciences, particularly within CLARIN (Common Language Resources and Technology Infrastructure). Since its start, ItAnt has been a project of interest for CLARIN-IT, also due to its potential contribution to the involvement of the community of historical linguists. At the end of the project, the hosting of the platform will transition to the ILC4CLARIN center, where it will be offered as a sustainable open service. Not only will the platform be preserved, but all data and software components will also be deposited and accessible in the long term through CLARIN, for documentation and re-use. The ILC4CLARIN repository[14] already stores copies of LexO-server (Bellandi, 2019), EpiLexO (Mallia et al., 2023) and ItAntDSL (Boschetti & Rigobianco, 2023). At the conclusion of the project, the inscription corpora, lexicons and bibliographies will also be deposited, making them easily discoverable and consumable via CLARIN channels.

Furthermore, due to ItAnt's focus on outputting Linguistic Linked Open Data (LLOD) compatible versions of the data, it will contribute to the development of a CLARIN(-IT) LLOD Platform. In this regard, DigItAnt is a candidate use case for one of the pilot projects to be developed in the context of a recently started large-scale Italian infrastructural initiative, the Humanities and Heritage Italian Open Science Cloud (H2IOSC)[15].

3 Domain-Specific Languages for the encoding of fragmentary archaic languages

3.1 Domain-Specific Languages

Domain-Specific Languages (DSLs) are programming or markup languages created specifically for a certain area of interest. Unlike general-purpose programming languages, which are made to handle a wide variety of programming tasks, DSLs are optimised for a specific field. They aim to provide more expressive power, simplicity, and efficiency for those specific areas. The main benefit of a DSL is its ability to let users describe concepts and actions in ways that closely match the specific abstractions of that domain.

DSLs in the domain of digital epigraphy provide scholars with a set of specialised tools for describing the structure and semantics of inscriptions, enabling precise and detailed digital representations of them. Furthermore, this may enhance the accessibility and dissemination of inscriptions in digital formats. Thanks to DSLs, digital epigraphists can more effectively engage with the textual data, automate repetitive tasks, and focus on the nuanced interpretation and study of the inscriptions.

3.2 How (ItAnt)DSL Facilitates the Encoding

Encoding epigraphic contextual metadata and textual data in XML-TEI/EpiDoc is a complex, error-prone task. Indeed, XML-TEI is quite verbose (because element names, attributes and values must be written in full) and redundant (because opening and closing tags repeat the element names). The percentage

[9] https://www.getty.edu/research/tools/vocabularies/aat/. Among other concepts, the AAT taxonomy defines useful terms for describing physical cultural objects such as materials (e.g., pottery, bronze, . . .) , object types (e.g., bowl, stele, . . .), and writing techniques (e.g., engraving, inscribing, . . .) . Additionally, the iDAI.thesauri provided by the Deutsches Archäologisches Institut (http://thesauri.dainst.org/de.html) is used as a supplement with regard to natural supports such as cliffs.

[10] https://www.eagle-network.eu/resources/vocabularies/typeins/

[11] https://pleiades.stoa.org/; https://www.geonames.org/

[12] https://www.trismegistos.org/tm/

[13] https://www.zotero.org/groups/2552746/

[14] https://ilc4clarin.ilc.cnr.it/

[15] https://www.h2iosc.cnr.it/home/

of informative and structural contents is unbalanced. XML-TEI ensures data interchange among software applications and promotes machine actionability and interpretability, but human readability of an encoded document decreases rapidly as complexity increases.

In ItAnt linguistic, philological and prosopographical data are highly entangled. Each word is associated to its part of speech, conjectural integrations to textual gaps (lacunae) are recorded, and named entities are identified. These chunks of information often overlap: for instance a lacuna in a line of text may extend between the end of the third token and the beginning of the fourth one, whereas a named entity defined by *praenomen* (partially conjectured), *gentilicium* and *patronymicus* may extend from the forth to the sixth token.

The problem of overlapping hierarchies in TEI is well-known and many solutions are available, both through manual encoding of stand-off annotations in XML (Spadini & Turska, 2019) and through alternative representations (e.g. in JSON), currently or planned to be convertible in XML-TEI (Neill & Schmidt, 2021). An experimental solution adopted in ItAnt for encoding part of the corpus, is based on a domain-driven approach, which involves the epigraphists to co-design a Domain-Specific Language (Parr, 2009), named ItAntDSL, to encode data and metadata.

The aims of this approach are twofold: a) optimising the encoding process and the encoded documents according to six dimensions (familiarity, transparency, completeness, compactness, consistency, and actionability (Zenzaro et al., 2022) and b) complying with the EpiDoc abstract model. With regard to the above mentioned dimensions, familiarity refers to the maintenance of the scholar's work habits and transparency indicates the level of cognitive effort and/or technical training required of the scholar. Completeness refers to the amount of information which may be expressed, while the ratio between completeness and formalisation is what is meant by compactness. In particular, what occurs more frequently is expected to be encoded with a smaller number of characters. Consistency assesses the coherence in describing the same phenomena in the same way, implying that the representation of the same type of information is unique and therefore unambiguous.

Finally, the ability to extract or deduce information from data is referred to as actionability, which is an intrinsic characteristic in formal languages described by a grammar and commonly accompanied by other components for code processing such as a lexer and a parser. It is evident that a DSL allows for a greater degree of familiarity, transparency, and compactness than an XML encoding. Specifically, once the DSL has been suitably designed by researchers in close contact with experts in the field in question (in our case the fragmentarily attested languages of ancient Italy), it may also be used by scholars who do not know XML nor the XML-TEI/EpiDoc standard, thus drastically reducing the training time necessary to proceed with text encoding. Furthermore, the encoding of contextual metadata (Fig. 3) and textual data (Fig. 4) is very compact. From the user's perspective, this guarantees greater readability and, therefore, the possibility of keeping the text under control, significantly reducing the risk of errors or omissions. In this regard it should also be noted that, although a DSL in itself provides less control over text insertion, the use of an editor may help the scholar by signalling syntactic errors, providing suggestions for their resolution as well as self-completion.

ItAntDSL is defined by a Context-Free Grammar (CFG) available on GitHub[16]. The documents encoded in ItAntDSL are then parsed by ANTLR (Parr, 2013), which first converts the Domain-Specific Language into XML with a proprietary schema (XML-ItAnt), based on the production rules of the CFG.

Then, a chain of XQuery scripts and XSLT stylesheets transforms XML-ItAnt documents into XML-TEI/EpiDoc documents. The transformations are not limited to the translation of element names and to structural modifications, but extend to the integration of a) automatically generated IDs; b) default values omitted in ItAntDSL documents; c) expansion of complex structured data encoded in ItAntDSL documents by reference (between quotation marks) and retrieved from the XML documents stored in an eXist-db[17]. A sample of the final result is shown in Fig. 5.

As already mentioned, the provision of lexical information is supplied in the companion lexicon and therefore, as far as lexical aspects are concerned, ItAntDSL is limited to explicitly mark up lexical items

[16]https://github.com/CoPhi/itantdsl/

[17]https://exist-db.org is a versatile native-XML database management system commonly used in DH projects for managing XML corpora and archives.

```
4
5  IDENTIFIERS
6  #place: "Schiavi d'Abruzzo (Chieti)"
7  #inst: "in situ (under the tutelage of the Soprintendenza Archeologia, Belle Arti e Paesaggio dell'Abruzzo)"
8  #msName: mosaic from the sanctuary of Schiavi
9  #tm: "TM_170843"
10 #trad: "ST_Sa_2" "ImIt_Teruentum_36"
11
12 SUMMARY
13 Inscription recording building and dedication of the paving from temple B of Pietrabbondante sanctuary.
14
15 SUPPORT
16 "temple floor" "tesserae (mosaic components)" #w: 350
17 #notRe-used #very_fragmentary (The inscription is damaged; reading is only possible through photographic material)
18
19 LAYOUT
20 #columns: 1 #writtenLines: 2
21 #exec: "mosaic (opus signinum)" #notOpistograph
22
23 HAND, SCRIPT, AND DECORATION
24 #palaeographicNotes: Letters measure 12 cm in height
25 #characterDimension: 12
26 #alphabet: "Oscan national alphabet"
27 #punctuation
28
```

Save

Figure 3: ItAntDSL: metadata

```
46 DIPLOMATIC EDITION
47 #face_a | #text_direction_r_to_l | #sinistrorse
48
49 1 m t ni d!e![.4]ú![.1] [.2] . [.10-12] s!t! legú . tanginúd
50 2 aama!nfed . es!í[.3] . [.6]e!d . ú!psed . g . paapi . g f
51
52
53 ***
54
55
56 INTERPRETATIVE EDITION
57 #face_a | #text_direction_r_to_l | #sinistrorse
58
59 1 * m(edíkúd) t(úvtíkúd) ni(umsiúd) d!e![kiti]ú![d] [mi](ínieís) . [10-12] s!t!(aatieís) legú . tanginúd
60 2 * aama!n(a)fed . es!í[dum] . [prúfat]e!d . ú!psed . g(aavis) . paapi(is) . g(aavieís) f()
61
62
63 #line: 1
64 1 m(edíkúd) = #word
65 2 t(úvtíkúd) = #word
66 3 ni(umsiúd) = #praenomen
67 4 d!e![kiti]ú![d] = #gentilicium
68 5 [mi](ínieís) = #patronymic
69 3;4;5 = @p1
70 6 . = #pc_word
```

Save Delete

Figure 4: ItAntDSL: textual data

as either words or names, uniquely identify them with an `@xml:id` attribute and, in the case of names, further specify them according to an appropriate taxonomy ('*praenomen*', '*gentilicium*', '*patronymic*', etc.).

4 Application of ItAnt DSL to Faliscan

A linguistics graduate, proficient in epigraphy of the fragmentary languages of ancient Italy but with only basic skills in DH and particularly in text encoding, was selected to collaborate in the testing phase. Specifically, she was entrusted with five Faliscan inscriptions and tasked with encoding them both in

```
<tei:div type="edition" subtype="interpretative" xml:space="preserve">
  <tei:div type="textpart" n="face_a" style="text-direction:r-to-l" rend="ductus:sinistrorse">
    <tei:ab>
      <tei:lb n="1" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1"/>
        <tei:w xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_1">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">m</tei:supplied></tei:abbr><tei:ex>edíkúd</tei:ex></tei:expan>
        </tei:w>
        <tei:w xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_2">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">t</tei:supplied></tei:abbr><tei:ex>úvtíkúd</tei:ex></tei:expan>
        </tei:w>
        <tei:name type="praenomen" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_3" ref="#p1">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">ni</tei:supplied></tei:abbr><tei:ex>umsiúd</tei:ex></tei:expan>
        </tei:name>
        <tei:name type="gentilicium" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_4" ref="#p1">
          <tei:unclear>de</tei:unclear>
          <tei:supplied reason="lost" evidence="previouseditor">kiti</tei:supplied>
          <tei:unclear>ú</tei:unclear><tei:supplied reason="lost" evidence="previouseditor">d</tei:supplied>
        </tei:name>
        <tei:name type="patronymic" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_5" ref="#p1">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">mi</tei:supplied></tei:abbr><tei:ex>íníeís</tei:ex></tei:expan>
        </tei:name>
        <tei:pc unit="word">.</tei:pc>
        <!-- ... -->
```

Figure 5: XML-TEI/EpiDoc

XML-TEI/EpiDoc and through ItAntDSL. Although the case study lacks scientific relevance, it nonetheless provided interesting qualitative insights. The results can be summarized as follows: the time required for training was significantly shorter for learning the DSL compared to learning XML-TEI encoding; the time needed for encoding was markedly lower; documents produced via ItAntDSL are approximately three times more compact than EpiDoc documents.

```
LANGUAGE
#l1: “Faliscan” (“Faliscan in Faliscan alphabet”)
```

Figure 6: Fragment of ItAntDSL metadata for a Faliscan inscription

During the text encoding phase, the need arose to introduce extensions to the original grammar, according to the planned workflow. The encoded documents (in Fig. 6 it is possible to see a couple of lines about the language of the inscriptions), processed by the ItAntDSL parser that generates an Abstract Syntax Tree (Fig. 7), produces XML files with a proprietary schema, as depicted in Fig. 8.

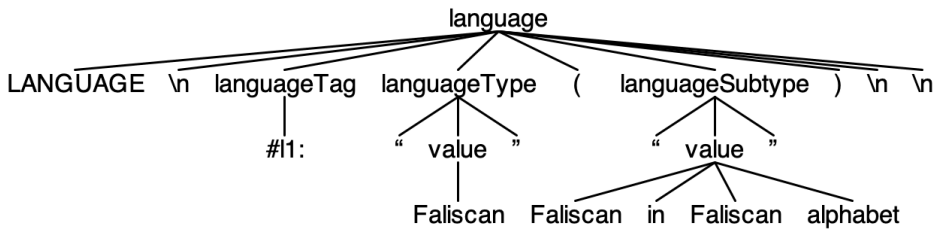


Figure 7: Example of Abstract Syntax Tree for ItAntDSL

To keep the language productions as compact as possible, shared information, which is typically repeated in the XML-TEI/EpiDoc files, is stored in a separated file. Fig. 9 shows a fragment of information related to the languages studied in this project, encoded in YAML and easily convertible to XML.

The XSLT transformation merges data from the XML file with proprietary schema with data from the YAML file converted to XML with a proprietary schema.

```
<language>
 LANGUAGE
 <languageTag>#l1:</languageTag>
 <languageType>“<value>Faliscan</value>”</languageType>
 (<languageSubtype> “<value>Faliscan in Faliscan alphabet</value>”</languageSubtype>)
</language>
```

Figure 8: XML encoding with proprietary schema

```
list:
  type: "language"
  items:
   - key: "Faliscan"
     values:
      - type: "ident"
        content: "xfa"
      - type: "source"
        content: "https://iso639-3.sil.org/code/xfa"
   - key: "Faliscan in Faliscan alphabet"
     values:
      - type: "ident"
        content: "xfa-Ital-x-xfafal"
      - type: "source"
        content: "https://www.prin-italia-antica.unifi.it/"
      - type: "ana"
        content: "https://unicode.org/iso15924/iso15924-codes.html"
```

YAML2XML

```
<list xmlns="http://itant.eu" type="language">
  <item key="Faliscan">
    <value type="ident">xfa</value>
    <value type="source">https://iso639-3.sil.org/code/xfa</value>
  </item>
  <item key="Faliscan in Faliscan alphabet">
    <value type="ident">xfa-Ital-x-xfafal</value>
    <value type="source">https://www.prin-italia-antica.unifi.it/</value>
    <value type="ana">https://unicode.org/iso15924/iso15924-codes.html</value>
  </item>
</list>
```

Figure 9: Fragment of the YAML file with look-up information

```
<xsl:template match="/dsl:start/dsl:language/dsl:languageType/dsl:value|/dsl:start/dsl:language/dsl:languageSubtype/dsl:value">
    <tei:language>
        <xsl:variable name="key" select="."/>
        <xsl:variable name="languageIdent" select="document('database.xml')/dsl:data/dsl:list/dsl:item[@key=$key]/dsl:value[@type='ident']"/>
        <xsl:variable name="languageSource" select="document('database.xml')/dsl:data/dsl:list/dsl:item[@key=$key]/dsl:value[@type='source']"/>
        <xsl:variable name="languageAna" select="document('database.xml')/dsl:data/dsl:list/dsl:item[@key=$key]/dsl:value[@type='ana']"/>
        <xsl:attribute name="ident">
            <xsl:value-of select="$languageIdent"/>
        </xsl:attribute>
        <xsl:if test="$languageSource">
            <xsl:attribute name="source">
                <xsl:value-of select="$languageSource"/>
            </xsl:attribute>
        </xsl:if>
        <xsl:if test="$languageAna">
            <xsl:attribute name="ana">
                <xsl:value-of select="$languageAna"/>
            </xsl:attribute>
        </xsl:if>
        <xsl:value-of select="$key"/>
    </tei:language>
</xsl:template>
```

Figure 10: Fragment of the XSLT file

The resulting XML-TEI/EpiDoc fragment is visible in Fig. 11.

```
<tei:langUsage>
   <tei:language ident="xfa" source="https://iso639-3.sil.org/code/xfa">Faliscan</tei:language>
   <tei:language ident="xfa-Ital-x-xfafal"
                 source="https://www.prin-italia-antica.unifi.it/"
                 ana="https://unicode.org/iso15924/iso15924-codes.html">Faliscan in Faliscan alphabet</tei:language>
</tei:langUsage>
```

Figure 11: XML-TEI/EpiDoc output

5 Conclusions

In this paper, we have presented a novel approach to addressing the challenges associated with encoding fragmentary languages of ancient Italy in digital formats, in ways that are user-friendly and meaningful within the discipline. Through the development and application of a Domain-Specific Language (DSL) called ItAntDSL, tailored to the needs of historical linguists and epigraphists, we have improved in efficiency, compactness, and ease of use compared to traditional XML-TEI/EpiDoc encoding methods.

Our experiment with encoding five Faliscan inscriptions using both ItAntDSL and XML-TEI/EpiDoc seems to reveal that the training time required for mastering the DSL is significantly shorter, while the encoding process itself is markedly more efficient. Furthermore, documents produced via ItAntDSL were approximately three times more compact than their EpiDoc counterparts, making them much more readable and accessible for humans as well as significantly reducing file size and complexity. The successful application of ItAntDSL in encoding ancient inscriptions underscores its potential as a powerful tool for digital epigraphy and historical linguistics. It advocates for an integration within platforms such as DigItAnt, thereby permitting a full editing experience from within a single online environment.

While the ItAnt project provided a valuable opportunity to develop methods and tools to facilitate the encoding activities of epigraphists, CLARIN offers not only the infrastructure to deposit research data, but also the means to disseminate and share new practices adequate to the domain of epigraphic studies. In recent years, collaborative effors within the CLARIN Knowledge Centre for Digital and Public Textual Scholarship (DiPText-KC)[18] have resulted in significant contributions to digital humanities projects. These efforts, involving the Venice Center for Digital and Public Humanities (VeDPH), the Institute for Computational Linguistics (CNR-ILC), and CLARIN-IT center ILC4CLARIN have addressed various kinds of resources, including DH projects related to collections of literary texts (Boschetti et al., 2021), and collections of epigraphic sources (Vagionakis et al., 2022).

5.1 Future works

The know-how acquired in the process of encoding and annotating inscriptions through ItAntDSL will be shared within our target disciplinary scientific community through the CLARIN Knowledge Center DiPText. Knowledge sharing may take the form of video tutorials, webinars, and/or workshops, and will thus continue also after the end of the ItAnt project.

The corpora containing the digital editions of the inscriptions will be deposited into CLARIN as soon as they are finalized. Additionally, the ItAnt DSL has already been deposited (Boschetti & Rigobianco, 2023) and it will be updated to integrate any improvements.

Regarding future improvements to the DigItAnt platform, we will explore the possibility of integrating the ItAntDSL system into the platform's interface, to allow users to create or revise editions of the inscriptions via the DSL directly within the DigItAnt web environment.

Acknowledgments

This work is supported by the Italian Ministry of the University and Research with the Italian National Strategic Research Grant PRIN 2017XJLE8J for the project: *Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models*. The project involves a consortium comprising the Ca' Foscari University of Venice, the University of Florence, and the Institute for Computational Linguistics "A. Zampolli" of the National Research Council of Italy. It also benefits from collaboration with the Venice Center for Digital and Public Humanities (VeDPH), which is part of the Department of Humanities of Ca' Foscari University of Venice (https://www.unive.it/pag/39287).
The project is also supported by and contributing to the CLARIN-IT research infrastructure for the Humanities and Social Sciences.

[18] https://diptext-kc.clarin-it.it/

References

Bellandi, A. (2019). LexO-server: REST services for linguistic linked data in OntoLex-lemon [ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa]. http://hdl.handle.net/20.500.11752/ILC-1004

Boschetti, F., Del Grosso, A. M., & Spinazzè, L. (2021). La galassia musisque deoque: Storia e prospettive. In *Paulo maiora canamus - raccolta di studi per Paolo Mastandrea* (pp. 405–419, Vol. 32). Edizioni CaFoscari. https://edizionicafoscari.unive.it/media/pdf/books/978-88-6969-558-2/978-88-6969-558-2-ch-26.pdf

Boschetti, F., & Rigobianco, L. (2023). ItAntDSL [ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa]. http://hdl.handle.net/20.500.11752/ILC-1003

Krummrey, H., & Panciera, S. (1980). Criteri di edizioni e segni diacritici. In G. Bevilacqua (Ed.), *Tituli* (Vol. 2). Edizioni di Storia e Letteratura.

Mallia, M., Bellandi, A., Tommasi, A., Zavattari, C., Bandini, M., & Quochi, V. (2023). EpiLexO [ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa]. http://hdl.handle.net/20.500.11752/ILC-1005

Marinetti, A. (2020). Venetico. *Palaeohispanica. Revista sobre lenguas y culturas de la Hispania Antigua*, (20), 367–401. https://doi.org/10.36707/palaeohispanica.v0i20.374

Murano, F., Quochi, V., Del Grosso, A. M., Rigobianco, L., & Zinzi, M. (2023). Describing Inscriptions of Ancient Italy. The ItAnt Project and Its Information Encoding Process. *Journal on Computing and Cultural Heritage*, *16*, 1–14. https://doi.org/10.1145/3593431

Neill, I., & Schmidt, D. (2021). SPEEDy. A Practical Editor for Texts Annotated with Standoff Properties. *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*, *15*, 45.

Parr, T. (2009). *Language implementation patterns: create your own domain-specific and general programming languages*. The Pragmatic Bookshelf.

Parr, T. (2013). *The definitive ANTLR 4 reference*. The Pragmatic Bookshelf.

Poccetti, P. (2017). The documentation of Italic. In J. Klein, B. Joseph, & M. Fritz (Eds.), *Handbook of Comparative and Historical Indo-European Linguistics* (pp. 733–742, Vol. 2). De Gruyter Mouton. https://www.degruyter.com/document/doi/10.1515/9783110523874-001/html

Prag, J. R. W., & Chartrand, J. (2019). I. Sicily: Building a Digital Corpus of the Inscriptions of Ancient Sicily. In A. D. Santis & I. Rossi (Eds.), *Crossing Experiences in Digital Epigraphy: From Practice to Discipline* (pp. 240–252). De Gruyter Open Poland. https://doi.org/10.1515/9783110607208-020

Quochi, V., Bellandi, A., Khan, F., Mallia, M., Murano, F., Piccini, S., Rigobianco, L., Tommasi, A., & Zavattari, C. (2022). From Inscriptions to Lexicon and Back: A Platform for Editing and Linking the Languages of Ancient Italy. *Proceedings of Second Workshop on Language Technologies for Historical and Ancient Languages LT4HALA 2022*, 59–67.

Quochi, V., Bellandi, A., Mallia, M., Tommasi, A., & Zavattari, C. (2022). Supporting Ancient Historical Linguistics and Cultural Studies with EpiLexO. *CLARIN Annual Conference Proceedings*, 39.

Sarullo, G. (2016). The encoding challenge of the ILA project. In A. E. Felle & A. Rocco (Eds.), *Off the beaten track. epigraphy at the borders* (pp. 15–17). Archaeopress. https://www.archaeopress.com/Archaeopress/download/9781784913229.pdf#page=25

Spadini, E., & Turska, M. (2019). XML-TEI Stand-off Markup: One Step Beyond. *Digital Philology: A Journal of Medieval Cultures*, *8*(2), 225–239.

Vagionakis, I., Del Gratta, R., Boschetti, F., Baroni, P., Del Grosso, A. M., Mancinelli, T., & Monachini, M. (2022). 'Cretan Institutional Inscriptions' Meets CLARIN-IT. *CLARIN Annual Conference*, 139–150.

Zenzaro, S., Grosso, A. M. D., Boschetti, F., & Ranocchia, G. (2022). Verso la definizione di criteri per valutare soluzioni di scholarly editing digitale: Il caso d'uso GreekSchools. In F. Ciracì, G.

Miglietta, & C. Gatto (Eds.), *Aiucd 2022 proceedings* (pp. 20–25). https://amsacta.unibo.it/id/eprint/6848/

Topics in Periodicals from the Swedish Diabetes Association 1949 – 1990: Extending the Topic Modelling Tool Topics2Themes with a Timeline Visualisation

Maria Skeppstedt[1], **Gijs Aangenendt**[1,2], **Vera Danilova**[2], **Ylva Söderfeldt**[2]
[1]Centre for Digital Humanities and Social Sciences Uppsala,
Department of ALM, Uppsala University, Sweden
maria.skeppstedt@abm.uu.se
[2]Department of History of Science and Ideas, Uppsala University, Sweden
{gijs.aangenendt,vera.danilova,ylva.soderfeldt}@idehist.uu.se

Abstract

Existing methods for visualising temporal topic models typically present the information in an aggregated form, and do not offer any possibility to track the specific texts responsible for the change in topic prevalence over time. We present a new type of topic modelling-based timeline visualisation. It still provides an overview with aggregated topic information suitable for distant reading, while also allowing the user to gradually zoom into the image for more detail. At the most detailed level, the individual texts can be reached, which makes it possible to switch to close reading. The timeline visualisation was implemented as an extension of the topic modelling tool Topics2Themes, but this visualisation technique can be adapted to other topic modelling tools and algorithms. We showcase the timeline visualisation on a corpus of periodicals from the Swedish Diabetes Association, which is one of the patient organisation corpora studied within the interdisciplinary project ActDisease. One timeline visualisation was generated for the entire corpus. Additionally, we generated a timeline focusing on the texts that contain the word "dietitian". The two timelines, including the functionality to zoom into the graphs and reach the texts, were used to analyse the topics and how they vary. It could be concluded that some of the topics and topic timelines were predictable, while others revealed content that might be less expected. These results indicate validity of the method applied, and they also show that this visualisation technique could help us learn something new.

1 Introduction

A text collection covering a longer time period offers a wealth of possibilities for studying temporal change. Among many potential aspects of change, we focus on exploring prominent topics and the longitudinal variation of their prevalence. With the support of natural language processing (NLP) and text visualisation methods, it is possible to carry out such explorations on large text collections as well.

In this paper, we showcase how the NLP technique *topic modelling* can be used to automatically extract topics from a large text collection, and how the topics extracted can be visualised on a timeline. For this aim, we use Topics2Themes, which is a topic modelling tool maintained and further developed at the CLARIN node at the Language Council of Sweden (Skeppstedt et al., 2018). Topics2Themes has previously been applied to many different types of text collections, and the main contribution of the work presented here is an extension of the tool in the form of a timeline visualisation of the tool's topic modelling output.

Maria Skeppstedt, Gijs Aangenendt, Vera Danilova and Ylva Söderfeldt 2023. Topics in Periodicals from the Swedish Diabetes Association 1949 – 1990: Extending the Topic Modelling Tool Topics2Themes with a Timeline Visualisation. *Selected papers from the CLARIN Annual Conference 2023. Ed. by Krister Lindén, Thalassia Kontino and Jyrki Niemi.* Linköping Electronic Conference Proceedings 210, pp. 203–216. DOI: https://doi.org/10.3384/ecp210

The main difference between Topics2Themes and the existing approaches for presenting topic modelling output is that the tool uses topic modelling as a means for selecting and topically sorting texts that might be relevant for a manual analysis. Our timeline visualisation takes the same approach. More specifically, building on the basic information visualisation principle of "Overview first – details on demand", the timeline first provides an overview of how the topics vary over time, and thereafter makes it possible to zoom into the image to gradually reach more detailed information. At the most detailed level, it is possible to access the actual texts upon which the timeline is built. (See Figure 1 for an illustration of this approach.) By providing a topic overview, as well as functioning as a point of departure for locating potentially interesting texts, we aim to create a visualisation that makes it possible to switch between a *distant* reading of temporal topic variation and a *close* reading of the actual texts.

Previous development iterations of this topic modelling-based timeline visualisation (Skeppstedt, 2022, 2023; Stede et al., 2023) had a technical focus and were carried out by groups consisting of only NLP researchers. In contrast, the work presented here was conducted in an interdisciplinary team, more specifically within the historical research project ActDisease that investigates the history of patient organisations in twentieth century Europe. We applied the timeline visualisation on one of the corpora investigated in the project – a collection of digitised periodicals issued by the Swedish Diabetes Association. This enabled us to test to what extent the visualisation was useful in exploring the content of the corpus. In addition to reporting on the results of this interdisciplinary work, we describe the design of the timeline visualisation and discuss the ideas underlying the design choices.[1]

2 Background

Topic modelling is a form of unsupervised technique for automatic text exploration and categorisation. That is, no manually annotated training data or pre-defined categories are fed to the topic modelling algorithm. Instead, patterns in the text collection itself, e.g. word co-occurrence patterns, are used to automatically extract frequently re-occurring topics. Each topic extracted by the algorithm is typically represented by (i) an ordered list of texts in which the topic occurs, and (ii) an ordered list of words prevalent in texts associated with the topic. The texts and words are ordered based on their closeness to the topic. There are a number of different topic modelling algorithms, from more classic ones such as LDA (D. Blei & Lafferty, 2006; D. M. Blei et al., 2003) and NMF (Greene & Cross, 2017; Lee & Seung, 2001) to transformer-based approaches, such as BERTopic (Grootendorst, 2022).

For the specific task of applying topic modelling on temporally extended text collections, a number of different strategies/topic modelling algorithms are available. Dynamic topic models is a frequently used method, which differs from standard topic models in that the timestamps of the texts are included as one of the parameters (D. Blei & Lafferty, 2006; Grootendorst, 2022; Wang & McCallum, 2006). It is also possible to first use standard topic models and then apply additional statistical methods for analysing their temporality (Meaney et al., 2022), or to apply standard topic models on many short time frames and combine those into topics spanning longer time periods (Greene & Cross, 2017). Another possibility is to combine dynamic and standard topic models (Hida et al., 2018), or to use a more straight-forward approach of applying standard topic modelling on a temporal text collection and simply visualise the temporal variation of the topics (Stede et al., 2023).

There are also many examples of approaches developed for the task of visualising topic models applied to temporal data. To the best of our knowledge, however, there is no standard or best practice yet. Examples of visualisation techniques used are line charts (D. Blei & Lafferty, 2006; Grootendorst, 2023), stacked bars (Sheehan et al., 2021), a Sankey diagram (Malik et al., 2013), a heatmap (Meaney et al., 2022) and horizontal trend lines (Gad et al., 2015). There is also the "Theme river" visualisation (Günnemann, 2013; Günnemann et al., 2013; Havre et al., 2000) where coloured "streams" (or "rivers") with changing widths represent topic variation, and (in a more developed format) these streams can split and merge as the topics they represent become more or less semantically close (Cui et al., 2011). Common to

[1] Topics2Themes is a language-independent, open source topic modelling tool, found at: https://github.com/sprakradet/topics2themes.
The open source programming code for generating the timelines is provided at: https://github.com/CDHUppsala/topic-timelines.

all these approaches – in their original form – is that they all aim to visualise aggregated topic prevalence over time, i.e. they provide a distant view of the corpus without any possibility to easily access the texts that contain the topics.

Topic modelling is one, among many, NLP methods that have been used in visualisations aimed to support *distant reading* (Jänicke et al., 2015). Distant reading emerged within the field of literary studies as a means of analysing literature at scale, often with the help of computational methods (Moretti, 2013; Underwood, 2017). Since its emergence, the approach has spread to other disciplines within the (digital) humanities and social sciences and is also applied on other forms of text besides literary data (Gelfgren & Drakman, 2022). Distant reading is often used in tandem with *close reading* to identify patterns in large textual datasets that merit further close-up investigation through the reading of individual texts.

Topic models are often criticised for being difficult to interpret due to the numerical and statistical nature of their outputs (Hagen, 2018). Assessing the quality of the extracted topics cannot solely be done based on the statistical output, but often requires domain knowledge and a close reading of the associated texts. Despite this, most topic modelling-based timeline visualisations only provide the user an aggregated view of the corpus. We believe that adding support for close reading could be one way to address this limitation. This hypothesis is supported by the usability evaluation of the Theme river approach, which showed that although the users appreciated the information provided by the aggregated distant view of the corpus, this information was not sufficient. Instead, the need to read the actual texts that contributed to a topic at a given timestamp was recognised (Havre et al., 2000).

The usability evaluation of Theme river also resulted in additional feature requests. These included the ability to see the total number of texts at any time period, and to be able to provide a user-defined ordering of the theme rivers. The original, static Theme river visualisation was later developed, and incorporated in an interactive text visualisation tool called D-Vita (Günnemann, 2013; Günnemann et al., 2013). Here, the timeline is divided into user-defined time periods, e.g. a year, and close reading is made possible by allowing the user to access the most typical texts for a topic for each such time period. The theme river visualisation is also incorporated in another interactive tool, called Tiara (Liu et al., 2009). In this tool, each stream is overlaid with word clouds, and a possibility to zoom into the word clouds using the fisheye view technique is provided. A form of close reading functionality is provided also for this tool, by making it possible for the user to read an automatically selected set of semantically diverse texts that represent the topics.

3 Design objectives and requirements

In addition to the overall design objective of creating a topic modelling-based timeline visualisation, we also had four main design requirements. These were based on studying previous research, as well as on reflections made in the preceding development iterations and on discussions within the team.

(1) The most important design requirement was to *not* create a visualisation limited to showing aggregated information, disconnected from the texts from which the topics are derived. This requirement is also what most clearly sets the visualisation created here apart from the previous examples of temporal topic visualisations described above. Instead, we aimed to employ the standard information visualisation workflow "overview first [...] then details on demand" (Shneiderman, 1996), to make it evident that the high-level/aggregated information has been derived by combining information from individual texts. That is, when the original-size version of the visualisation is presented to the user, aggregated information from many texts should be shown (overview first). It should then be possible to demand more details by zooming into interesting areas of the visualisation, letting representations of the individual texts – and topics that occur in these texts – be shown. Finally, in search for potentially interesting text content, it should be possible for the user to demand even more detailed information, by shifting from exploring the representations of the texts to reading the actual text content. (See Figure 1 for an illustration of this approach.) We do not aim to limit the texts that can be directly accessed to a subset of the texts from which the timeline is created, as has been the case for previous tools that include some support for close reading (Günnemann, 2013; Günnemann et al., 2013; Liu et al., 2009). Instead, all texts for which the occurrence of one or more topics is indicated in the graph should be possible to access.

(2) Another design objective was to avoid showing any concrete numerical values (except timestamps) in the visualisation, or using any visualisation technique that is closely associated with the visualisation of numerical values, such as bar charts or line graphs. The rationale for this objective is that when concrete numerical values are introduced, the importance of their exact meaning might easily be overestimated. For topic modelling, the output generated is more fuzzy than e.g. word count statistics. Not only because many topic modelling algorithms are randomised, but also because the output generated is very sensitive to what configuration parameters are used (Da, 2019, p 625). Topic modelling offers *a* window through which a text collection can be viewed, rather than offering *the* window for viewing it. Also, each numerical output value of the topic modelling algorithms is often not meaningful in itself, but only in relation to other output values. Yet a reason to communicate uncertainty or fuzziness has to do with the nature of the underlying datasets used in humanities research. These datasets often consist of incomplete historical sources and are created through subjective and implicit decisions made during the data collection process, e.g. when selecting what sources to include in the dataset (Panagiotidou et al., 2022). For these reasons, we aimed for a visualisation that would somehow convey a fuzziness, and that would encourage active exploration of the data as well as an interpretation of the topic modelling output in relation to other values in the graph.

(3) It was also important for the resulting visualisation to consist of one, static image. That is, the topic timeline should be displayed in a single static graph, without using any form of dynamic functionality to convey the information required to explore the topic modelling output. The only dynamic element of the exploration should consist of the user zooming in and out in the image. This restriction enforces a simplicity (and thereby hopefully also an increased usability) of the timeline design. In addition, we believe that if a visualisation can be included as a zoomable image in an article or as a large image on a printed poster, it lowers the threshold for using and re-using it. Although an interactive graphical user interface is needed for some types of text exploration tasks, we believe that one dimension is lost when the simplicity of a static image is traded for an interactive system. An example of this is the original Theme river visualisation (Havre et al., 2000) in relation to its interactive version (Günnemann, 2013; Günnemann et al., 2013). We, however, employed one important exception to this restriction: To let the user easily switch from exploring the graph to reading the texts upon which the graph is based, we allowed the image to be dynamic in the sense that it can be configured to contain links to web pages where the texts can be read.

(4) The final main design objective consisted of including information about the temporal text frequency in the visualisation. That is, the timeline should visualise the variation in the number of texts that stem from different time periods. This information is interesting in itself, e.g. to inform on variation in text publication frequency or on the temporal prevalence of certain keywords that have been used to extract the text collection. The text frequency also helps the user to interpret the variation in topic prevalence, since the number of texts associated with a topic during a time period is in part dependent on how many texts stem from that period (Da, 2019, p 627). This design objective is also in line with one of the feature requests from the user evaluation of the Theme river approach (Havre et al., 2000).

In addition to the four main design requirements, we also had a number of smaller requirements. It should, for instance, be possible to compare different timelines, e.g. timelines from different corpora or those resulting from different topic modelling configurations. To make visualisations resulting from texts in various languages understandable to an international audience, it must also be possible to translate the topic labels. Building on our own reflections when using the first versions of the timeline, as well as on results from previous studies (Baumer et al., 2017; Havre et al., 2000; Stede et al., 2023), we also saw the usefulness of making it possible to manually categorise and re-order the automatically extracted topics. Finally, previous research has shown the importance of the visualisations being fairly scalable, i.e allowing both few and many topics to be visualised within the same graph (Gad et al., 2015).

4 Implementing the design

The basic components of the design are described in the caption of Figure 1. This figure also includes a symbolic illustration of how the design makes it possible to gradually move from “overview first” to

more "details on demand". Our approach is to (i) let the original-size graph consist of *combinations* of small graphical components that represent the texts and their topic associations, and (ii) make it possible to see *each individual* graphical component by zooming into the graph (**1:H**). For instance, that a topic occurs in a text is indicated by a *vertical bar* (**1:D**) and by a *circle* (1:**G**), both with a size proportional to the strength of the text-topic association. In the original-size graph, the combination of partly overlapping circles provides an overview of topic prevalence, while zooming in makes it possible to see each vertical bar (**1:D**) and each circle (1:**G**) that represents the topic-strength for the text.

The most detailed level, i.e. the text itself, can not be reached by zooming into the static image. We therefore implemented the possibility to associate a unique HTML link to each of the texts represented by a bar/circle in the graph. When clicking on the circle (**1:I**), the web page associated with the text is opened, e.g. a web page that contains the original text with the original page layout (**1:J**). This overview first-details on demand approach is our solution for how to make it possible to switch between distant and close reading.

In addition to these two most important features, i.e. to be able to **gradually zoom in for more detail** and **click in the graph to reach the actual texts**, we also provided several other configuration options to meet the design requirements.

One configuration regards **texts that share the same timestamp**. That is, to be able to plot texts on a timeline, each text must be provided with a timestamp. If several texts share the same timestamp, they will be plotted on the exact same position in the graph, obscuring each other. We therefore implemented a configuration possibility that spreads texts with the same timestamp along the x-axis. The configuration lets the user specify the following three pieces of information: (i) A small time fraction indicating the distance by which to move the x-position of a text when another text is already positioned at this timestamp, (ii) the width of the vertical bar indicating the topic-strength for the text, and (iii) the transparency of the topic-strength indicators. By configuring these three parameters, it is possible to use overlapping transparent circles to indicate topic-strength for each individual text. This results in a pattern of partly overlapping, transparent circles, which will not only show topic-strength variations for texts with the exact same timestamp, but also for texts that are positioned very close to each other on the x-axis (**1:H**).

Another configuration option regards **how to scale the vertical bar** (and circle) that indicates topic-strength. The height is scaled to make sure the tallest bar fits within the horizontal lane that represents the topic (**1:L**). This scaling can either be configured to be performed on a graph-global level, using the overall maximum topic-strength as the factor with which to scale the bar. It can also be carried out on a topic-local level, using the maximum topic-strength for each specific topic in question as the scaling factor. The first option makes it possible to compare association strengths between topics, but it also makes it difficult to see temporal variations for topics with weaker text associations.

Finally, a configuration which made it possible to **manually order the topics**, and/or to create groups of topics, was implemented.[2] The original topic sorting is still shown by the number associated with the topic, but the topics are resorted based on the user input. The user-defined groups of similar topics are indicated by alternating colours (**1:M**).

5 Applying the timeline visualisation on a corpus

As mentioned in the introduction, the timeline visualisation development presented here was conducted within an interdisciplinary team, as part of the historical research project ActDisease. The project investigates the history of patient organisations in twentieth century Europe, and has digitised a number of patient organisation periodicals from four different countries (Aangenendt et al., 2024). To develop and apply the timeline within the context of a historical research project made it possible to (i) add functionality to the timeline visualisation that would be immediately useful to historical research, and (ii) apply the timeline visualisation on a corpus upon which research had already been carried out using traditional historiographical methods. We chose the journal *Diabetes*, published by the Swedish Diabetes

[2]Practically, it was implemented by providing a parameter to the timeline generation function, which takes the form of a nested list with numbers associated with the topics.

Association.[3]

This corpus contains 8 891 pages from 233 individual issues, covering the period 1949-1990 (Aangenendt et al., 2024). The raw files for the corpus derive from scans made by Gothenburg University Library, available in full through GUPEA[4].

In addition to applying the timeline visualisation on the entire corpus, we also applied it on a subset. This subset was selected to be more focused on one specific issue important to the organisation, and only included text data from those pages that contain the word "dietitian". The dietitian profession was introduced in Sweden in the 1960s through a process that the SDA was heavily involved in. From the early 1950s, when they first started using the term, until the 1980s when the role was a clearly defined healthcare profession, the SDA participated in defining and negotiating the boundaries and position of the dietitian. Since this process is one that we had studied extensively through close reading of the corpus and archival material, a dietitian sub-corpus allowed us to view the topic model visualisation from a standpoint of familiarity with the text it represents.

The Diabetes corpus is organised with pages as the text unit and the same unit was used for topic modelling. The texts were lemmatised using Efselab[5] (Östling, 2018), and thereafter the topic modelling tool Topics2Themes (Skeppstedt et al., 2018) was applied on the corpus. We used the Swedish stop word list included in NLTK (Bird, 2002), which we expanded by iteratively running the topic modelling tool on the corpus and inspecting the output. Topics2Themes provides an interactive graphical user interface, which can be used to inspect the output, e.g. regarding which words should be added to the stop word list. In the final iteration, we used the functionality in Topics2Themes where the topic modelling algorithm is run several times, and only stable topics are retained[6], an approach previously used by Baumer et al., 2017. We instructed the topic modelling algorithm to return 70 topics for the full corpus and 30 topics for the sub-corpus with texts containing the word "dietitian". This resulted in 51 stable topics being generated for the full corpus and 21 topics for the sub-corpus. We configured the tool to extract the 20 most typical words for each topic. For the texts, we used a cut-off of the 200 most closely associated texts in the visualisation, provided that the text contained at least one of the top 20 most typical words.

We automatically assigned the timestamp based on the issue in which the text appeared, assuming that the publication dates for the issues were evenly distributed over the year. All texts from one issue were thus assigned the same timestamp, but were configured to be slightly moved along the x-axis when visualised, in order not to collide. The texts within an issue were ordered according to page number, resulting in the first page being positioned on the timestamp for the issue and subsequent pages being slightly moved to the right. A local scaling of the topic strengths was applied, which makes it easier to see temporal variations for a topic, but more difficult to compare topic strengths between different topics. When clicking on the topic-strength symbols, the visualisation was configured to direct the user to a web page containing a pdf with the original page layout (see Figure **1:J**). English translations for the topic labels were obtained by applying Google translate on the Swedish labels and then manually correcting the automatic translations.

6 Exploring the timeline visualisations

As a first step in using the timeline visualisations created, the topics were manually combined into larger groups based on the topic labels. For the full corpus, five groups were created, and there were three topics not included in any of these groups. Five groups were also created for the dietitian-subset, and one topic could not be assigned to any of the five groups. The timelines using these groups are shown in Figure 2 (for the full corpus) and in Figure 3 (for the sub-corpus containing the word "dietitian"). The grouping of topics, as well as the exploration of the timeline described below, was carried out by the historian leading the project, who through close reading of the corpus and archival material had the knowledge required to

[3] *Svenska Diabetesförbundet* in Swedish (*Riksförbundet för sockersjuka* before 1956).

[4] https://gupea.ub.gu.se/handle/2077/64597

[5] https://github.com/robertostling/efselab

[6] More specifically, the topic modelling algorithm was run 50 times, and from these re-run outputs we kept the five most typical ones, and only retained topics that occurred in all five outputs. Similarity between topics was measured based on their associated words, and two topics with a 60% overlap of the top 20 most similar words were counted as the same topic.

assess the information visualised.

The difference between the timeline characteristics for different topics is immediately evident when looking at the original-size graphs. E.g. in Figure 2, it can be seen that some topics occur more or less the entire time period studied, e.g. the first two topics in the image, while others are limited to certain time periods – e.g. topic 49 – or to only having strong occurrences in certain time periods, e.g. topics 16 and 10. There are topics that occur with regular intervals, e.g. topics 3 and 6, while others do not show any obvious regularity, e.g topics 51 and 13. By zooming in, it is possible to see that some of the topics are represented with a single transparent circle per aggregated text-line, which shows that these topics occur only once per issue for a period of time, e.g. topics 2, 16 and 17. Zooming in further shows that topic 2 generally occurs towards the end of the publication and topic 17 in the beginning, while 16 at first occurs in the beginning and then later in time towards the end of the publications. Other topics, in contrast, have opaque segments created by many, partly overlapping circles positioned close to each other, which shows that the topic occurs several times in an issue, e.g. topics 6, 19 and 21. The two semantically similar topics 6 and 28 seem to be the same topic, since they have very similar timeline characteristics, and since topic 28 replaces topic 6, around 1979. As expected, the vertical lines that represent the texts and their timestamps occur with a regular frequency in Figure 2, except for a short time period in the beginning of the 1950s, when no SDA periodicals were published. In Figure 3, in contrast, the frequency of the text-lines vary, corresponding to a variation in how often there are texts published that contain the word "dietitian".

Exploring the visualisation of topics in a familiar corpus makes it possible to evaluate the extent to which it captures known trends in the material. In Figure 3, topic 16 relates to the training of dietitians, an issue that the SDA pioneered and were instrumental in implementing. Their campaign to establish the dietitian as a new profession in Swedish healthcare began around 1960, and led to a trial course in 1969, which was followed by repeated efforts for a permanent solution. Finally, in 1978, a higher education program for dietitians started (Söderfeldt, 2024). Here, the visualisation closely aligns with the period that this issue was prominent in the SDA, and shows spikes in those phases when important steps took place.

Grouping the topics in categories makes it possible to see trends within thematic fields. For instance, the topics 2-51 in Figure 2 are related to diabetes treatment. It can be seen how certain manufacturers of insulin, glucose tests, and injection material come and go over time (e.g. 5, 16, 33) whereas other appear more stable (e.g. 42 and 51, regarding fundamental pathophysiology and treatment). Similarly, the group 4-50, which relates to food, has different brands of artificial sweeteners and sugar-free foods that appear for particular, limited time periods (e.g. 18, 30, 34), but also some more stable topics like 4, 10, and 39 that are more generally related to diets. By grouping topics, we are able to view a chronological map of a particular theme in the corpus, which provides an initial orientation in overall trends. An interesting find are also topics that were not related to particular themes, such as topic 9 in Figure 2, which represents texts with a certain type of language. It relates to texts dealing with personal experience, which show a remarkable increase in the 1980s.

In the case of dietitian education (topic 16 in Figure 3), the visualisation did not allow us to make new discoveries in the corpus. Rather, it served to confirm that the method depicts meaningful trends. Similarly, topic 46 in Figure 2, relating to nursing (and dietitian) education, is in zenith from the early 1970s to the early 1980s, during which time the SDA also engaged in nursing education by arranging annual courses on diabetes care. Topic 7 in Figure 2 regards the SDA owned and operated retreat centre Nordanede, which came to them through a donation in 1963 and was sold in 1984. Also topic 49 corresponds to known events, since insulin pumps, which were first developed in 1974, hardly appear at all until 1982 and then quickly become a frequent topic. Other topics, however, give indications of change that we had not previously detected and point towards interesting fields of further study. Using the feature of zooming in and accessing the texts, we can determine that the more transient treatment-related topics in Figure 2 (5, 16, 33) are found in advertisements that are repeated over a certain period of time. The more stable topics (42 and 51), however, are drawn from a mix of both editorial and advertising content. A close reading of the texts from these topics can therefore help in characterising the interplay between

the way that the patient organisation discussed their illness, and how pharmaceutical companies marketed their products. Findings like these provide helpful directions for qualitative research in the corpus texts as well as supplementary sources like archival material. In the thematic group related to the association (topics 3-45 in Figure 2) we see a trend in that many of these administrative topics taper off in the 1980s. It is crucial to note that changes like these could be due to changes in layout and format, for instance title pages or other elements that are frequently repeated in the corpus. Therefore, the possibility to directly zoom in on the texts that the topics appear in offer a more transparent solution that allow the historian to not only view trends, but to investigate what they consist of. Sampling some of the text pages from this thematic field, we find that several topics are indeed drawn from elements like address lists (6, 28) or front matter (29). However, topics 3 and 19, both relating to annual meetings, reveal a drastic change in how these matters are presented in the periodical not just quantitatively, but also in style. From using the publication as an outlet for meeting reports, the significance of internal congresses not only decreases, but also becomes more journalistic than administrative. This gives indications to cultural changes in the organisation.

7 Final discussion and future directions

A method previously suggested for evaluating the validity of topic models is to investigate whether the model is able to extract known topics (Da, 2019, p. 628). If that is the case, also previously unknown topics and trends indicated by the model might potentially be interesting to explore. The timeline visualisation created helped us to not only detect such known topics, but also to investigate to what extent the prevalence of these topics corresponded to known historical events. The timeline and the possibility to read the actual texts giving rise to the topic-trends was equally useful in helping us to interpret unknown topics.

We are not aware of any previous approaches for the topic-timeline visualisation task that offer the same possibilities for text and topic exploration. This applies both to the functionality which allows the user to click on the timeline to access the actual texts, as well as to the level of expressiveness when showing aggregated topic prevalence in the original-size timeline visualisation. None of the approaches described in the background makes it possible to display this type of timeline characteristic for each topic, i.e. the characteristic that we achieve by letting partly overlapping transparent circles indicate the interplay between topic strength and topic prevalence. Our approach also scales well to displaying a timeline for many topics, and fulfils the requirement to visualise how the text frequency varies over the time period studied. The use of partly overlapping transparent circles as topic strength indicators does not only result in expressiveness, but we also believe – albeit rather subjectively – that they help us communicate the fuzziness we aim for.

Despite this practical usefulness of the timeline visualisation created, there are still many potential improvements for future development iterations. The approach used here to direct the user to a pdf that contains the original page layout – without providing any indications from the topic modelling output – has the drawback that the relevance of some texts to the topic is enigmatic to the human reader. The feature provided by, e.g. the graphical user interface of Topics2Themes, to markup words associated with the topic, might provide a help to understand the model better. This would be especially helpful for the type of corpus used here, where the text unit consists of a page, since a page sometimes includes one or more separate texts and it is unclear if the topic has been detected in one, both, or the texts together. Another problem associated with the pdf files that were linked to, was that they lacked the context of neighbouring pages, i.e. it would have been practical to be able to easily reach the previous or following page in the periodical. A problem, which is more related to the actual timeline visualisation, is that labels for the timeline and topics are no longer visible when the user has zoomed in. To facilitate navigation, it might therefore be useful with local indications of timestamps and topics that are provided with a font size small enough to only be visible when the user has zoomed in.

A number of additional features might also be added to the timeline visualisation. One such feature, which could be added alongside the manual categorisation and ordering of the topics, is the option to select an automatic ordering of the topics, by using a hierarchical topic model or a semantic similarity

measure of texts or words associated with the topics. Another potentially useful feature would be to allow the researcher to build sub-corpora for close reading, based on texts that contain interesting topics.

Future work could also focus on making the timeline visualisation compatible with additional types of topic modelling algorithms and outputs. The implementation that we currently provide uses the json output format generated by Topics2Themes. However, the timeline design is by no means specific to Topics2Themes and could easily be adapted to another output format. This would also make it easier to compare the usefulness of different types of topic modelling algorithms for creating the topic timeline.

It can also be noted that the focus of the work presented here has been to visualise the *texts*, in particular how the prevalence of the topics occurring in the texts vary over time. We have used the words associated with the topics only as topic-labels in the visualisations developed. To also create a static timeline visualisation of the *words* associated with the topics is an equally relevant task. For instance, it might be interesting to create visualisations that convey the contrast of the fuzziness of the topic modelling output with more objective – but not necessarily fully objective in all aspects (Panagiotidou et al., 2022) – measures, such as the frequency count of the words associated with the topics.

Finally, we intend to perform a more structured user evaluation, which could include the visualisation technique presented here and some of the updates we have suggested. Such an evaluation might, for instance, be performed on the Diabetes corpus – possibly with developed methods for selecting and segmenting texts – or on some of the other corpora investigated within the ActDisease project.

Acknowledgements

We thank Andrew Burchell for his constructive comments that greatly helped to improve the manuscript.

The work described here was conducted within the ActDisease project. ActDisease is funded by the European Union (ERC ActDisease, ERC-2021-STG 101040999). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

The development of the topic-timeline programming code was also partly supported by InfraVis (the Swedish Research Council, grant 2021-00181) and Huminfra (the Swedish Research Council, grant 2021-00176). The Topics2Themes tool is maintained by Swe-CLARIN/The National Language Bank of Sweden (the Swedish Research Council, 2017-00626).

References

Aangenendt, G., Skeppstedt, M., & Söderfeldt, Y. (2024). Curating a historical source corpus of 20th century patient organization periodicals. *Proceedings of the Huminfra Conference (HiC 2024)*, 76–82. https://doi.org/10.3384/ecp205011

Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, *68*(6), 1397–1410.

Bird, S. (2002). NLTK: The natural language toolkit. *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*.

Blei, D., & Lafferty, J. (2006). Dynamic topic models. *ACM International Conference Proceeding Series; Vol. 148: Proceedings of the 23rd international conference on Machine learning; 25-29 June 2006*, 113–120.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*.

Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., Qu, H., & Tong, X. (2011). Textflow: Towards better understanding of evolving topics in text. *IEEE Trans. Vis. Comput. Graph.*, *17*(12), 2412–2421. https://doi.org/10.1109/TVCG.2011.239

Da, N. (2019). The computational case against computational literary studies. *Critical Inquiry*, *45*, 601–639. https://doi.org/10.1086/702594

Gad, S., Javed, W., Ghani, S., Elmqvist, N., Ewing, T., Hampton, K. N., & Ramakrishnan, N. (2015). Themedelta: Dynamic segmentations over temporal topic models. *IEEE Transactions on Visualization and Computer Graphics*, *21*(5), 672–685. https://doi.org/10.1109/TVCG.2014.2388208

Gelfgren, S., & Drakman, A. (2022). How to combine close and distant reading within the history of science and ideas: Two examples from ongoing research. *Lychnos*, (1), 85–108.

Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political analysis*, *25*(1), 77–94.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Grootendorst, M. (2023). Dynamic topic modeling, visualization.

Günnemann, N. (2013). D-vita: A visual interactive text analysis system using dynamic topic mining. *Datenbanksysteme für Business, Technologie und Web*. https://api.semanticscholar.org/CorpusID:15848321

Günnemann, N., Derntl, M., Klamma, R., & Jarke, M. (2013). An interactive system for visual analytics of dynamic topic models. *Datenbank-Spektrum*, *13*(3), 213–223. https://doi.org/10.1007/s13222-013-0134-x

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information Processing & Management*, *54*(6), 1292–1307. https://doi.org/10.1016/j.ipm.2018.05.006

Havre, S., Hetzler, B., & Nowell, L. (2000). Themeriver: Visualizing theme changes over time. *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, 115–123. https://doi.org/10.1109/INFVIS.2000.885098

Hida, R., Takeishi, N., Yairi, T., & Hori, K. (2018, July). Dynamic and static topic model for analyzing time-series document collections. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 516–520). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2082

Jänicke, S., Franzini, G., Scheuermann, G., & Cheema, M. (2015). On close and distant reading in digital humanities: A survey and future challenges. a state-of-the-art (star) report. *Eurographics Conference on Visualization (EuroVis)*.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 556–562.

Liu, S., Zhou, M. X., Pan, S., Qian, W., Cai, W., & Lian, X. (2009). Interactive, topic-based visual text summarization and analysis. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 543–552. https://doi.org/10.1145/1645953.1646023

Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., & Shneiderman, B. (2013). Topicflow: Visualizing topic alignment of twitter data over time. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 720–726. https://doi.org/10.1145/2492517.2492639

Meaney, C., Escobar, M., Stukel, T. A., Austin, P. C., & Jaakkimainen, L. (2022). Comparison of methods for estimating temporal topic models from primary care clinical text data: Retrospective closed cohort study. *JMIR medical informatics*.

Moretti, F. (2013). *Distant reading*. Verso Books.

Östling, R. (2018). Part of speech tagging: Shallow or deep learning? *North. Eur. J. Lang. Technol.*

Panagiotidou, G., Lamqaddam, H., Poblome, J., Brosens, K., Verbert, K., & Moere, A. V. (2022). Communicating uncertainty in digital humanities visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 1–11. https://doi.org/10.1109/TVCG.2022.3209436

Sheehan, S., Luz, S., & Masoodian, M. (2021, April). TeMoTopic: Temporal mosaic visualisation of topic distribution, keywords, and context. In H. Toivonen & M. Boggia (Eds.), *Proceedings of the eacl hackashop on news media content analysis and automated report generation* (pp. 56–61). Association for Computational Linguistics. https://aclanthology.org/2021.hackashop-1.8

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings of the IEEE Symposium on Visual Languages*, 336–343. https://doi.org/10.1109/VL.1996.545307

Skeppstedt, M. (2022). The topic modelling tool Topics2Themes applied to different types of climate change-related texts. [The CLARIN Bazaar 2022. https://www.clarin.eu/content/clarin-bazaar-2022].

Skeppstedt, M. (2023). Topics in Swedish news on climate change: A timeline 2016 – 2023. *CLARIN Annual Conference Proceedings 2023*, 150–154.

Skeppstedt, M., Kucher, K., Stede, M., & Kerren, A. (2018). Topics2Themes: Computer-assisted argument extraction by visual analysis of important topics. *Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, 9–16.

Söderfeldt, Y. (2024). Joint efforts in the Swedish model: The Swedish Diabetes Association under Nancy Eriksson (1956-1978) [Manuscript submitted for publication]. Department of History of Science; Ideas, Uppsala University.

Stede, M., Bracke, Y., Borec, L., Kinkel, N. C., & Skeppstedt, M. (2023). Framing climate change in Nature and Science editorials: applications of supervised and unsupervised text categorization. *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-023-00199-7

Underwood, T. (2017). A genealogy of distant reading. *DHQ: Digital Humanities Quarterly*, *11*(2).

Wang, X., & McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. *Conference on Knowledge Discovery in Data: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining; 20-23 Aug. 2006*, 424–433.

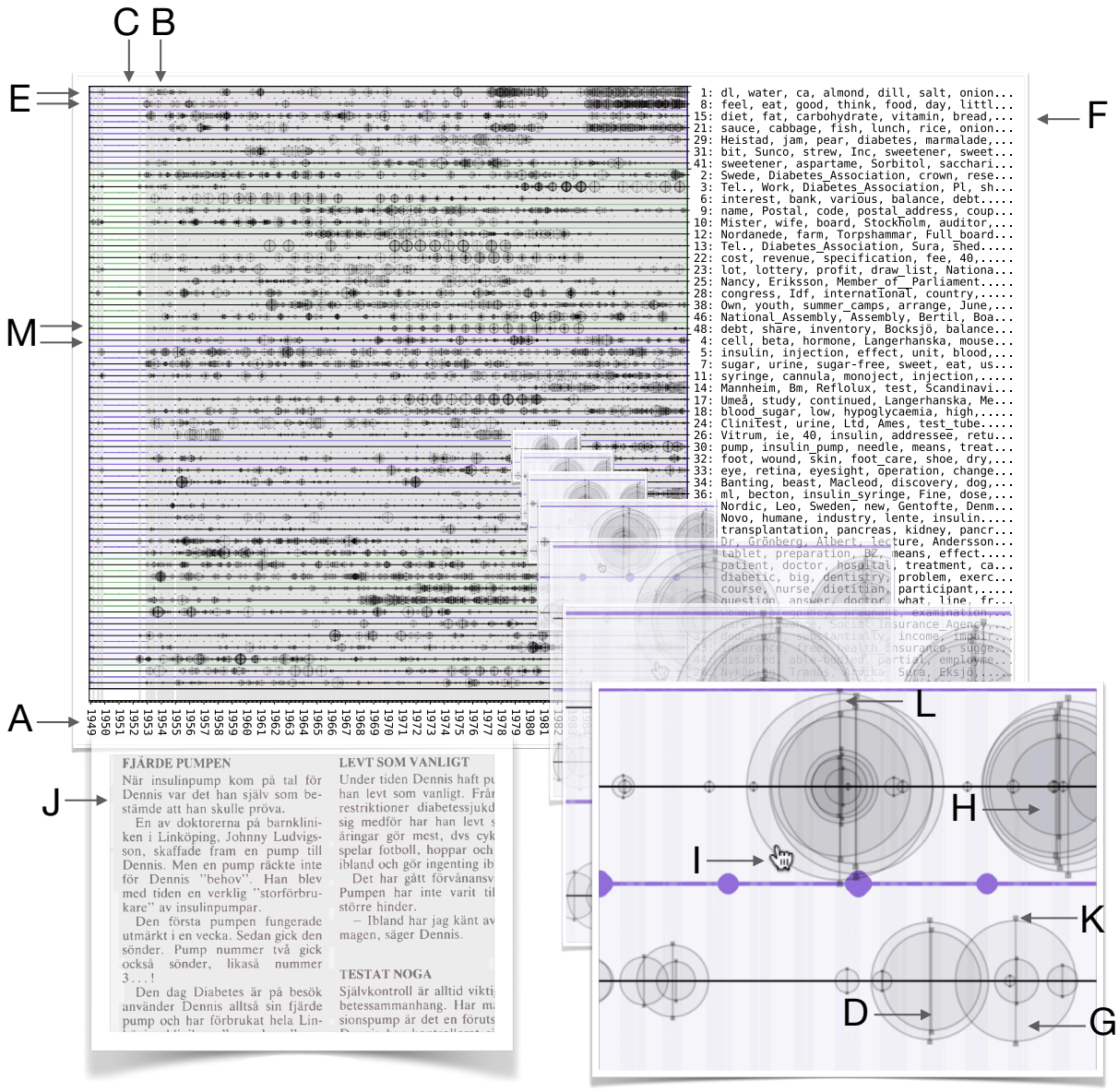


Figure 1: A visualisation of the approach to gradually zoom in for more detail and then finally click on the topic-strength indicator to reach the text.

Texts: Time is represented by position on the x-axis (**A**). Each text in the collection is visualised by a long vertical line positioned at the timestamp for the text (**B**). The visualisation of how the number of texts vary between different time periods is exemplified by the contrast between (**B**) and (**C**), i.e. many texts were published at (**B**) and none at all at (**C**).

Topics: The topics are represented by position on the y-axis, i.e. each topic is represented by a horizontal lane. (**E**) indicates the first and second of the horizontal topic lanes. The topic labels, which are created by the words most closely associated with the topic, are positioned to the right of the horizontal lanes (**F**). Ellipsis indicates that not all of the most closely associated words fit into the space available for labels. The centre of each horizontal topic lane is marked by a thin, black horizontal line, which we refer to as the *topic-line* (**E**).

Text-topic associations: If a text is one of the *n* texts most closely associated with a topic, this is indicated at the point where the *text-line* intersects the *topic-line*, by two indicators: by a vertical bar (**D**) – which runs along the text-line, and which has a height proportional to the text's topic-strength – and by a circle (**G**) – which has its centre at the intersection and which has a radius proportional to the topic-strength.

Reach the text: By clicking on the circle representing topic-strength (**I**), the web page associated with the text is opened (**J**). It is also possible to click on the small circles at the top and bottom of the topic-strength bars to reach the linked page (**K**). (**H**, **L** and **M** are described elsewhere in the text.)

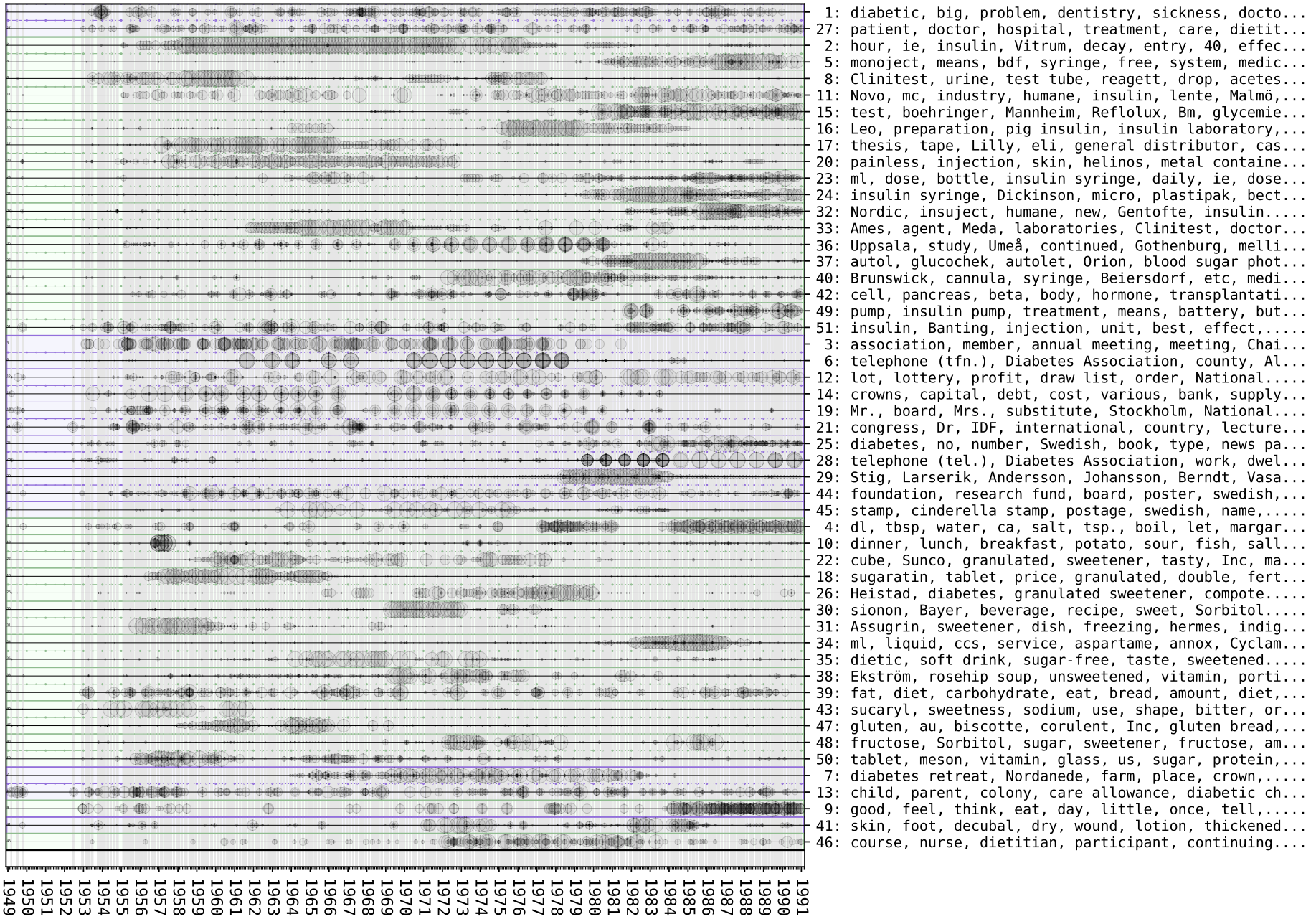


Figure 2: Timeline for the entire corpus from the Diabetes periodical. Each vertical line represents a text. The circles on the horizontal topic-lines indicate the occurrence of a topic in the text.

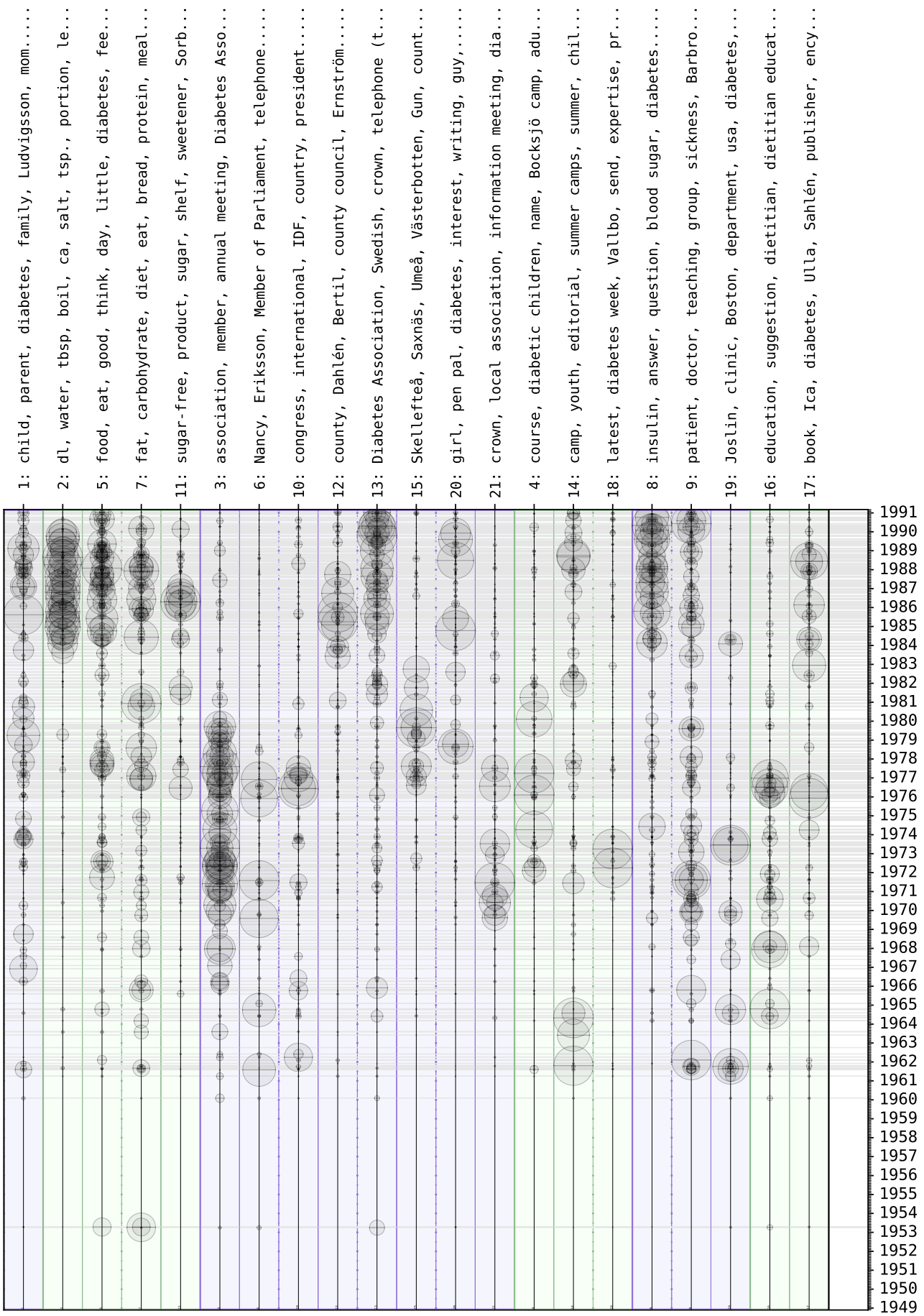


Figure 3: Timeline for a subset of the corpus containing the word "dietitian". Each vertical line represents a text. The circles on the horizontal topic-lines indicate the occurrence of a topic in the text.