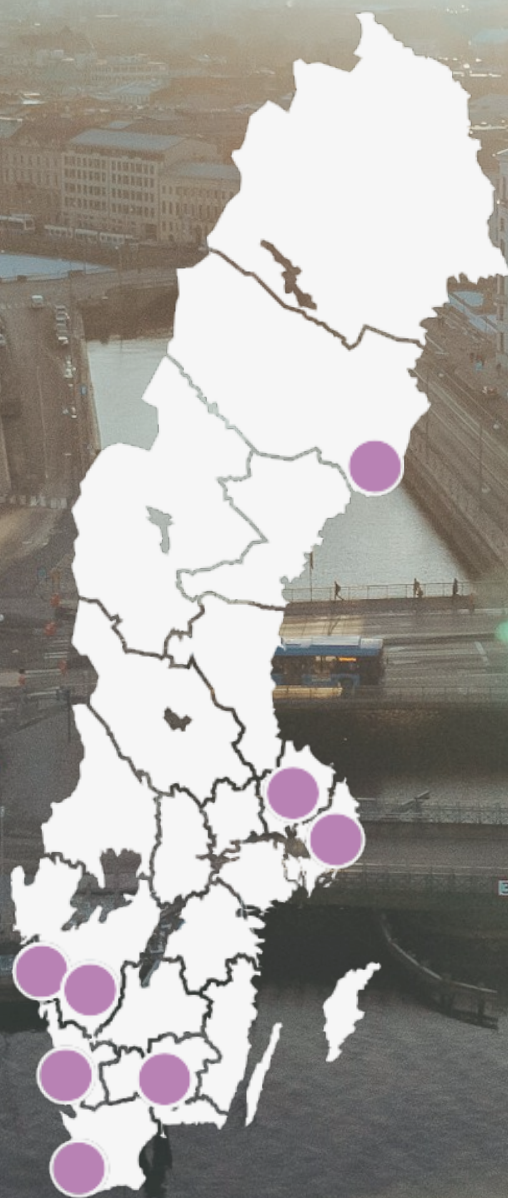


HiC 2024

Huminfra Conference

10-11 January, 2024
Gothenburg, Sweden



HUM
INFRA



Swedish
Research
Council

Proceedings of the
Huminfra Conference (HiC 2024)

edited by

Elena Volodina, Gerlof Bouma, Markus Forsberg, Dimitrios Kokkinakis,
David Alfter, Mats Fridlund, Christian Horn, Lars Ahrenberg, Anna Blåder

Proceedings and all papers therein
published under a CC-BY 4.0 license:
<https://creativecommons.org/licenses/by/4.0>

Cover image by
Jonas Jacobsson/Unsplash

Linköping Electronic Conference Proceedings 205
ISBN 978-91-8075-512-2
ISSN 1650-3686
eISSN 1650-3740
<https://doi.org/10.3384/ecp205>

Preface

Huminfra¹ is a Swedish national research infrastructure supporting digital and experimental research in the Humanities by providing users with a single entry point for finding existing Swedish materials and research tools, as well as developing national methods courses. On January, 10-11, 2024, Huminfra had its first conference with the aim to showcase the variety of infrastructural tools, resources and initiatives aimed at supporting digital and experimental research in the Humanities.

We invited Huminfra researchers and associated colleagues to submit presentations for the Huminfra Conference (HiC) which took place on 10-11 January, 2024 in Gothenburg. We invited two types of contributions, to be presented orally or as a demo presentation (i.e., work-in-progress or finished software, hardware technology, tools, datasets, and so forth):

- short papers of approximately 3-6 pages (excluding references) or
- short abstracts of approximately 300 words

We accepted 22 short papers, and 25 abstracts.

We solicited contributions that cover any aspect of research infrastructure components for Digital Humanities (DH) or for experimental research in the Humanities, focusing primarily, but not exclusively, on:

- Resources and tools aimed at DH research, including collaborative tools and platforms
- Infrastructures (e.g., digital archives and repositories; experimental tasks)
- Description of DH centers, networks, and related projects
- DH expertise and expertise in experimental Humanities
- Practical aspects of how tools/services/resources have been and can be used
- User training and education (e.g., "how-to" tutorials)
- Availability of tools/services/resources, and associated ethical, privacy or legal constraints
- Funding opportunities and strategies for building and sustaining DH infrastructure
- User involvement and/or case studies where research infrastructure components have been used
- Emerging trends and future directions in DH infrastructure development (e.g., AI)
- Dissemination of digital and experimental infrastructures, data, software, and/or research output in and beyond knowledge institutions
- Demos of any of the above
- Other, related to infrastructure, issues and challenges such as DH at the intersection of disciplines within and beyond the arts and humanities

The present volume collects all short papers. Abstracts are collected in a separate non-archival volume available at <https://www.huminfra.se/HiC-2024>.

HiC and Huminfra gratefully acknowledges funding and support from the Swedish Research Council (grant number 2021-00176) and all the partner institutions: Lund University, Umeå University, University of Gothenburg, the Royal Institute of Technology, the National Library of Sweden, Stockholm University, the Swedish National Archives, Uppsala University, the Swedish School of Library and Information Science, Linnaeus University, and Halmstad University.

¹<https://www.huminfra.se/>

Program committee

- Elena Volodina, Språkbanken Text (SBX), University of Gothenburg (Chair)
- Lars Ahrenberg, Swe-Clarin, Linköping university
- David Alfter, Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg
- Gerlof Bouma, Språkbanken Text (SBX), University of Gothenburg
- Daniel Brodén, Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg
- Coppélie Cocq, Humlab, Umeå University
- Jens Edlund, Språkbanken Tal (SBT), The Royal Institute of Technology (KTH)
- Anna Foka, Centre for Digital Humanities (CDHU), Uppsala University
- Markus Forsberg, Språkbanken Text (SBX), University of Gothenburg
- Mats Fridlund, Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg
- Koraljka Golub, Digital humaniora, Linnaeus University
- Marianne Gullberg, Lund University Humanities Lab (Humanistlaboratoriet), Lund University (LU)
- Justyna Sikora, KB-labb, National Library of Sweden (Kungliga biblioteket)
- Christian Horn, Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg
- Olof Karsvall, Riksarkivet (SNA)
- Dimitrios Kokkinakis, Språkbanken Text (SBX), University of Gothenburg
- Cecilia Lindhé, Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg
- Matts Lindström, Centre for Digital Humanities (CDHU), Uppsala University
- Gustaf Nelhans, The Swedish School of Library and Information Science, University of Borås
- Tomas Nilson, Digital Laboratory Centre (DLC), Halmstad University
- Harko Verhagen, Digital humanvetenskap (DHV), Stockholm University
- Jonathan Westin, Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg

Organizing committee

- Elena Volodina, Språkbanken Text (SBX), University of Gothenburg (Chair)
- Gerlof Bouma, Språkbanken Text (SBX), University of Gothenburg
- Markus Forsberg, Språkbanken Text (SBX), University of Gothenburg
- Dimitrios Kokkinakis, Språkbanken Text (SBX), University of Gothenburg
- David Alfter, Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg
- Mats Fridlund, Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg

- Christian Horn, Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg
- Lars Ahrenberg, Swe-Clarin, Linköping university
- Anna Blåder, Lund University Humanities Lab (Humanistlaboratoriet), Lund University

Content

Preface	i
<i>Elena Volodina, Gerlof Bouma, Markus Forsberg, Dimitrios Kokkinakis, David Alfter, Mats Fridlund, Christian Horn, Lars Ahrenberg, Anna Blåder</i>	
Open Brain AI: An AI Research Platform	1
<i>Charalambos Themistocleous</i>	
Profiles for Swedish as a Second Language: Lexis, Grammar, Morphology	10
<i>Elena Volodina, David Alfter, Therese Lindström Tiedemann</i>	
Digital History and Immaterial Infrastructure: A Bottom-Up Approach	20
<i>Sune Bechmann Pedersen, Marie Cronqvist, Kajsa Weber</i>	
Documentation of data making, processing and use facilitates future reuse of research data: the CAPTURE project	26
<i>Isto Huvila, Stefan Ekman</i>	
Queerlit – a bibliography of Swedish fiction with LGBTQI topics	31
<i>Siska Humlesjö, Jenny Bergenmar, Arild Matsson</i>	
From Zipf distribution to Universal Dependencies – Interactive Notebooks for Swedish Text Analysis	36
<i>Dimitrios Kokkinakis</i>	
Tradita innovare, innovata tradere	41
<i>Lars Borin, Louise Holmer</i>	
Collectio: a software especially designed for creating dynamic libraries for fluid and multilingual text traditions	51
<i>Britt Dahlman</i>	
AI, Data Curation and the Data Readiness of Heritage Collections: Exploring the Swedish Newspaper Archive at KBLab	60
<i>Justyna Sikora, Chris Haffenden</i>	
SAOL och svensk språkvetenskaplig infrastruktur – nu och i framtiden	68
<i>Louise Holmer, Ann Lillieström, Emma Sköldberg, Jonatan Uppström</i>	
Curating a historical source corpus of 20th century patient organization periodicals	76
<i>Gijs Aangenendt, Maria Skeppstedt, Ylva Söderfeldt</i>	
On two SweLL learner corpora – SweLL-pilot and SweLL-gold	83
<i>Elena Volodina</i>	
STUnD: ett Sökverktyg för Tvåspråkiga Universal Dependencies-trädbanker	95
<i>Arianna Masciolini, Márton A. Tóth</i>	
DASH Swedish National Doctoral School in Digital Humanities: From Local Expertise to National Research Infrastructure	110
<i>Matti La Mela, Daniel Brodén, Coppélie Cocq, Anna Foka, Koraljka Golub, Clelia LaMonica, Jonathan Westin</i>	

Research stories on Twitter <i>David G. Lorentzen, Gustaf Nelhans</i>	115
Humanistic AI: Towards a new field of interdisciplinary expertise and research <i>Mats Fridlund, David Alfter, Daniel Brodén, Ashely Green, Aram Karimi, Cecilia Lindhé</i>	122
Designing digitally-driven integrative interdisciplinarity: Professionalism between protocol and judgement <i>Daniel Brodén, Mats Fridlund, Cecilia Lindhé</i>	128
From the Arctics to Antarctica - A multimodular visualisation of data <i>Jonathan Westin, Tristan Bridge, Matteo Tomasini</i>	135
The DIGARV Platform: A collaborative platform for working with cultural heritage data and research data <i>Johan Åhlfeldt, Arild Matsson</i>	141
Samförfattande som datadriven tvärvetenskap: Pragmatiska lärdomar från SweTerror-projektet <i>Daniel Brodén, Mats Fridlund, Leif-Jöran Olsson, Magnus P. Ängsal, Patrik Öhberg</i>	148
Accessing centuries of documentation - Resources to improve access to Swedish rock art documentation and metadata <i>Ashely Green, Tristan Bridge, Christian Horn, Siska Humlesjö, Aram Karimi, Johan Ling, Jonathan Westin</i>	154
Konsten att bedriva svensk ordforskning utan att kränka upphovsrätten <i>Gerlof Bouma, Markus Forsberg, Justyna Sikora, Emma Sköldberg</i>	161

Open Brain AI: An AI Research Platform

Charalambos Themistocleous¹

¹ University of Oslo, Helga Engs hus 4. etg Sem Sælands vei 70371, Oslo, Norway

Abstract

Language assessment is pivotal in identifying therapeutic interventions for speech, language, and communication disorders stemming from neurogenic origins, developmental or acquired, and student performance in the classroom. Traditional assessment techniques, however, are predominantly manual, necessitating extensive time and effort for administration and scoring. Such procedures can exacerbate the stress experienced by patients. In response to these inherent challenges, we introduced Open Brain AI (<https://openbrainai.com>). This state-of-the-art computational platform leverages advanced AI methodologies, encompassing machine learning, natural language processing, large language models, and automated speech-to-text transcription. These capabilities enable Open Brain AI to autonomously analyze multilingual spoken and written language productions. This work aims to present the development and evolution of Open Brain AI, elucidating its AI-driven language processing components and the intricate linguistic metrics it employs to evaluate the overarching and granular discourse structures. Open Brain AI significantly reduces the workload on researchers, clinicians, and teachers by facilitating rapid and automated language analysis. It allows healthcare and education professionals to optimize their operational processes, reallocating precious time and resources to more personalized user interactions. Moreover, Open Brain AI provides clinicians, researchers, and educators the autonomy to undertake essential data analytics, freeing up more bandwidth to focus on other vital facets of therapeutic intervention and care.


Keywords

Open Brain AI, Large Language Models, NLP

1. Introduction

Assessing speech, language, and communication is critical for clinicians and researchers. It informs clinicians about the neurologic functioning of their patients, provides early linguistic biomarkers of conditions, such as Mild Cognitive Impairment (MCI), and guides treatment (1-7). Furthermore, speech and language assessments are critical for evaluating the classroom performance of first and second-language students. Nevertheless, the manual evaluation of speech, language, and communication is cumbersome, time-consuming, and subjective, as it depends on the expertise and training of those who perform the evaluation. Moreover, manual assessments, such as the Boston Naming Test (BNT; 8), Western Aphasia Battery-Revised (WAB-R Kertesz (9)), Boston Diagnostic Aphasia Examination (BDAE; 10), and Psycholinguistic Assessment of Language Processing in Aphasia (PALPA; 11) often focus on a single language domain, such as confrontational naming and fluency, and do not offer an ecological depiction of speech, language, and communication. Therefore, it is critical to provide tools to enable researchers, clinicians, and educators to conduct assessments of speech, language, and communication informed by ecologically reliable data. Currently, machine learning models, natural language processing techniques, signal processing methodologies, and advanced statistics collectively named Artificial Intelligence can easily automate language assessment and offer the means for more robust, accurate, and quantitative language assessments that can generalize for speakers with different linguistic backgrounds.

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

 charalampos.themistocleous@isp.uio.no (C. Themistocleous)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In our previous research, we demonstrated that a computational system with four computational pipelines for performing automated acoustic analysis, speech-to-text transcription, automatic morphosyntactic and linguistic analysis of transcripts, and machine learning could enable the identification of Swedish patients with Mild Cognitive Impairment and Alzheimer’s Disease from healthy controls (12-17) and the subtyping of patients with Primary Progressive Aphasia into variants (nonfluent PPA, semantic PPA, and logopenic PPA) (18). The machine learning model of the classification of patients with PPA was based on deep neural networks (DNN), and its performance was better than that of Random Forests, Support Vector Machines, Decision Trees, and expert clinicians’ classifications (18). We have also employed morphological and syntactic evaluation to analyze transcripts using natural language processing (NLP) and to provide automated part-of-speech (POS) tagging and syntactic parsing. For example, Themistocleous, Webster (19) analyzed connected speech productions from 52 individuals with PPA using a morphological tagger and showed differences in POS production in patients with non-fluent Primary Progressive Aphasia (nfvPPA), logopenic variant of Primary Progressive Aphasia (lvPPA), and the semantic variant of Primary Progressive Aphasia (svPPA). Also, we have employed machine learning to identify speakers with different dialects, from speech acoustics, namely prosody (20-22), vowels (23, 24), and consonants (21, 25-28). Machine learning was used to track the learning of L1 dialectal learners of the standard language variety in classrooms (29), showing the implication of using machine learning applications in diverse populations. The end-to-end automated machine learning approaches we developed for these works inspired the development of Open Brain AI to enable clinicians and researchers to provide an easy, quick, and inexpensive assessment of speech, language, and communication.

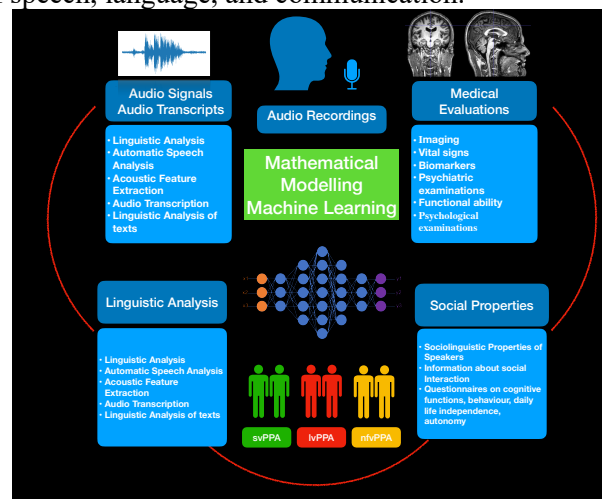


Figure 1. Multimodal Analysis of speech, language, cognition, the brain, and sociolinguistic properties in Open Brain AI.

2. Open Brain AI

Open Brain AI (<http://openbrainai.com>) is online computational platform application that aims to provide automated linguistic and cognitive assessments and tools that can be employed by researchers, clinicians, and educators to inform their daily practice and automate their tasks (30). Open Brain AI relies on Artificial Intelligence (AI) methods and tools for assessing speech, language, and communication. Clinicians can use Open Brain AI to automate spoken and written language analysis and provide informative linguistic measures of discourse and conversation to support diagnosis, prognosis, therapy efficacy evaluation, and treatment planning. Teachers can use Open Brain AI to analyze the speech and language of their students and elicit meaningful markers from essays and other materials, estimate student performance, and assess the efficacy of teaching methodologies. Researchers can produce quantitative measures of speech, language, and communication, provide results that can be compared across studies, collaborate, share ideas, and evaluate novel technologies for patient care and student learning.

3. Technologies

Open Brain AI assesses written text, speech recordings, neurolinguistic assessments, and other documents researchers, clinicians, and teachers use. These documents are first pre-processed and analyzed depending on the application. A speech-to-text component enables the multilingual transcription of speech recordings into texts. Texts are subsequently parsed using large language models, morphological taggers/parsers of the analysis of grammar, and semantic analysis tools. These tools also provide quantitative measures of the linguistic domains, such as phonology, morphology, syntax, semantics, and lexicon. Additional tools incorporate IPA transcription and acoustic analysis tools. *Open Brain AI* enables end-to-end spoken and written speech production analysis by combining the different computational pipelines to provide automated and objective linguistic measures (1, 16, 18, 19, 24, 26, 31, 32).

To achieve this Open Brain AI, incorporate language-specific Natural Language Processing (NLP) tools for analyzing written and oral texts (33, 34). These include tokenizers, which split texts into individual tokens, such as words, punctuation marks, and numbers; stemmers, to analyze words to their stems, which are the primary forms of words; lemmatizers, to identify the lemmas of words, which are the canonical or dictionary forms of words; part-of-speech (POS) taggers to assign a POS tag to each word in a sentence; named entity recognizers (NERs), to identify and classify named entities, such as people, dates, places, and organizations; parsers, to analyze the grammatical/syntactic structure of sentences; semantic role labelers to assign semantic roles to the constituents of a sentence, such as agent, patient, and recipient; and coreference resolvers: Identify and resolve coreferences in text, which are expressions that refer to the same entity (33, 34). Open Brain AI also incorporates state-of-the-art language models used to analyze texts in specific Open Brain AI applications, such as the discourse analysis of texts.

Open Brain AI provides acoustic analysis to enable the transcription of texts. Sound is first passed to Open Brain AI's acoustic analysis modules and Speech-to-Text for automatic transcription. Then, it segments speech into words and speakers and extracts acoustic measures, such as prosody and voice quality. An independent component allows the acoustic analysis tools to plot sound waveforms, spectrograms, and F0.

Machine Learning and statistical models, especially Deep Neural Network architectures, are employed to find patterns from texts and characterize language impairment. The computational can be integrated with multimodal data for research, clinical, and educational applications (Figure 1).

4. Principles

First, Open Brain AI provides access to language assessment to all individuals independently of language. For this reason, Open Brain AI offers multilingual support in different languages and language varieties (e.g., dialects). It offers automatic transcription and comprehensive grammar analysis in English, Norwegian, Swedish, Greek, and Italian. The complete grammar analysis extends to Danish, Dutch, Finnish, French, German, Portuguese, and Spanish, whereas other tools work with a wider range of languages and language varieties. Additional language varieties will be supported over time. The ability of Open Brain AI to scale concerning new languages and language variety support highlights a critical difference between computational models and traditional manual assessment techniques, which require expert knowledge for translation, standardization, and evaluation to maintain crosslinguistic psychometric properties, such as the reliability and validity of tests. The Open Brain AI platform offers access to these trained models for clinicians and teachers and makes them available.

Second, Open Brain AI does not collect data provided for analysis. Data are analyzed on the server or locally on the user's machine. Data uploaded on the server for analysis are removed immediately after processing. Information provided in Open Brain AI for accessing the site is not shared with third parties. Open Brain AI takes data privacy and security very seriously and follows industry standards to protect the confidentiality and security of personal health

information. However, no data transmission over the Internet is guaranteed to be completely secure. Therefore, Open Brain AI cannot guarantee the security of any information transmitted through the service, and you use the service at your own risk. Open Brain AI provided for healthcare purposes is not intended to replace or substitute for professional medical advice, diagnosis, or treatment.

5. Backend Infrastructure

Open Brain AI is developed using the Django framework in Python and SQL server. It is hosted on Google Cloud Run and utilizes several Google Cloud services, such as Cloud Run, Cloud Secret Manager, and Cloud Storage, to ensure consistent performance and scalability. Specifically, the system architecture includes an SQL database connection for user and post management, configurations for static file storage, email backend, and a template pack for front-end design and accessibility. The Django application of the project is deployed on Google Run, a serverless computing platform, allowing it to scale based on demand without manual server management. The application's secret key is maintained in the Cloud Secret Manager for security measures, and all static files are stored in Cloud Storage. Furthermore, the Open Brain AI backend supports the Natural AI text and sound processing models.

The backend design of Open Brain AI allows for scalability. Namely, the infrastructure supports automatic scaling depending on demand without server intervention. It provides security as secret keys are safeguarded in a dedicated location, and static files are retrieved from a scalable object storage service. The system ensures high reliability and availability.

Finally, it is flexible, as the configuration via Django settings and environment variables enables customization to address specific user requirements or scenarios. For example, the backend infrastructure allows users to access the online platform from different devices, e.g., Computers, Tablets, and Mobile phones, without specific configuration.

6. Applications

Open Brain AI offers applications for three distinct groups: researchers, clinicians, and teachers. Researchers may want to access raw data to analyze further, whereas applications for clinicians and teachers provide applications that can analyze the patients and students in clinical or teaching environments.

6.1. Research Applications

Computational Discourse Analysis Application. Discourse provides multidomain data and information on language production, perception, planning, and cognition (35-38). Thus, discourse can explain brain functioning and provide recommendations on whether there is evidence for a possible speech, language, and communication impairment. Open Brain AI's discourse module employs large AI language Models to analyze texts and metrics from discourse, semantics, syntax, morphology, phonology, and lexical distribution elicited using NLP and machine learning. Subsequently, it combines its internal knowledge of the world based on its training to provide a comprehensive analysis of speech, language, and communication for the textual transcripts based on quantified measures from part of speech analysis, syntactic phrase identification, semantic analysis (e.g., named entity recognition), and linguistic distribution.

Linguistic Measures Application. Open Brain AI provides objective measures of written speech production that clinicians, teachers, and researchers compare a patient with a targeted population concerning discourse, phonology, morphology, syntax, semantics, and lexicon (18, 39-45). Specifically, this module analyzes the text or the transcripts from the speech-to-text module and conducts measures on the following linguistic domains:

1. Phonology: It elicits measures, such as the number and type of syllables and the ratio of syllables per word.
2. Morphology: It provides counts and their ratio of parts of speech (e.g., verbs, nouns, adjectives, adverbs, and conjunctions) concerning the total number of words.
3. Syntax: It provides counts and their ratio of syntactic constituents (e.g., noun phrases and verb phrases).
4. Lexical Measures: it provides measures such as the number of words, hapax legomena, and Type Token Ratio (TTR) measures.
5. Semantic Measures: It provides counts and their ratio of semantic entities in the text (e.g., persons, dates, and locations).
6. Readability Measures: It provides readability measures about the text and grammar.

Recordings are analyzed using different applications.

1. *Automatic transcription.* Open Brain AI employs Automatic Speech Recognition (ASR) to process audio files. The process begins by uploading an audio file on Open Brain AI. The transcription of the audio file is conducted using speech-to-text. The system is modular, so it employs different speech-to-text applications.
2. *Linguistic Analysis & AI Discourse Analysis.* The transcripts are further analyzed using the automatic morphosyntactic analysis and by a GPT3 Large Language Model. The module combines the text and metrics from discourse, semantics, syntax, morphology, phonology, and linguistic distribution.
3. *Acoustic Analysis.* The spoken speech assessment module provides transcription and grammatical analysis of these transcripts. The grammatical study replicates that of written speech productions. Namely, it offers total phonology, morphology, syntax, semantics, and lexicon scores.
4. *Speakers Segmentation.* The Open Brain AI platform allows splitting the audio, dividing patients from clinicians in the audio recordings. When there is more than one speaker in the audio file. The diarization output is exported as a coma delimited file or Praat TextGrid for researchers wanting to perform acoustic analysis.
5. *Word Alignment and Pause Detection.* The platform enables the alignment of words with the sound wave to allow further acoustic analysis for measures, such as word duration, and the elicitation of the specific acoustic measures on acoustic production. The automatically segmented sounds are exported in various formats, such as Praat TextGrids.

6.2. Clinical Applications

The clinical toolkit provides scoring tools and comprises four primary tools:

Picture Description Task. A picture description task is a standard assessment tool for evaluating individuals with aphasia or other language disorders. In a picture description task, the patient is presented with a picture and is asked to describe it in as much detail as possible. The picture typically depicts a scene with multiple elements, actions, and interactions to allow for various linguistic constructions and vocabulary. The task assesses the patient's ability to produce spontaneous speech. It can reveal difficulties in forming grammatically correct sentences, using appropriate vocabulary, or maintaining coherence. Open Brain AI incorporates tools to conduct the picture description task and evaluate the content (what the patient says) and the structure (how they say it). This can provide insights into the type and severity of the aphasia. Patients with distinct types of aphasia (e.g., Broca's, Wernicke's, Global) may produce different patterns of errors and difficulties in the picture description task. Repeated assessments using Open Brain AI over time can track a patient's recovery and the effectiveness of therapeutic interventions.

Automatic conversion to the International Phonetic Alphabet. The tool converts words written in standard orthography into the International Phonetic Alphabet with extensive support to languages and language varieties and provides measures of their sounds in the transcribed texts.

Spelling Scoring App. The evaluation of spelling is a complex, challenging, and time-consuming process. It relies on comparing letter-to-letter, the words spelled by the patients to the target

words, using the Levenshtein Distance. It processes both words and non-words (1). It specifically Themistocleous, Neophytou (1) developed a spelling distance algorithm that automatically compares the inversions, insertions, deletions, and transpositions required to make the target word and response identical (1, 46). To determine phonological errors in patients with aphasia, we have developed a phonological distance algorithm that quantifies phonological errors automatically.

Phonological Scoring Tool. The tool converts the target and response words into the International Phonetics Alphabet, compares their differences using the Levenshtein Distance, and provides scores changes, namely deletions, insertions, transpositions, and substitutions. It processes both words and non-words.

Semantics Scoring Tool. The semantic distance scoring tool employs embeddings to score naming tasks involving semantic memory access (47-49).

6.3. Educational Applications

Open Brain AI provides infrastructure for educational applications. The underlying system for automatic spoken and written language analysis is being employed to assess students' performances in different settings, to track students' performance over time, and to assess teaching methods' efficacy. The Open Brain Education platform incorporates phonology, semantics, spelling, and essay assessment scoring applications.

Essay Assessment. This application, powered by advanced language models, evaluates Content and Argumentation by examining the thesis statement's clarity and strength, logical argument progression, depth of analysis, and evidence backing claims. The tool reviews essays' structure, ensuring logical flow, cohesion, and the presence of a clear introduction, body, and conclusion. It checks for Grammar and Mechanics, including punctuation, spelling, sentence construction, verb tense adherence, style, and voice for uniqueness, consistency, and appropriateness. The tool provides feedback on essay clarity and precision, flagging ambiguous language or jargon. Lastly, it highlights potential grammatical and stylistic errors.

6.4. Offline Open Brain AI Applications

Accurate diagnosis and prognosis are vital for personalized intervention in speech, language, and communication disorders, enhancing the quality of life (50, 51). Prognosis involves predicting a patient's trajectory and outcomes (52). Offline Open Brain AI harnesses machine learning and multimodal data to distinguish patients with MCI from healthy controls (12, 15, 16), students with different learning needs (19, 53), and speakers with contextualized speech patterns based on age, gender, and sociolinguistic factors (21, 23, 24, 26-28) (Figure 1). Computational tools offer a comprehensive analysis of patients and students (19) and can be extended to naturalistic speech analysis (54).

7. Conclusions

Speech and written language are distinct communication modalities, and accurate diagnosis and prognosis of speech, language, and communication disorders and student assessment in classroom settings require understanding their unique characteristics. Continued collaboration between experts in education, clinical research, computer science and AI will enhance our understanding and capabilities in assessing and treating neurocognitive disorders and support students learning, improving the lives of affected individuals. By considering the factors above and leveraging technological advancements offered by Open Brain AI, clinicians, educators, and researchers can develop effective intervention plans in the clinic and the classroom and make informed prognostic judgments. Ultimately, Open Brain AI empowers clinicians, educators, and researchers to deliver effective and inclusive support to patients with speech, language, and communication impairments and language students, improving their overall well-being and learning.

References

1. Themistocleous C, Neophytou K, Rapp B, Tsapkini K. A tool for automatic scoring of spelling performance. *Journal of Speech, Language, and Hearing Research*. 2020;63:4179-92.
2. de Aguiar V, Zhao Y, Ficek BN, Webster K, Rofes A, Wendt H, et al. Cognitive and language performance predicts effects of spelling intervention and tDCS in Primary Progressive Aphasia. *Cortex*. 2020;124:66-84.
3. Neophytou K, Wiley RW, Rapp B, Tsapkini K. The use of spelling for variant classification in primary progressive aphasia: Theoretical and practical implications. *Neuropsychologia*. 2019;133:107157.
4. Tsapkini K, Webster KT, Ficek BN, Desmond JE, Onyike CU, Rapp B, et al. Electrical brain stimulation in different variants of primary progressive aphasia: A randomized clinical trial. *Alzheimers Dement (N Y)*. 2018;4:461-72.
5. Purcell JJ, Rapp B. Local response heterogeneity indexes experience-based neural differentiation in reading. *Neuroimage*. 2018;183:200-11.
6. Rapp B, Fischer-Baum S. Uncovering the cognitive architecture of spelling. *The Handbook of Adult Language Disorders: Psychology Press*; 2015. p. 59--86.
7. Fischer-Baum S, Rapp B. The analysis of perseverations in acquired dysgraphia reveals the internal structure of orthographic representations. *Cognitive neuropsychology*. 2014(ahead-of-print):1--29.
8. Kaplan E, Goodglass H, Weintraub S. *Boston Naming Test*. Austin, TX: Pro-ed; 2001.
9. Kertesz A. *Western aphasia battery-revised (WAB-R)*. New York: Pearson; 2006. null p.
10. Goodglass H, Kaplan E. *Boston diagnostic aphasia examination (BDAE)*. Philadelphia: Lea & Febiger; 1983.
11. Kay J, Lesser R, Coltheart RM. *PALPA. Psycholinguistic Assessments of Language Processing in Aphasia*. New York: Psychology Press; 1992.
12. Themistocleous C, Eckerström M, Kokkinakis D. Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls. *PLoS One*. 2020;15(7):e0236009.
13. Themistocleous C, Eckerström M, Kokkinakis D. Automated speech analysis enables MCI diagnosis. *ExLing 2020*. 2020:201.
14. Fraser KC, Linz N, Li B, Lundholm Fors K, Rudzicz F, König A, et al. Multilingual prediction of Alzheimer's disease through domain adaptation and concept-based language modelling. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019:3659-70.
15. Themistocleous C, Kokkinakis D. Speech and Mild Cognitive Impairment detection. In: Botinis A, editor. *Proceedings of the 9th Tutorial & Research Workshop on Experimental Linguistics (ExLing2019)* 2019. p. 201.
16. Themistocleous C, Eckerström M, Kokkinakis D. Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural Networks. *Frontiers in Neurology*. 2018;9:975.
17. Fraser KC, Lundholm Fors K, Eckerström M, Themistocleous C, Kokkinakis D. Improving the Sensitivity and Specificity of MCI Screening with Linguistic Information. *Proceedings of the LREC 2018 Workshop "Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2)"*. 2018(2015):19-26.
18. Themistocleous C, Ficek B, Webster K, den Ouden D-B, Hillis AE, Tsapkini K. Automatic Subtyping of Individuals with Primary Progressive Aphasia. *Journal of Alzheimer's Disease*. 2021;79:1185-94.
19. Themistocleous C, Webster K, Afthinos A, Tsapkini K. Part of Speech Production in Patients With Primary Progressive Aphasia: An Analysis Based on Natural Language Processing. *American Journal of Speech-Language Pathology*. 2020:1-15.

20. Themistocleous C. Seeking an Anchorage. Stability and Variability in Tonal Alignment of Rising Prenuclear Pitch Accents in Cypriot Greek. *Language and Speech*. 2016;59(4):433-61.
21. Themistocleous C, Savva A, Aristodemou A, editors. Effects of stress on fricatives: Evidence from Standard Modern Greek. *Interspeech 2016*; 2016; San Francisco, September 8-12.
22. Themistocleous C. Edge-Tone Effects and Prosodic Domain Effects on Final Lengthening. *Linguistic Variation*. 2014;14(1):129-60.
23. Themistocleous C. The Nature of Phonetic Gradience across a Dialect Continuum: Evidence from Modern Greek Vowels. *Phonetica*. 2017;74(3):157-72.
24. Themistocleous C. Dialect classification using vowel acoustic parameters. *Speech Communication*. 2017;92:13-22.
25. Themistocleous C, Fyndanis V, Tsapkini K. Sonorant spectra and coarticulation distinguish speakers with different dialects. *Speech Communication*. 2022:1-14.
26. Themistocleous C. Dialect Classification From a Single Sonorant Sound Using Deep Neural Networks. *Frontiers in Communication*. 2019;4:1-12.
27. Themistocleous C. Effects of two linguistically proximal varieties on the spectral and coarticulatory properties of fricatives: Evidence from Athenian Greek and Cypriot Greek. *Frontiers in Psychology*. 2017;8(NOV).
28. Themistocleous C. The bursts of stops can convey dialectal information. *The Journal of the Acoustical Society of America*. 2016;140(4):EL334-EL9.
29. Grohmann K, Papadopoulou E, Themistocleous C. Acquiring Clitic Placement in Bilectal Settings: Interactions between Social Factors. *Frontiers in Communication*. 2017;2(5).
30. Themistocleous C. Computational Language Assessment: Open Brain AI. *arXiv*. 2023;2306.06693:1-17.
31. Themistocleous C, Webster K, Tsapkini K. Effects of tDCS on Sound Duration in Patients with Apraxia of Speech in Primary Progressive Aphasia. *Brain Sciences*. 2021;11(3).
32. Themistocleous C, Kokkinakis D. THEMIS-SV: Automatic classification of language disorders from speech signals. *ESOC 2018: European Stroke Organisation Conference; Gothenburg2018*.
33. Manning C. *Last Words*. Computational Linguistics and Deep Learning. Association for Computational Linguistics. 2015.
34. Clark A, Fox C, Lappin S, Blackwell Reference Online (Online service). *The handbook of computational linguistics and natural language processing*. Chichester, West Sussex ; Malden, MA: Wiley-Blackwell;; 2010. Available from: http://www.blackwellreference.com/subscriber/uid=3/book?id=g9781405155816_9781405155816.
35. Stark BC, Bryant L, Themistocleous C, den Ouden D-B, Roberts AC. Best practice guidelines for reporting spoken discourse in aphasia and neurogenic communication disorders. *Aphasiology*. 2022:1-24.
36. Stark Brielle C, Dutta M, Murray Laura L, Bryant L, Fromm D, MacWhinney B, et al. Standardizing Assessment of Spoken Discourse in Aphasia: A Working Group With Deliverables. *Am J Speech Lang Pathol*. 2020:1-12.
37. Cunningham KT, Haley KL. Measuring Lexical Diversity for Discourse Analysis in Aphasia: Moving-Average Type-Token Ratio and Word Information Measure. *J Speech Lang Hear Res*. 2020;63(3):710-21.
38. Fyndanis V, Arcara G, Capasso R, Christidou P, De Pellegrin S, Gandolfi M, et al. Time reference in nonfluent and fluent aphasia: a cross-linguistic test of the PAST Discourse Linking Hypothesis. *Clinical linguistics & phonetics*. 2018:1--21.
39. Stockbridge MD, Matchin W, Walker A, Breining B, Fridriksson J, Hickok G, et al. One cat, Two cats, Red cat, Blue cats: Eliciting morphemes from individuals with primary progressive aphasia. *Aphasiology*. 2021;35(12):1-12.
40. Breining BL, Lala T, Martínez Cuitiño M, Manes F, Peristeri E, Tsapkini K, et al. A brief assessment of object semantics in primary progressive aphasia. *Aphasiology*. 2015;29(4):488--505.

41. Tsapkini K, Frangakis C, Gomez Y, Davis C, Hillis AE. Augmentation of spelling therapy with transcranial direct current stimulation in primary progressive aphasia: Preliminary results and challenges. *Aphasiology*. 2014;28(8-9):1112-30.
42. Miceli G, Capasso R, Caramazza A. The interaction of lexical and sublexical processes in reading, writing and repetition. *Neuropsychologia*. 1994;32(3):317--33.
43. Badecker W, Hillis A, Caramazza A. Lexical morphology and its role in the writing process: evidence from a case of acquired dysgraphia. *Cognition*. 1990;35(3):205--43.
44. Hillis AE, Rapp B, Romani C, Caramazza A. Selective impairment of semantics in lexical processing. *Cognitive Neuropsychology*. 1990;7(3):191-243.
45. Hillis AE, Caramazza A. The graphemic buffer and attentional mechanisms. *Brain and language*. 1989;36(2):208--35.
46. Neophytou K, Themistocleous C, Wiley R, Tsapkini K, Rapp B. Understanding and classifying the different variants of Primary Progressive Aphasia based on spelling performance. *Frontiers in Human Neuroscience*. 2018;12.
47. Sebastian R, Thompson CB, Wang NY, Wright A, Meyer A, Friedman RB, et al. Patterns of Decline in Naming and Semantic Knowledge in Primary Progressive Aphasia. *Aphasiology*. 2018;32(9):1010-30.
48. Riello M, Faria AV, Ficek B, Webster K, Onyike CU, Desmond J, et al. The Role of Language Severity and Education in Explaining Performance on Object and Action Naming in Primary Progressive Aphasia. *Frontiers in Aging Neuroscience*. 2018;10:346.
49. Afthinos A, Themistocleous C, Herrmann O, Fan H, Lu H, Tsapkini K. The Contribution of Working Memory Areas to Verbal Learning and Recall in Primary Progressive Aphasia. *Frontiers in Neurology*. 2022;13:1-11.
50. Grasemann U, Peñaloza C, Dekhtyar M, Miikkulainen R, Kiran S. Predicting language treatment response in bilingual aphasia using neural network-based patient models. *Scientific Reports*. 2021;11(1):10497.
51. Johnson JP, Ross K, Kiran S. Multi-step treatment for acquired alexia and agraphia (Part I): efficacy, generalisation, and identification of beneficial treatment steps. *Neuropsychological Rehabilitation*. 2019;29(4):534-64.
52. Diogo VS, Ferreira HA, Prata D, Alzheimer's Disease Neuroimaging I. Early diagnosis of Alzheimer's disease using machine learning: a multi-diagnostic, generalizable approach. *Alzheimers Res Ther*. 2022;14(1):107.
53. Themistocleous C, Ficek B, Webster K, den Ouden D-B, Hillis AE, Tsapkini K. Automatic subtyping of individuals with Primary Progressive Aphasia. *bioRxiv*. 2020:2020.04.04.025593.
54. Themistocleous C. A review of discourse and conversation impairments in patients with dementia arXiv preprint arXiv:221107971. 2022.

Profiles for Swedish as a Second Language: Lexis, Grammar, Morphology

Elena Volodina¹, David Alfter¹ and Therese Lindström Tiedemann²

¹University of Gothenburg, Sweden

²University of Helsinki, Finland

Abstract

This article gives a short introduction to the Swedish Second Language Profile, a tool that visualizes language in Swedish learner corpora from different angles, such as vocabulary, grammar and morphology. The tool is aimed at research on Second Language Acquisition, development of NLP models, teaching of Swedish as a second language, automatic approaches for second language teaching and learning, and at a number of other fields.

Keywords

Second Language Profile for Swedish (SweL2P), Lexical profile, Grammatical profile, Morphological profile

1. Introduction

Learner corpus researchers, NLP researchers, as well as Digital Humanities and Social Sciences in general rely on access to various data sets for empirical analysis, statistical insights and/or for model building. However, interpretation of data is a non-trivial task and there is a need for data visualization tools [1, 2, 3, 4]. One such attempt is the Swedish Second Language (L2) profile¹ - a project setting up the first digital tool allowing users to explore written Swedish learner language from a linguistic point of view.

The aim of the project was to describe the learning paths of Swedish as a second language with a focus on vocabulary and grammar. To do this, we analysed two text collections - course books and learner essays - for various patterns and their statistics at different levels of proficiency, and organized them into so-called profiles (section 2).

Access to the profiles facilitates studies into, among others, which vocabulary or grammar is central for a certain level of proficiency, which linguistic features can be discriminative at different levels, in which order the various linguistic aspects are introduced into reading materials and into writing, and many others. The practical application of the profiles can take many different forms. For Intelligent Computer-Assisted Language Learning they facilitate, for example, generation of learning materials (exercises, test items, etc.) of appropriate linguistic complexity for given

Huminfra Conference 2024, Gothenburg, 10–11 January 2024.

✉ elena.volodina@svenska.gu.se (E. Volodina); david.alfter@gu.se (D. Alfter);

therese.lindstromtiedemann@helsinki.fi (T. Lindström Tiedemann)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹The tool is a by-product of the research project *Development of lexical and grammatical competences in immigrant Swedish*, <https://spraakbanken.gu.se/en/projects/l2profiles>

Multi Word Expressions [🔗](#)

Search word Word Class Saldo Word Class Type 1: Syntactic cont... Type 2: Lexical catego... Type 3: Verbal subcat...

A1 A2 B1 B2 C1 C2

Tables Graphs Statistics Download Clear all

Extend columns in the table Show only first occurrence [Tables - description](#) [Filters - description](#)

Figure 1: Filters for Multiword Expressions (Lexical profile)

CEFR level ↑↓	Word ↑↓	Lemgram ↑↓	Sense ↑↓	Word Class ↑↓	Saldo Word Class ↑↓	Receptive ↑↓ 🔗	Productive ↑↓ 🔗
A1	jeans	jeans..nn.1	jeans..1	Noun (NN)	Noun (nn)	0.77 (2)	0.00 (0)
A1	kläder	kläder..nn.1	kläder..1	Noun (NN)	Noun (nn)	3.09 (8)	7.32 (3)
A1	shorts	shorts..nn.1	shorts..1	Noun (NN)	Noun (nn)	1.16 (3)	0.00 (0)
A2	makaroner	makaroner..nn.1	makaroner..1	Noun (NN)	Noun (nn)	0.44 (3)	0.00 (0)

Figure 2: Table view for Sen*Lex (Lexical profile), filtered for 'Always plural' nouns

levels; or to classify learner texts according to proficiency levels reached using information of the vocabulary or grammar scope per level as input features. For Second Language Acquisition (SLA), it is possible to study the relation between vocabulary acquisition and previous morphological (derivational) knowledge; the order of introduction of various grammatical structures at different levels, and many others.

As such, the users can be researchers, developers, language assessors, course book writers, teachers or learners. The fact that any filtering allows download of a filtered subset of structures makes us believe that such lists will be used on their own as spin-off resources.

2. SweL2P - overview

The Swedish Second Language Profile (SweL2P) features the following profiles:

- A lexical profile, organized into subprofiles by words, multiword expressions, adjectival declensions and adjectival & adverbial structure;
- A grammatical profile, including noun patterns and verb patterns;
- A morphological profile, organized into word families and morpheme families.

2.1. Source data

SweL2P is empirical in nature since it is based on data from two corpora: **COCTAILL** – a corpus consisting of coursebooks used for teaching Swedish to L2 students [5]; and **SweLL-pilot** – a corpus of learner essays written by L2 learners of Swedish [6]. Texts in both corpora have

Lexical profile | **Grammatical profile** | Morphological profile

Verb patterns

Pattern (description) | Pattern (structure) | Tense | Mood | Voice | Form

A1 | A2 | B1 | B2 | C1 | C2

Extend columns in the table
 Show only first occurrence
 Tables - descr

Tense
 Mood
 Voice
 Form

Total rows : 18

CEFR level	Pattern (description)	Pattern (structure)	Tense	Mood	Voice	Receptive	Productive
A2	Past tense (preterite) - ...	Simple	Past (Preterite)	Indicative	Passive	0.03 (4)	0.04 (1)
A2	Perfect tense - s-passive...	Complex	Perfect	Indicative	Passive/Deponent	0.01 (2)	0.00 (0)
A2	Perfect tense - s-passive...	Complex	Perfect	Indicative	Passive	0.01 (2)	0.00 (0)
B1	Perfect tense - deponent ...	Complex	Perfect	Indicative	Deponent	0.04 (7)	0.00 (0)

Figure 3: Table view for Verb Patterns (Grammatical profile)

been graded with CEFR [7] levels by experts, starting from A1 (beginner) to C1 (advanced), the C2 level being largely absent. COCTAILL has been used to get an approximation of the vocabulary and grammar L2 learners meet when reading, and therefore what they are expected to understand **receptively**. SweLL-pilot has been used to get an approximation of the vocabulary and grammar L2 learners are able to produce actively when writing, and therefore represents learners' **productive** abilities.

The two corpora have been used to create a sense-disambiguated word list, Sen*Lex [8, 9, 10], as the main input for the Lexical profile. Sen*Lex was subsequently manually enriched with morphological analysis giving rise to the CoDeRoMo resource [11] of which an updated version has been used as the main input for the Morphological profile. Furthermore, both corpora have been semi-automatically parsed for verb and noun patterns that currently constitute the core of the Grammatical profile.

2.2. Annotation

To prepare the data, we combined **automatic** processing [12, 13, 14] and **manual** annotation [11, 15, 16]. To support manual annotation and visualization of the process, the tool LEGATO [17] was implemented and a range of resources for Swedish were reused to inter-link the lexicographic and other information for each item on the Sen*Lex list [18, 19, 20].

Manual annotation was used for: (1) classification of multiword expressions (MWEs) into subtypes [15] giving rise to a MWE subprofile under the Lexical profile; (2) morphological analysis of lemmas and word-formation patterns [11, 21] creating the basis for the Morpheme family and Word family under the Morphological profile; and (3) setting up regular expressions to extract noun and verb patterns from linguistically annotated texts used for Verb and Noun patterns under the Grammatical profile [16, 22].

2.3. User Interface

The User Interface for browsing the SweL2P² has some features that are shared by all its modules and subprofiles, such as a filter for CEFR levels, a possibility to enter a search item (except in the grammatical profile) and an option to see frequencies and samples from receptive/productive data for one's search. Some other features are specific for a (sub)profile in question.

Filters appear at the top of the page, providing a set of filters for each subprofile, e.g. *MWE Types 1-3* in Figure 1. The resource **can be explored** using several views: Table, Graphical and Statistical. The **Table** view (Figure 2 and 3) lists all items with associated information. Columns contain descriptive information, among others, a clickable word (e.g. 'jeans' in Figure 2) that opens a link to an entry with this item at <https://svenska.se/> and manual morphological analysis, as well as clickable receptive and productive (relative and absolute) frequencies that open a corpus search tool **Korp** [23] containing hits with those items. The range of the columns depends on the profile – notably, for the *Grammatical profile*, 'pattern' is listed instead of 'word' (Figure 3). **Graphical** and **Statistical** views summarize the statistics and distribution of various features for the current selection in the two sources, receptive and productive. For example, Figure 4 shows a graph view over the use of the Past tense (Sw. preteritum) across levels and a related table with the statistics; while Figure 5 demonstrates an excerpt from the statistics over the *Morpheme family* subprofile. In Figure 5, instead of graphs and distributions, we rather see counts in terms of types, tokens and type-token ratios per filter category so that we can study a statistical break-down of each selection contrasting receptive and productive competences.

The entire dataset or filtered data selection can be **downloaded** from the website.

3. Profiles

SweL2P covers three distinct areas of learner language – lexis, grammar and morphology.

3.1. Lexical profile

The *Lexical profile* includes four subprofiles, each representing a subsection of *Sen*Lex*. **Adjectival declension** features one specialized filter for *declension* with four declensions: 1st, 2nd, suppletive and indeclinable. Through it, we can see how many of the adjectives in the two corpora that belong to the different declensions. The assignment of items into the different declensions was made semi-automatically. Similarly, the related subsection **Adjectival and Adverbial structure** presents adjectives and adverbs by types of *comparison* (morphological, periphrastic, both or unclassified) and *regularity* (regular, regular with umlaut, irregular). The items were semi-automatically categorized. **Multiword Expressions** allow the user to filter MWEs for (1) *syntactic construction* (contiguous or non-contiguous); (2) *lexical sub-categories*, including some which are closely related to parts of speech (POS), e.g. nominal, adjectival and verbal; (3) *verbal sub-categories*, with 4 types, e.g. particle verbs. MWEs were identified automatically, but the classification of MWEs into subgroups was done manually [15].

²<https://spraakbanken.gu.se/larkalabb/svlp>, login: demo

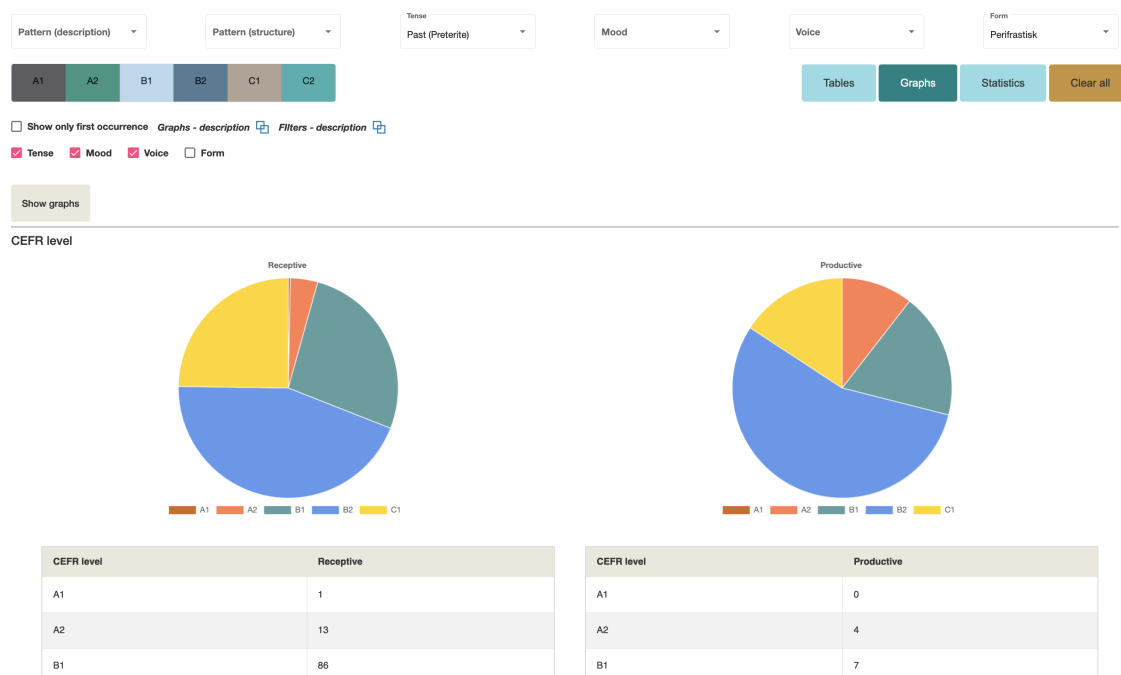


Figure 4: Graphical view over Verb patterns (Grammatical profile): Statistics of (simple) past tense (Sw. preteritum) over CEFR levels

CEFR level	Receptive (Type)	Productive (Type)	Receptive (Token)	Productive (Token)	Receptive (TTR)	Productive (TTR)
A2	5153	715	16395	1870	31.43	38.24
B1	10365	1189	27296	2835	37.97	41.94
B2	10961	1318	22606	2778	48.49	47.44

Morpheme Category	Receptive (Type)	Productive (Type)	Receptive (Token)	Productive (Token)	Receptive (TTR)	Productive (TTR)
Derivational prefix	1941	337	5166	620	37.57	54.35
Root morpheme	21448	3109	70004	8954	30.64	34.72
Derivational suffix	9133	1491	23780	3246	38.41	45.93

Figure 5: Statistical view (Morpheme family)

Sen*Lex features a list of lexical items from the two corpora, ordered by sense-disambiguated lemmings [8, 9, 10] (see an example in Figure 2). The three specific filters, all automatically assigned to the items but manually checked and sometimes corrected, include: (1) Noun declension; (2) Gender; and (3) Conjugation.

Verb patterns	Noun patterns
<i>Pattern</i> – 40 unique patterns, e.g. Imperative	<i>Pattern</i> – 28 head patterns, 81 unique patterns, e.g. Nsg (Jag har körkort)
<i>Pattern structure</i> – 3 categories, e.g. Elliptic	<i>Pattern structure</i> – 2 categories, e.g. Simple noun phrase
<i>Tense</i> – 7 categories, e.g. Pluperfect	<i>Clause position</i> – 2 categories, e.g. Initial
<i>Mood</i> – 4 categories, e.g. Subjunctive	<i>Definiteness</i> – 2 categories, e.g. Definite
<i>Voice</i> – 5 categories, e.g. Passive/Deponent	<i>Gender</i> – 3 categories, e.g. Neuter
<i>Form</i> – 3 categories, e.g. S-form	<i>Number</i> – 2 categories, e.g. Plural
	<i>Attribute</i> – 4 categories, e.g. Relative clause
	<i>Article</i> – 5 categories, e.g. Definite affix
	<i>Other definite attributes</i> – 5 categories, e.g. Demonstrative pronoun

Table 1
Overview of linguistic filters for Noun and Verb patterns (Grammatical profile)

3.2. Grammatical profile

The `Grammar profile` [16] features **Verb patterns** and **Noun patterns**. Both are based on structures parsed from the two source corpora, and are grouped under different patterns. The two subprofiles have the same general filtering options, such as CEFR levels and hits in receptive and productive data. However, they also have a number of unique linguistic filters (see Table 1 for details), some of which are based on automatic annotation.

3.3. Morphological profile

The `Morphological profile` is based on an updated version of `CoDeRoomor` [11] - a list containing 16 230 sense-based morphologically analyzed lemgrams organized into 4 986 morpheme families, of which 4 429 are word families (i.e. root families). To create `CoDeRoomor`, we used `Sen*Lex` list. A team of annotators analyzed each item on the list for their constituent morphemes (e.g. prefix, root) and word formation mechanism (e.g. compounding, derivation). The original `CoDeRoomor` resource can be freely downloaded as csv. or xlsx. files³ or can be browsed and downloaded in a slightly updated form in the `SweL2P`.²

The `Morphological profile` includes **Morpheme family** and **Word family**. They can both be filtered for a morpheme, word, word class and word formation, and `Morpheme family` can additionally be filtered by `morpheme category`. Searching for a particular morpheme (e.g. root `bröd`), shows all lemgrams in the family (once for each level where they appear). Ticking `Only first occurrence`, shows unique items and their count. Selecting a proficiency level(s) shows how that morpheme/word family is represented there.

Morphemic analysis of the vocabulary facilitates various studies into language learning patterns, some examples presented in Volodina et al. [24, 25]. The resource is useful in a broader context, e.g. in NLP for training models for word segmentation into morphemes; in ICALL for the generation of exercises aimed at word-formation patterns, etc.

³<https://spraakbanken.gu.se/en/resources/coderoomor>

4. Potential for research and teaching

The SweL2P has great potential for both research and teaching. Some initial research possibilities have to been explored in relation to morphology in [25], [24] and [26], for MWEs in [15], for place names in [27], and in relation to some initial grammatical patterns in [16]. The resource can be used by researchers but also by teachers to explore what is common or rare at different levels in coursebooks and/or learner essays. Coursebooks can be taken as a proxy for input and help us explore language acquisition in relation to usage-based theories (e.g. [28]). Similarly, a teacher who finds that students find it challenging to learn how to use e.g. double definiteness (Sw. *den gula bilen* [DEF yellow-DEF car-DEF] 'the yellow car') or a particular subjunction can use the resource to explore the actual empirical usage. Thanks to the fact that statistics from both corpora are presented in parallel there is an excellent possibility to ascertain if a certain challenge is likely to be related to a lack of examples in input. Furthermore, since the resource also contains links to the searches in Korp this can then easily be explored further also in reference corpora with the same annotation.

5. Concluding remarks and future prospects

Language learning profile resources exist predominantly for English, e.g. English profile [29], CEFR-J [30] and Pearson's GSE Teacher Toolkit [31]. Most languages have nothing similar, the L2 Estonian Teacher's Tools [32] being one of the first non-English profiles. Even when the existing profiles have been based on empirical corpus data, this data is not openly provided, rendering them rather prescriptive. SweL2P takes a descriptive view of the language and provides access to the empirical evidence, i.e. all corpus hits and statistics of actual usage. It lets users zoom in on actual data and draw their own conclusions based on the empirical data with the help of visualizations. Including both receptive and productive frequencies side-by-side the resource gives a more nuanced picture of language learning. Through that and the special efforts invested in the visualization of the data, the SweL2P tool is more readily appropriate for research on SLA than any predecessor known to us. Furthermore, the inclusion of links to searches in Korp makes it easy to compare with other corpora from different varieties of Swedish. Finally, the open nature of the resource makes it highly useful for future learning apps, for training of automatic tools, and for teaching.

In the future, we envisage efforts invested into the disambiguation of morphemes, adding more patterns to the grammar profile (e.g. word order), adding reference corpus statistics, expanding visualization techniques (e.g. word clouds), and user upload of data for analysis.

Acknowledgments

Work on the Swedish L2 Profile has been supported by a research grant from the Swedish Riksbankens Jubileumsfond *Development of lexical and grammatical competences in immigrant Swedish*, P17-0716:1 (2018–2021). Work on the article has been supported by Nationella språkbanken and HUMINFRA, both funded by the Swedish Research Council (2018–2024, contract 2017-00626; 2022–2024, contract 2021-00176) and their participating partner institutions.

References

- [1] M. Islam, S. Jin, An overview of data visualization, in: 2019 International Conference on Information Science and Communications Technologies (ICISCT), IEEE, 2019, pp. 1–7.
- [2] C. Sievert, Interactive web-based data visualization with R, plotly, and shiny, CRC Press, 2020.
- [3] S. Coulange, M.-P. Jouannaud, C. Cervini, M. Masperi, From placement to diagnostic testing: Improving feedback to learners and other stakeholders in SELF (Système d’Evaluation en Langues à visée Formative), *Language Learning in Higher Education* 10 (2020) 195–205.
- [4] M. L. Waskom, Seaborn: statistical data visualization, *Journal of Open Source Software* 6 (2021) 3021.
- [5] E. Volodina, I. Pilán, S. R. Eide, H. Heidarsson, You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language, in: *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*, Linköping University Press, 2014.
- [6] E. Volodina, I. Pilán, I. Enström, L. Llozhi, P. Lundkvist, G. Sundberg, M. Sandell, SweLL on the rise: Swedish learner language corpus for European reference level studies, in: *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia, 2016.
- [7] C. of Europe, Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion Volume with new descriptors, Council of Europe Publishing, 2020.
- [8] T. François, E. Volodina, I. Pilán, A. Tack, SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 2016, pp. 213–219.
- [9] E. Volodina, I. Pilán, L. Llozhi, B. Degryse, T. François, SweLLex: second language learners’ productive vocabulary, in: *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, 2016, pp. 76–84.
- [10] D. Alfter, Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective, *Data Linguistica* 31, University of Gothenburg, 2021.
- [11] E. Volodina, Y. A. Mohammed, T. Lindström Tiedemann, CoDeRoomor: A new dataset for non-inflectional morphology studies of Swedish, in: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 2021, pp. 178–189.
- [12] L. Borin, M. Forsberg, M. Hammarstedt, D. Rosén, R. Schäfer, A. Schumacher, Sparv: Språkbanken’s corpus annotation pipeline infrastructure, in: *The Sixth Swedish Language Technology Conference (SLTC)*, Umeå University, 2016, pp. 17–18.
- [13] L. Nieto Piña, Splitting rocks: Learning word sense representations from corpora and lexica, *Data Linguistica* 30, University of Gothenburg, 2019.
- [14] E. Volodina, D. Alfter, T. Lindström Tiedemann, M. S. Lauriala, D. H. Piipponen, Reliability of automatic linguistic annotation: native vs non-native texts, in: *Selected papers from the CLARIN Annual Conference 2021*, Linköping University Electronic Press (LiU E-Press), 2022.
- [15] T. Lindström Tiedemann, D. Alfter, Y. A. Mohammed, D. Piipponen, B. Silén, E. Volodina,

Multiword expressions in Swedish as second language: taxonomy, annotation and initial results, in: Multiword expressions in language resources. Linguistic, lexicographic and computational considerations, acc.

- [16] T. Lindström Tiedemann, Y. A. Mohammed, E. Volodina, Swedish Grammar profiling for empirical L2 research and teaching (subm.).
- [17] D. Alfter, T. Lindström Tiedemann, E. Volodina, LEGATO: A flexible lexicographic annotation tool, in: Proceedings of the 22nd Nordic Conference on Computational Linguistics, 2019, pp. 382–388.
- [18] L. Borin, M. Forsberg, L.-J. Olsson, J. Uppström, The open lexical infrastructure of Språkbanken., in: LREC, 2012, pp. 3598–3602.
- [19] L. Borin, M. Forsberg, L. Lönngrén, SALDO: a touch of yin to WordNet’s yang, Language resources and evaluation 47 (2013) 1191–1211.
- [20] L. Borin, M. Forsberg, Swesaurus; or, The Frankenstein approach to Wordnet construction, in: Proceedings of the Seventh Global Wordnet Conference, 2014, pp. 215–223.
- [21] E. Sköldberg, L. Holmer, E. Volodina, I. Pilán, State-of-the-art on monolingual lexicography for Sweden, Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave 7 (2019) 13–24.
- [22] U. Teleman, S. Hellberg, E. Andersson, Svenska akademiens grammatik, Svenska Akademien and Norstedts ordbok, 1999.
- [23] L. Borin, M. Forsberg, J. Roxendal, Korp-the corpus infrastructure of Språkbanken., in: LREC, volume 2012, 2012, pp. 474–478.
- [24] E. Volodina, Y. Ali Mohammed, T. Lindström Tiedemann, Swedish Word Family Resource – Construction, Applicability, Strengths and first Experiments (subm.).
- [25] E. Volodina, Y. Ali Mohammed, T. Lindström Tiedemann, Lyxig språklig födelsedagspresent from the Swedish Word Family, in: In Volodina, Dannélls, Berdicevskis, Forsberg and Virk (editors). Live and Learn: Festschrift in honor of Lars Born, GU-ISS-2022-03, Department for Swedish, Multilingualism, Language Technology, University of Gothenburg, 2022, pp. 153–160.
- [26] A. Ingves, T. Lindström Tiedemann, Prefix i svenskan – en ordinlärningsresurs för inlärare av svenska som andraspråk?, in: Svenskans beskrivning: Förhandlingar vid trettioåttonde sammankomsten, volume 1, in print.
- [27] T. Lindström Tiedemann, Egennamn, morfologi och andraspråksinläring, in: Namn och gränser: rapport från den sjuttonde nordiska namnforskarkongressen den 8–11 juni 2021, NORNA-förlaget, 2023.
- [28] J. Bybee, Usage-based grammar and second language acquisition, in: Handbook of cognitive linguistics and second language acquisition, Routledge, 2008, pp. 226–246.
- [29] A. O’Keeffe, G. Mark, The English grammar profile of learner competence: Methodology and key findings, International Journal of Corpus Linguistics 22 (2017) 457–489.
- [30] Y. Tono, Coming full circle—from CEFR to CEFR-J and back, CEFR Journal 1 (2019) 5–17.
- [31] V. Benigno, J. de Jong, Developing the global scale of English vocabulary for young learners (6 to 11), 2017.
- [32] T. Üksik, J. Kallas, K. Koppel, K. Tsepelina, R. Pool, Estonian as a second language teacher’s tools, in: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, 2021, pp. 130–134.

A. Online Resources

Online resources mentioned in the article:

- Swedish L2 Profile (login: demo)
- CoDeRoom resource
- English Profile
- CEFR-J toolkit
- L2 Estonian Teacher's Tools

Guidelines for manual annotation:

- Lexicographic annotation guidelines: multi-word expressions, adjectives and adverbs
- Official L2P morphology annotation guidelines
- Noun pattern descriptions for Swedish
- Verb pattern descriptions for Swedish

Digital History and Immaterial Infrastructure: A Bottom-Up Approach

Sune Bechmann Pedersen¹, Marie Cronqvist² and Kajsa Weber³

¹ *Stockholm University, Sweden*

² *Linköping University, Sweden*

³ *Lund University, Sweden*

Abstract

This paper argues for an expanded view of research infrastructure. Drawing on our experiences leading the research platform *DigitalHistory@Lund*, it shows how research capacity can be unlocked “bottom-up”, by providing scholars with comparatively cheap—yet often inaccessible—technological support. By engaging researchers in digitally enabled scholarly practices, the platform yielded a multiplying effect that has seen participants produce highly competitive grant applications and eventually bring home external funding currently worth eight times the platform’s original costs. The platform thus demonstrates the importance of “immaterial” infrastructure in the sense of basic organisational structures that facilitate collaboration and communication.

Keywords

Computational history, digital skills training, research infrastructure, *DigitalHistory@Lund*

1. Introduction

Digital technology is profoundly reshaping all aspects of the historian’s craft. Digital tools are affecting the collection, organisation, interpretation, and presentation of sources as well as the communication of historians with colleagues, students, and the broader public. Digital literacy is thus crucial for today’s historians [1]–[3]. New media are also reshaping archives and libraries in unpredictable ways. It is certain, though, that the current politics of digitalisation greatly influences future historical research. Critical discussions of cultural heritage and digitisation—in close collaboration with archives and libraries—are thus essential to ensure the accessibility and usability of historical collections [4], [5]. This paper details how the authors tackled these challenges as leaders of the Lund University research platform *DigitalHistory@Lund* (2021–2023). The platform, funded by the Joint Faculties of Humanities and Theology at Lund University, aimed to support research, promote skills, strengthen partnerships, and critically reflect on the implications of digitalisation. The paper demonstrates the strengths of working bottom up with a broad and inclusive definition of digital history. It argues for the importance to historical research of “immaterial infrastructure.” By this we refer to people providing researchers with technical support to perform comparatively simple tasks that nevertheless unlock significant potential for the individual research project.

2. Digital history in Sweden

Digital history is a flexible term referring to the nature and organisation of historical sources, the tools of analysis, and the means of presenting results. While new reproduction technologies have a long history of prompting historians to consider technology’s implication for the discipline [6], [7], the past decade has seen digital history grow rapidly, sparking renewed discussions about the future of the historical discipline [8]–[10] as well as historiographical enquiries into the origins of computer-assisted historical analysis [11]–[13]. As a proof of the field’s consolidation, the first issue of the *Journal of Digital History* (De Gruyter Press/CD2H Luxembourg) appeared in 2021.

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

✉ sune.bechmann@historia.su.se (S. Bechmann Pedersen); marie.cronqvist@liu.se (M. Cronqvist); kajsa.weber@hist.lu.se (K. Weber)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Proceedings of the Huminfra conference (HiC 2024)

In Sweden, however, we found that at the time we prepared our application for the research platform in 2020, the challenges and opportunities of digital history generally remained under-explored by professional historians [14], [15]. Informal talks with senior professors about digital history often brought up memories of past flagship projects involving computational analysis, e.g. “Sverige under andra världskriget” (1966–78), and dismissive attitudes reminiscent of the biting critique Lawrence Stone directed against quantitative computational history in his classic 1979 article “The revival of narrative: reflections on a new old history”.

It is just those projects that have been the most lavishly funded, the most ambitious in the assembly of vast quantities of data by armies of paid researchers, the most scientifically processed by the very latest in computer technology, the most mathematically sophisticated in presentation, which have so far turned out to be the most disappointing ... On any cost-benefit analysis the rewards of large-scale computerized history have so far only occasionally justified the input of time and money [16, pp. 12–13].

In our mapping of the history curriculum at Swedish universities we found that digital skills training was conspicuously absent, as was research building on digital methods and historiographical literature considering the dilemmas and affordances posed by digital history. While several universities had launched broad Digital Humanities initiatives, concerted efforts centred on the specific issues and challenges pertaining to digital history were scarce.¹ A few large research projects had received five-year funding under the DIGARV funding scheme in 2018 and 2020 (e.g. Terrorism in Swedish Politics; Welfare State Analytics; Swedish Caribbean Colonialism; Mapping Lived Religion)², while others had been funded by the Wallenberg foundations (e.g. The Digital Periagesis). Yet these bold projects did not aim to systematically build capacity in digital history in Sweden. Inspired by long-term initiatives by colleagues at universities in Aarhus and Luxembourg, the aim of our proposed platform was thus to grow a sustainable, bottom-up interest in the questions pertaining to digital history among historians who do not self-identify as “digital”.³

3. DigitalHistory@Lund as immaterial infrastructure

The call for inter-departmental research platform proposals by the Joint Faculties of Humanities and Theology at Lund University was aimed at two types of projects. On the one hand, the Faculties sought to support projects building on existing strong research groups that aimed for large scale funding schemes such as “RJ Programme” and “ERC Advanced Grant”. On the other hand, the call invited applications in support of research infrastructure. We applied for the second type of project, though not with a typical, “built network” in mind [17]. Instead, we argued that what was needed to unlock historically oriented research was an immaterial infrastructure. Most historians do not require expensive equipment or expansive technical support beyond what most universities already provide in terms of basic software packages, modest server space, and solid research libraries. Rather, what was needed was an “immaterial infrastructure of human relations” [18, p. 2]. In other words, the opportunity to bounce ideas with computationally skilled colleagues and basic support to start using entry-level software would (currently) satisfy the vast majority of historians. Our application therefore emphasized the need for platform staff whose

¹ For instance, Lund University’s Humanities Lab; Umeå University’s HumLab; University of Gothenburg’s Centre for Digital Humanities; Centre for Digital Humanities and Social Sciences Uppsala; and Linnaeus University’s Digital Humanities Initiative.

² DIGARV was a joint funding scheme of the Swedish Research Council, Riksbankens Jubileumsfond, and the Royal Swedish Academy of Letters, History and Antiquities. <https://www.digarv.se/>

³ <https://cas.au.dk/en/cedhar> and <https://www.c2dh.uni.lu/>

primary task it would be to support researchers in embracing digital methods to ask new research questions or answer old questions in novel ways.

In today's research landscape, access to computationally skilled human resources is usually predicated on external funding. The large projects outlined above and the digital humanities centres with which they are affiliated all have research engineers and systems developers on their payrolls. However, these members of staff rarely have time to provide more than rudimentary support to researchers with unfunded project ideas. The gambit of our project was that an investment in staff whose sole task it would be to support (currently) unfunded projects with research coordinators and a research engineer would act as a multiplier effect and eventually pay off in terms of external funding secured by the platform-supported projects. Organisationally, the platform application was sponsored by the Section for Media History, Department of Communication and Media, in close collaboration with Lund Centre for the History of Knowledge (LUCK) and Lund University Humanities Lab. The application also involved the Lund University Libraries and other GLAM-partners in Sweden.

After an extensive external review process, DigitalHistory@Lund received funding for two years in September 2020 (later extended through 2023). The Faculty's 3.3 million SEK, combined with the overhead cost covered by the participating departments, allowed for the funding of a full-time research engineer (Mathias Johansson), plus a part-time platform coordinator (Sune Bechmann Pedersen) and deputy coordinator (Kajsa Weber).

4. Building digital history from the bottom up

The other four research platforms funded under the same call were all built on existing research groups under full professors. By contrast, the digital history research platform started virtually from scratch. To engage our colleagues, we embraced a broad definition of digital history and sought to reach researchers whose projects had a historical dimension, disregarding their departmental affiliation.

The declared ambition of the platform was to integrate existing digital history projects and support the design and execution of new projects. At the same time, the platform strove to build local digital history capacity through skills training workshops, research seminars, research engineer support, and by developing new BA and MA courses in digital history.

The defined aims of the platform were:

- To support the invention, design, and execution of digital history research projects
- To promote digital history skills in research, teaching, and dissemination
- To strengthen partnerships with digital history stakeholders outside academia
- To critically examine the implications of digital history on historical work

In practice, we pursued these aims by offering tool training seminars introducing methods and software such as *Transkribus* (HTR tool), *Tropy* (tool for managing photos of archival sources), *Voyant* (GUI text mining tool) and topic modelling. The aim was to provide a collegial space for overcoming learning thresholds and getting familiarised with relevant tools. It was important for us to highlight that we required no previous skills or experience to take part in our training seminars. The idea is, as many of these analytical tools are foreign to the historian's mythological tool box, that it is only possible to win over historians by convincing them of the practical utility of an expanded digital tool box. Digital history must communicate accessibility and practical relevance. We also aided researchers with concrete tasks, collecting, cleaning, and analysing data, for instance by setting up web scraping, structuring large data with RegEx, or geocoding with QGIS. Doctoral students, often early in their programme, proved particularly receptive and interested in integrating digital tools in their research.

In addition to the workshops, we organised a regular research seminar series inviting domestic and international speakers from the field of digital history, broadly defined. In total, the platform organised 27 seminars and workshops between 2021 and 2023 attended by more than 120 different individuals, many of whom attended more than one event. We also organised a local “inspirational conference” for work in progress presentations and hosted several conferences including the 5th Digital History in Sweden conference in November 2022.⁴

5. Third stream collaborations

The platform also engaged in projects and initiatives to digitise historical sources. These efforts were mostly directed towards the project *Digitized Swedish Print* (Digitaliserat svenskt tryck)—a collaboration between the five largest research libraries in Sweden (Gothenburg, Lund, Stockholm, Umeå and Uppsala), the National Library, and the Swedish Academy launched in 2020. The digitisation of library and archival holdings is a critical infrastructure and a necessary condition for future Swedish digital history projects. The platform thus engaged in various initiatives to ensure knowledge exchange between researchers and librarians on how to build these large research infrastructures. Among these, the platform helped organise the conference “Swedish Retrospective National Bibliography (SRNB), 1483–1599” at the National Library September 27–28 2022.

6. Results and conclusions

In conclusion, we posit that the DigitalHistory@Lund platform during its initial period of operation 2021–2023 formed a *distinct, productive, and valuable complement to existing large-scale research infrastructures* at Lund University. The platform shows how digital history can work in practice, overcome learning thresholds and engage researchers in the use of digital research practices. We built up an infrastructure that was largely *immaterial and bottom-up*, demonstrating that vital research infrastructures are not limited to large systems or physical networks of equipment and material capabilities. Alluding to the distinction between “hard” and “soft” infrastructures, the concept of infrastructure could in our context also be used to highlight *organisational or institutional structures that facilitate social collaboration, learning, information dissemination, and communication*. In fact, we hold, vital infrastructures in the humanities and social sciences are as much about exploratory workshops and seminar discussions as libraries, servers, and archives. In DigitalHistory@Lund, through seminars, workshops, conferences and course development, we created a range of common arenas that have proven to bear fruit in the form of new research collaborations and ideas. Moreover, our work has contributed to the securing of substantial external research funding. At the time of writing, the platform has actively supported successful applications in excess of 24.4 million SEK.

A cornerstone of our engagement in DigitalHistory@Lund has been to form a platform guided by *generosity and openness* [10]. The platform has welcomed all researchers at the HT faculties and invited them to take advantage of its resources and competencies. However, there are scientific and methodological issues that are specifically important for the historical sciences but have not been targeted by the much broader field of digital humanities. These issues have been seized and articulated by the DigitalHistory@Lund platform.

In the autumn of 2023, DigitalHistory@Lund entered its second organisational phase. Sune Bechmann Pedersen and Marie Cronqvist have left Lund University for positions at other universities, but Kajsa Weber remains in place to ensure the sustainability of the initiative as the

⁴ For a full list of past events, see the platform’s website <https://projekt.ht.lu.se/digitalhistory>.

platform's main coordination. Exchange continues with Bechmann Pedersen's and Cronqvist's new departments in Stockholm and Linköping as well as other universities in southern Sweden through common seminar series as well as workshops and third-stream joint collaborations.⁵ There are thus excellent conditions for DigitalHistory@Lund to not only survive but thrive in the future. As initiators and coordinators of the platform we are very proud that the HT faculties' investment in this platform turned out so well. At the same time, we are convinced that the greatest footprint of this type of investment—mainly in the form of granting even larger research grants and researchers increasingly and naturally incorporating digital history methods into their scientific toolbox—will only become visible in the longer term.

Acknowledgements

The authors would like to thank the Joint Faculties of Humanities and Theology at Lund University for the generous funding of the DigitalHistory@Lund research platform. We are also grateful to Helle Strandgaard Jensen for her invaluable advice and support throughout the project.

References

- [1] L. Putnam, 'The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast', *Am. Hist. Rev.*, vol. 121, no. 2, pp. 377–402, 2016.
- [2] H. Salmi, *What is Digital History?* Cambridge: Polity Press, 2020.
- [3] I. Milligan, *The Transformation of Historical Research in the Digital Age*. Cambridge: Cambridge University Press, 2022. Accessed: Feb. 08, 2023. [Online]. Available: <https://www.cambridge.org/core/elements/transformation-of-historical-research-in-the-digital-age/30DFBEEA3B753370946B7A98045CFEF4>
- [4] J. Jarlbrink and P. Snickars, 'Cultural Heritage as Digital Noise: Nineteenth Century Newspapers in the Digital Archive', *J. Doc.*, vol. 73, no. 6, pp. 1228–1243, 2017.
- [5] H. S. Jensen, 'Digital Archival Literacy for (All) Historians', *Media Hist.*, vol. 27, no. 2, pp. 251–265, Apr. 2021, doi: 10.1080/13688804.2020.1779047.
- [6] C. G. Andræ, 'Clio inför automationen', *Hist. Tidskr.*, vol. 86, no. 1, pp. 47–79, 1966.
- [7] E. L. R. Ladurie, 'La fin des érudits', *Nouvel Observateur*, May 08, 1968.
- [8] A. Fickers, 'Towards a New Digital Historicism? Doing History in the Age of Abundance', *VIEW J. Eur. Telev. Hist. Cult.*, vol. 1, no. 1, pp. 19–26, 2012.
- [9] M. Fridlund, 'Digital History 1.5: A Middle Way between Normal and Paradigmatic Digital Historical Research', in *Digital Histories: Emergent Approaches within the New Digital History*, M. Fridlund, M. Oiva, and P. Paju, Eds., Helsinki University Press, 2020, pp. 69–87. doi: 10.33134/HUP-5.
- [10] R. Ahnert, E. Griffin, M. Ridge, and G. Tolfo, *Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project*, 1st ed. Cambridge University Press, 2023. doi: 10.1017/9781009175548.
- [11] D. Greenstein, 'Bringing bacon home: The divergent progress of computer-aided historical research in Europe and the United States', *Comput. Humanit.*, vol. 30, no. 5, pp. 351–364, Sep. 1996, doi: 10.1007/BF00054018.
- [12] G. Zaagsma, 'On Digital History', *BMGN-Low Ctries. Hist. Rev.*, vol. 128, no. 4, pp. 3–29, 2013.
- [13] J. Jarlbrink, 'Historievetenskapens mediehantering', in *Massmediaproblem: mediestudiets formering*, Mats Hyvönen, Pelle Snickars, and Per Vesterlund, Eds., Lunds universitet, 2015, pp. 225–247.
- [14] J. P. Essen and K. Nyberg, *Historia i en digital värld*. 2014. [Online]. Available: https://digihist.files.wordpress.com/2014/05/hdv_v1_0_1.pdf

⁵ For instance, through the Digital History Seminar hosted by the Department of History at Stockholm University in collaboration with Lund University and Linnaeus University. <https://www.su.se/departement-of-history/research/conferences-and-seminars/digital-history-seminar-1.640988>.

- [15] T. Karlsson, 'Let's Make the "Digital Turn" a "Narrative Turn"! On the Gap between Two History Disciplines and How to Bridge It', *Scandia*. [Online]. Available: <https://journals.lub.lu.se/scandia/announcement/view/66>
- [16] L. Stone, 'The Revival of Narrative: Reflections on a New Old History', *Past Present*, no. 85, pp. 3–24, 1979.
- [17] B. Larkin, 'The Politics and Poetics of Infrastructure', *Annu. Rev. Anthropol.*, vol. 42, no. 1, pp. 327–343, 2013, doi: 10.1146/annurev-anthro-092412-155522.
- [18] L. Kanoi, V. Koh, A. Lim, S. Yamada, and M. R. Dove, "'What is infrastructure? What does it do?': Anthropological perspectives on the workings of infrastructure(s)", *Environ. Res. Infrastruct. Sustain.*, vol. 2, no. 1, p. 012002, Feb. 2022, doi: 10.1088/2634-4505/ac4429.

Documentation of data making, processing and use facilitates future reuse of research data: the CAPTURE project

Isto Huvila¹ and Stefan Ekman,¹

¹ Department of ALM, Uppsala University, Thunbergsvägen 3H, Uppsala, Sweden

Abstract

Reuse of research data requires knowing what the data is about but also of how it was created and previously processed, interpreted and used. The major challenges in capturing enough – but not too much – such process information, termed *paradata*, are to know what to document and how to document it in adequate detail and form. This paper showcases research and findings from the ERC-funded research project CAPTURE, which develops in-depth understanding of how paradata is being created and used today and which elicits and explores methods for capturing paradata. From a research infrastructure perspective, the most challenging question for managing paradata is how to enable and support the creation of paradata that is sufficient, relevant for its future reusers, and not too labour-intensive to produce and maintain. Considering the significant extent to which paradata is coincidental and exists because of the lack of data cleaning and management, a major challenge is also how to strike a balance between too much and too little standardisation.

Keywords

paradata, research data, research process, process documentation, research data management

1. Introduction

Reuse of research data requires not only knowing what the data is about but also a comprehensive understanding of how the data has been created and previously processed, interpreted and used (e.g. [1,2]). Without sufficient documentation of data, we risk ending up in a digital dark age [3] where hard-to-(re)use “dark data” dominates [4]. In the worst case, lack of context around the creation of archaeological research data – both digital and analog – can lead to substandard datasets that are difficult to interpret. Such datasets may not support future research and the creation of new archaeological knowledge to a sufficient extent.

The research project Capturing Paradata for documenting data creation and Use for the REsearch of the future (CAPTURE) is funded by the European Research Council. It investigates the previously fairly unexplored question of exactly what information about the creation, management, and use of research data is necessary for the data to be reusable in the future. CAPTURE also examines how this information can be captured in a way that is both efficient and comprehensive enough to support data reuse. The empirical focus of the project is archaeology. As a transdisciplinary and paradigmatically diverse field that operates with a broad range of data types, from textual evidence to measurements, visual information, and physical evidence, it provides an outstanding context to delve into the complexities of data documentation.

In the project, data about data creation and manipulation processes is referred to as paradata [5]. The concept, first introduced in survey research to refer to data that describes or concerns processes [6], was introduced in the early 2000s in cultural-heritage visualisation research through the London and Seville principles [7], which stipulate fundamentals for the documentation of visual (re)presentations in archaeological and cultural heritage-related contexts. More recently, paradata has been instituted in information and data management [8] and AI and records management contexts [9].

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

✉ isto.huvila@abm.uu.se (I. Huvila); stefan.se.ekman@abm.uu.se (S. Ekman)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The aim of this paper is to highlight the significance of capturing, documenting, and preserving research-data-related paradata in research infrastructures and, on the basis of the ongoing research in CAPTURE, to outline key challenges and opportunities relating to the management of paradata.

2. New knowledge on paradata that supports research data management and open science

The purpose of CAPTURE is to create new knowledge and increased understanding of how paradata is created and used today, as well as to develop and test methods for working with paradata. Based on the results, CAPTURE contributes to creating standards and tools for paradata and advances in data-intensive research areas that use heterogeneous research data of different origins. The project does this by creating knowledge that supports the implementation of national, European, and global policies for data management and open data (cf. [10–12]).

CAPTURE also contributes to the effective sharing and reuse of research data in discipline-specific, thematic, and interdisciplinary knowledge ecosystems and repositories (cf. [13]). The project develops a critical understanding of the social contexts and use of infrastructures emphasised in recent research agendas (e.g. ARIADNE [14–16]) and empirical research (e.g. [17]). It creates new knowledge about what data creators and users find important to document regarding data-related processes, what explicit and implicit needs for documentation there are, and how these needs can be satisfied in practice.

3. Document enough, not too much

A major challenge for capturing and preserving paradata is that different data users have different needs in different situations. Literature on data reuse has identified differing needs across various disciplines and how these needs depend on what methods and theoretical perspectives underpin the scholarly enterprise [18]. At the same time, it is both practically impossible to document everything and very hard to decide what should and should not be documented. The variety of needs along with the difficulty of predicting what needs exist make it complicated to document data-related processes.

Determining how to document and preserve just enough therefore becomes a key issue. Like all data about data [19], paradata will be incomplete. As a consequence, it is necessary to focus on striking an acceptable balance between what can be captured automatically and what has to be documented manually (e.g. [20]). It is thus important to examine what information is already embedded in the data itself [21,22] and what can be left to future users to find out for themselves through various types of “archaeological” or “forensic” post-hoc methods [23] for “excavating” existing data. To date, a great deal of research has explored each of these approaches but there has been a lack of research covering the entire paradata phenomenon and how it can be used to support the reuse of research and survey data.

The CAPTURE project uses several methods to investigate the intellectual processes that underlie the creation and use of research data within and outside of archeology and to propose and develop strategies for capturing them. This palette of methods consists of document and documentation studies (e.g. [24–26]), conceptual [27] and citation analysis [28], ethnography, review and testing of previously proposed and newly developed methods for documenting paradata, as well as interviews [29] and focus group discussions with key stakeholders.

4. Much paradata is available in existing documentation

The results from interview and survey studies and the analysis of research publications and data show that a great deal of paradata is already available in the existing scholarly output. In archaeology, survey reports constitute an important source of paradata. They are expected to document both research results and the investigation process. In addition to regular job descriptions, they convey knowledge of work processes, for example in description of results and in information about participating actors. Photographs provide an important paradata source, especially those showing work in progress, environment, and physical conditions at investigation sites [24]. Even a close reading of the dataset can contribute information about underlying processes. Word choices, descriptions, and time stamps are just a few examples of elements in databases that can yield process knowledge. Much of this

information is not documented explicitly but is inherent in the messiness of primary research data. Standardised data and metadata formats lack the flexibility to document all possible forms of process information and are unsuited to preserve such implied or inherent paradata. Perhaps somewhat counter-intuitively from a research data management perspective, therefore, extensive standardisation and data cleaning therefore risk resulting in a loss of essential paradata [26].

The fact that much paradata can already be found in existing documentation means that the main challenge with process documentation is not necessarily to expand its quantity or scope. One of the problems is that the paradata is fragmented across different parts of the documentation and that it can prove complicated to get an overall grasp of what paradata is available. Key challenges involve finding the paradata, understanding what is missing, and complementing it with the necessary additional information.

5. Documenting useful information

Another problem in finding paradata is that it is not always available or that available paradata do not correspond with user needs. In particular, information about data management procedures, standards, and structuring of data is rarely documented in detail. Results also demonstrate that data creators and users often have different views on what paradata is needed [26]. When paradata is documented, the data creator would probably focus on those elements that are obvious to them, that accord with their ideas about what is central to data creation, and that are easy to document. Data users, on the other hand, expect and need paradata that help them understand the data based on their particular situation.

The apparent gap between what data creators and data users consider to be important makes it difficult to create and provide paradata that is meaningful to both parties. Data creators have to understand how users think, anticipate what paradata is likely to be helpful, and consider data usage when creating and documenting data. The users similarly need insight into how data creation has taken place and capacity to understand how the data creation process works. An additional complication is that the specific needs depend on the purpose, context and situation of data use. Reproducing research and reanalysing data again for the same purpose and in the same research field as the original study to verify or disprove results require a different set of paradata than if the purpose is to extend the original analysis temporally, spatially, or for example socially, by combining data with other (possibly new) data in the same research field. The same applies if the data is used, possibly in combination with other or new datasets from the same or other research fields, for analysis in another research field, to produce new results using new analytical methods, or to conduct a historical study of a phenomenon related to the dataset or to the research itself.

6. Conclusions

The practical key challenges in providing enough – but not too much – paradata to make research data usable relate to documenting data creation, processing and use: what to document, but also how to document it in adequate detail and form. It is equally crucial to realise what is understood as paradata. The term “paradata” is used with different meanings in different contexts [27]. Therefore, it is necessary to clarify exactly what is meant when the term is used in theory as well as in practice. From a research infrastructure perspective, the most challenging question is how to enable and support the creation of paradata that is sufficient, relevant for its future reusers, and not too labour-intensive to produce and maintain. Considering the significant extent to which paradata is coincidental and exists because of the lack of data cleaning and management, a major challenge is also how to strike a balance between too much and too little standardisation.

7. Acknowledgments

Capturing Paradata to document data creation and use for future research projects (CAPTURE) has received funding from the European Research Council (ERC) Grant number 818210.

8. References

- [1] B.L. Voss, Curation as research: A case study in orphaned and underreported archaeological collections, *Archaeol. Dialogues*. 19 (2012) 145–169. <https://doi.org/10.1017/S1380203812000219>.
- [2] I. Faniel, E. Yakel, Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation, in: L.R. Johnston (Ed.), *Curating Res. Data Vol. One Pract. Strateg. Your Data Repos.*, ACRL, Chicago, IL, 2017: pp. 103–126.
- [3] K.D. Bollacker, Avoiding a Digital Dark Age: Data longevity depends on both the storage medium and the ability to decipher the information, *Am. Sci.* 98 (2010) 106–110. <https://doi.org/10.1511/2010.83.106>.
- [4] G. Geser, F. Niccolucci, D2.4: Final innovation agenda and action plan, ARIADNE. (2016).
- [5] L. Börjesson, O. Sköld, I. Huvila, The politics of paradata in documentation standards and recommendations for digital archaeological visualisations, *Digit. Cult. Soc.* 6 (2020) 191–220. <https://doi.org/10.14361/dcs-2020-0210>.
- [6] M.P. Couper, Usability Evaluation of Computer-Assisted Survey Instruments, *Soc. Sci. Comput. Rev.* 18 (2000) 384–396. <https://doi.org/10.1177/089443930001800402>.
- [7] J.M. Carrillo Gea, A. Toval, J.L.F. Alemán, J. Nicolás, M. Flores, The London Charter and the Seville Principles as sources of requirements for e-archaeology systems development purposes, *Virtual Archaeol. Rev.* 4 (2013) 205–211. <https://doi.org/10.4995/var.2013.4275>.
- [8] I. Huvila, Improving the usefulness of research data with better paradata, *Open Inf. Sci.* 6 (2022) 28–48. <https://doi.org/10.1515/opis-2022-0129>.
- [9] J. Davet, B. Hamidzadeh, P. Franks, Archivist in the machine: paradata for AI-based automation in the archives, *Arch. Sci.* 23 (2023) 275–295. <https://doi.org/10.1007/s10502-023-09408-8>.
- [10] A. Beck, C. Neylon, A vision for open archaeology, *World Archaeol.* 44 (2012) 479–497. <https://doi.org/10.1080/00438243.2012.737581>.
- [11] DCC, An analysis of open data and open science policies in Europe, May 2017, SPARC Europe & DCC, Apeldoorn, 2017.
- [12] E. Kansa, S. Kansa, Toward a do-it-yourself cyberinfrastructure: Open data, incentives, and reducing costs and complexities of data sharing, in: E. Kansa, S. Kansa, E. Watrall (Eds.), *Archaeol. 20 New Approaches Commun. Collab.*, CA: Cotsen Institute of Archaeology, UC Los Angeles, Los Angeles, CA, 2011: pp. 57–91.
- [13] G. Bruseker, N. Carboni, A. Guillem, Cultural heritage data management: The role of formal ontology and CIDOC CRM, in: M.L. Vincent, V. Manuel. López-Menchero Bendicho, Marinos. Ioannides, T.E. Levy (Eds.), *Herit. Archaeol. Digit. Acquis. Curation Dissem. Spat. Cult. Herit. Data*, Springer, Cham, 2017: pp. 93–131.
- [14] N. Aloia, C. Binding, S. Cuy, M. Doerr, B. Fanini, A. Felicetti, J. Fihn, D. Gavrilis, G. Geser, H. Hollander, C. Meghini, F. Niccolucci, F. Nurra, C. Papatheodorou, J. Richards, P. Ronzino, R. Scopigno, M. Theodoridou, D. Tudhope, A. Vlachidis, H. Wright, Enabling european archaeological research: The ARIADNE E-infrastructure, *Internet Archaeol.* 43 (2017). <https://doi.org/10.11141/ia.43.11>.
- [15] G. Lambourne, L. Stoakes, M. Cassar, K.V. Balen, M. Rhisiart, M. Thomas, R. Miller, L. Burnell, *Strategic Research Agenda, JPI Cultural Heritage and Global Change*, Rome, 2014. <http://www.jpi-culturalheritage.eu/wp-content/uploads/SRA-2014-06.pdf>.
- [16] G. Geser, Achievements of the ARIADNE Initiative for Archaeological Data Sharing and Research, *Internet Archaeol.* (2023). <https://doi.org/10.11141/ia.64.2>.
- [17] M.S. Mayernik, D.L. Hart, K.E. Maull, N.M. Weber, Assessing and tracing the outcomes and impact of research infrastructures, *JASIST*. 68 (2017) 1341–1359.
- [18] K. Gregory, L. Koesten, Data Needs, in: *Hum.-Centered Data Discov.*, Springer International Publishing, Cham, 2022: pp. 19–32. https://doi.org/10.1007/978-3-031-18223-5_3.
- [19] M.S. Mayernik, A. Acker, Tracing the traces: The critical role of metadata within networked communications, *J. Assoc. Inf. Sci. Technol.* 69 (2018) 177–180. <https://doi.org/10.1002/asi.23927>.
- [20] M. Stamatogiannakis, P. Groth, H. Bos, Looking inside the black-box: Capturing data provenance using dynamic instrumentation, in: *Proven. Annot. Data Process. 5th Int. Proven. Annot.*

- Workshop IPAW 2014 Cologne Ger. June 9-13 2014 Revis. Sel. Pap., Springer, Cham, 2015: pp. 155–167. https://doi.org/10.1007/978-3-319-16462-5_12.
- [21] J. Huggett, Promise and paradox: Accessing open data in archaeology, in: Proc. Digit. Humanit. Congr. 2012 Humanit. Res. Inst. Sheff., 2012.
- [22] S. Gant, P. Reilly, Different expressions of the same mode: a recent dialogue between archaeological and contemporary drawing practices, *J. Vis. Art Pract.* 17 (2017) 100–120. <https://doi.org/10.1080/14702029.2017.1384974>.
- [23] M.G. Kirschenbaum, R. Ovenden, R. Gabriela, Digital forensics and born-digital content in cultural heritage collections, Council on Library and Information Resources, Washington, D.C., 2010.
- [24] I. Huvila, O. Sköld, L. Börjesson, Documenting information making in archaeological field reports, *J. Doc.* 77 (2021) 1107–1127. <https://doi.org/10.1108/JD-11-2020-0188>.
- [25] I. Huvila, L. Börjesson, O. Sköld, Archaeological information-making activities according to field reports, *Libr. Inf. Sci. Res.* 44 (2022) 101171. <https://doi.org/10.1016/j.lisr.2022.101171>.
- [26] L. Börjesson, O. Sköld, Z. Friberg, D. Löwenborg, G. Pålsson, I. Huvila, Re-purposing Excavation Database Content as Paradata: An Explorative Analysis of Paradata Identification Challenges and Opportunities, *KULA Knowl. Creat. Dissem. Preserv. Stud.* 6 (2022) 1–18. <https://doi.org/10.18357/kula.221>.
- [27] O. Sköld, L. Börjesson, I. Huvila, Interrogating paradata, *Inf. Reseach Proc. 11th Int. Conf. Concept. Libr. Inf. Sci. Oslo Metrop. Univ. May 29 - June 1 2022.* 27 (2022) paper colis2206. <https://doi.org/10.47989/ircolis2206>.
- [28] I. Huvila, L. Andersson, O. Sköld, Citing methods literature: citations to field manuals as paradata on archaeological fieldwork, *Inf. Res.* 27 (2022) paper941. <https://doi.org/10.47989/irpaper941>.
- [29] L. Börjesson, I. Huvila, O. Sköld, Information needs on research data creation, *Inf. Res.* 27 (2022) isic2208. <https://doi.org/10.47989/irisic2208>.

Queerlit – a bibliography of Swedish fiction with LGBTQI topics

Siska Humlesjö¹, Jenny Bergenmar² and Arild Matsson¹

¹ GRIDH: Gothenburg Research Infrastructure for Digital Humanities, University of Gothenburg, Renströmsgatan 6, Göteborg, 40530 Sverige

² Department of Literature, History of Ideas and Religion, University of Gothenburg, Box 200, 40530, Sweden

Abstract

This paper summarizes the project Queerlit: Metadata and Searchability for LGBTQ+ Literary Heritage 2020-2023 and discusses some challenges in the development of this resource.

The Queerlit project consist of four parts:

1. Creating a bibliography of Swedish fiction with LGBTQI themes
2. Creating a Swedish thesaurus (QLIT), adapted from the of the linked open data thesaurus Homosaurus
3. Assigning all material in the bibliography with subject headings from QLIT.
4. A web user interface for searching the material

All four parts are integrated with the Swedish union catalog, Libris, making the results of the project available for all under a CC0 license. QLIT is the first external thesaurus integrated in the linked open data framework used in the technical platform of Libris, XL.

The bibliography spans from rune stones from the 7th century to recently published fiction. When applying subject headings for the material both general aspects of the work and specific LGBTQI topics are described, making this the most comprehensive retrospective indexing project of Swedish literature to date. The underlying knowledge organization is made a prominent method of interacting with the search interface, which is empirically designed around the needs of various user groups.

Keywords


Linked open data, LGBTQI, bibliography, metadata

1. Introduction

How has LGBTQI-themed subject matter, identities, desires, and actions been portrayed in Swedish fiction literature? The three-year research project, Queerlit, has compiled existing works and provided them with subject headings to enhance their searchability for both scholars and the interested general public. This paper delineates the project, discusses the specific challenges that bibliographic metadata for literary fiction entails, and presents the achieved outcomes, as well as the functionalities available within the interface.

Queerlit originates from a perceived lack of information and discoverability of LGBTQI literature, experienced both from librarians' and scholars' perspectives [9]. Lacking adequate search tools, Swedish queer literature scholars have developed their own methods for identifying relevant literature, including searching full-texts resources for relevant terms, and using methods depending on contextual knowledge about certain intellectual environments or reading communities. Queerlit builds on these previous bibliographical efforts, and on information activism in digital environments, such as tagging and listing LGBTQ+ literature on social media to create a bibliography with descriptive metadata in order to serve scholars in the field, and a broader public. Queerlit is a cooperation between three Swedish universities and two libraries, KvinnSam, National resource library for gender studies at the University of Gothenburg, and the

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

 siska.humlesjo@lir.gu.se (S. Humlesjö); jenny.bergenmar@lir.gu.se (J. Bergenmar); arild.matsson@gu.se (A. Matsson)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

National Library.² The project is transdisciplinary and involves the participation of literary scholars, researchers in library and information science, librarians, and a research engineer [8].

2. The bibliography

The bibliography compiles fiction written in the Swedish language, encompassing the geographical boundaries of contemporary Sweden, as well as literature produced in Sweden's national minority languages. In addition to fiction, some related literary forms have been included, such as letters and autobiographies. This literature must incorporate some form of LGBTQI thematic content, which, as defined by the project, pertains to the subversion of binary gender norms, same-sex sexual practices, emotions, and/or identities. Given that the purpose of the bibliography is to provide a comprehensive overview of how these themes have been portrayed in literature throughout various historical periods, all depictions are included, irrespective of their prominence and whether they are portrayed positively or negatively.

Throughout the project, additional inclusion criteria have been introduced, such as whether the works have been read or consumed by the LGBTQI community, or if they contain subtextual references to LGBTQI-related terminology, as exemplified by Tove Jansson's works featuring the Moomins, which include subtle allusions to LGBTQI sociolect.

To compile the bibliography and ensure its comprehensiveness, a range of methods have been employed, including:

- Previous literary research.
- Workshops involving interested readers.
- Full-text searches in the Literature Bank, a database for copyright-free Swedish fiction.
- Searches for relevant subject terms in library catalogs.
- Previous bibliographies in print and online, primarily created by interested readers.

In November 2023, the Queerlit bibliography contained 1719 indexed works, spanning from runestones to recently published works. The majority of the included works are published between 2000 and 2023 (approx. 1400), and only 16 titles were published before 1800, indicating both the difficulty of identifying relevant older literature, and the increased possibility to include manifest representation of LGBTQI motifs and characters from the later part of the 20th century until today.

3. The thesaurus

The thesaurus created, Queer Literature Indexing Thesaurus (QLIT), is a Swedish translation and adaptation of the Homosaurus thesaurus. The Homosaurus is maintained by the Digital Transgender Archive, built using linked open data, and has been in existence since 1997 [1]. The purpose of the Homosaurus is to supplement other thesauri (for example Library of Congress Subject Headings, LCSH). To develop a thesaurus tailored to the needs of the Queerlit bibliography, a list of subject terms from the Homosaurus that had been used to describe fiction in the Digital Transgender Archive and the Dutch IHLIA was compiled [2]. These terms were reviewed to assess their relevance to Swedish contexts, and the associated broader and narrower terms were included for these terms. Currently, QLIT comprises 876 terms, of which 750 are translations from Homosaurus, while 126 are unique to QLIT.

The unique terms can be broadly categorized as follows:

1. Differences in the use of the Swedish and English languages, such as between the English word "sex" and the Swedish word "kön."
2. Swedish context, including terms related to the Sami people, specific Swedish legal concepts like "gender identity assessment," and significant events in

² QUEERLIT Database: Metadata Development and Searchability for LGBTQI Literary Heritage, Infrastructure for research project funded by Riksbankens jubileumsfond 2021-2023, IN20-0013.

Swedish LGBTQI history, such as the 1979 occupation of the National Board of Health and Welfare.

3. Terms describing actions rather than identity. In fiction, actions are often described without detailing the identity of the person performing the action. For example, it might involve a woman in a relationship with another woman where the action does not reveal whether the woman is lesbian or bisexual.
4. Terms describing symbolism. E.g. terms for rainbow symbolism and transformation symbolism.

Each term in the thesaurus is accompanied by a scope note that describes how the term should be used and under what circumstances. Term definitions are drawn from sources like organizations, such as the National Association of Afro-Swedes for terms related to individuals of African descent, and authorities like the WHO for terms related to young and older individuals. The thesaurus was developed iteratively, with new terms added during the indexing process, when the need for more precise descriptions was discovered. In this way, QLIT is not just a translation of the Homosaurus, it has also entailed the inclusion and definition of new terms that were not included in the Homosaurus. The entire thesaurus is published as linked data and integrated into the National Library's cataloging interface. The fact that the thesaurus is published as linked data allows for the terms to be modified as language evolves.

4. Indexing fiction in Queerlit

Fiction is typically indexed based on concrete aspects such as geographical location, the time period in which it is set, and genre [3]. While fiction in itself poses a challenge for topical descriptions, LGBTQI literature further complicates this issue. Capturing queer meanings in texts will entail a modification of common indexing norms, such as only capturing explicit representation. Further, Swedish fiction published before 2010 seldom has any subject headings applied to it. Swedish fiction published during 2005 has in average 1.5 subject headings per work, while work published in 2020 has 3.6. subject headings per work. As Queerlit indexes older literature, this means that the subject terms we assign often are the first terms applied.

In Sweden fiction for an adult audience is indexed using Svenska ämnesord (SAO), a broad subject headings list able to be applied to all types of literature, and fiction for children using the Children's subject headings list.

The general rule for indexing is that the subject matter needs to constitute a significant portion of the work, with a requirement of 20% for non-fiction literature. However, as Queerlit aims to encompass all literature with LGBTQI themes, even works where the LGBTQI theme is very peripheral are included. To address this issue without deviating from the cataloging formats in use, a solution has been implemented in which the peripheral terms are added to the holdings record, a record specific to Queerlit.

To reduce the risk of the applied terms being subjective interpretations by the indexing librarian, all indexed works are reviewed by another librarian. Furthermore, to enhance consistency, regular meetings are held to address issues and questions related to indexing.

5. The search interface

To improve accessibility of the bibliography and the thesaurus, a search interface website³ was created. Naturally, like other data in Libris, the bibliography is available in the Libris website⁴. However, a number of important weaknesses form the need for a separate search website, which was clear at an early stage and formulated as one of the project goals. These include difficulties

³ The search interface is available at <https://queerlit.dh.gu.se/>

⁴ The Libris website is available at <https://libris.kb.se/>

with searching in sub bibliographies (such as Queerlit), inadequate support for searching around subject headings and a generally outdated user experience.

To kickstart the work with the search interface, two project-internal workshops were held. In these, we took inventory of the needs and expectations among various user groups, and compiled a list of desired functionality based on that. Somewhat later in the project, two surveys were carried out, directed towards general readers [5] and librarians [6]. The results indicated that existing library catalogues often lack opportunities to explore collections around subject headings.

Data in Libris is published freely under CC0, and is open for programmatic use through an Application Programming Interface (API). This provided functionality for nearly all of the primary needs, so a frontend-backend architecture was employed, relying on Libris as a backend service. Thus, development work was reduced to building a frontend application⁵. Additionally, another smaller backend service was created for the QLIT thesaurus, as described in [2].

In technical terms, the frontend application is built in Vue 3, a JavaScript framework. From the user's interaction with the webpage, it builds queries as HTTP GET requests for the Libris API, which responds with data in the JSON format. The frontend and the thesaurus backend are served to the web by GRIDH at the University of Gothenburg.

The search interface first took the form of a mockup site containing placeholders for core functionality. Soon, it was extended with a connection to the Libris API, adding interactive searching and real data. Since then, work has been ongoing progressively, adding functionality and revising elements of interaction design. Continuous collaboration with the Libris developer team at the National Library has ensured extended feature support in the Libris API for our benefit and theirs alike.

Previous research on user interfaces to collections was helpful in guiding this work. [4] A list of 21 "subject access features" has been developed [4], such as full-text search, auto-completion and hierarchical term search. Of these, ten are fully implemented and four partially implemented. Three items are still "on the wish list" (but one of these requires substantial changes to the Libris API), and four are not applicable to this data. [6]

As another method to guide the process, several rounds of user testing have been performed. In each of these sessions, a single user is given a few tasks, such as "*find what children's literature is in the database*". Their usage of the interface is observed, and notes are taken and then used to identify potential improvements that merit further development. These notes have also been shared with the Libris developer team, as input for their planned development of a new Libris search interface.

A primary means for navigation and searching are the QLIT subject headings. This works in conjunction with the extensive indexing practice described in **Section 4**, and as such, reflects the underlying knowledge organization. The QLIT subject headings are shown in a prominent yellow color throughout the interface, as can be seen in **Figure 1**.

6. Conclusions

QLIT was the first external thesaurus to be integrated into Libris XL. This integration necessitated adjustments to functionality of Libris XL, and questions regarding the use of linked data were raised when practical issues emerged, such as the implications of an exact match between two concepts from different vocabularies. The ability to elucidate peripheral subjects without deviating from cataloging rules is another issue that has been brought to the forefront by the Queerlit project and has been raised as desirable for other databases as well. This will likely require a new relationship between works and vocabulary terms in the BIBFRAME cataloging format.

The Queerlit search interface also functions effectively as a model for the new Libris search interface that is currently under preparation. The relatively small scale of this project has been

⁵ Code for the frontend application is published open-source at <https://github.com/gu-gridh/queerlit-gui/>

favorable in this regard. Through our collaboration, developers at the National Library are gaining insights from a real use case based on the same backend system.

For a project like Queerlit, building on past information activism, and community knowledge, navigating the constraints in various knowledge organization and technical systems is a question of ethics. What space do such systems allow for the lived experience of marginalized people, and how can we move beyond previous efforts to “correct” classification and indexing systems? [7] Developing more terms, and terms closer to the vernaculars in use in communities can be seen as an act of worldmaking [1], make visible actions and identities that have previously been misrepresented or silenced. As a linked open system, these terms can be re-used on other materials, as well as modified for future needs.

References

- [1] M. Cifor, K.J. Rawson, Mediating Queer and Trans Pasts: The Homosaurus as Queer Information Activism, *Information, Communication & Society*, 26:11 (2023): 2168-2185, DOI: 10.1080/1369118X.2022.2072753
- [2] A. Matsson, O. Kriström, Building and Serving the Queerlit Thesaurus as Linked Open Data. *Digital Humanities in the Nordic and Baltic Countries Publications* 5, no. 1 (October 10, 2023): 29–39. <https://doi.org/10.5617/dhnpub.10648>.
- [3] J. Saarti, Fiction indexing and the development of fiction thesauri, *Journal of Librarianship and Information Science*, 31 (1999): 85-92.
- [4] K. Golub, P. M. Ziolkowski, G. Zlodi, Organizing Subject Access to Cultural Heritage in Swedish Online Museums. *Journal of Documentation* 78 (2021): 211–47. <https://doi.org/10.1108/JD-05-2021-0094>.
- [5] K. Golub, J. Bergenmar, S. Humelsjö, Searching for Swedish LGBTQI Fiction: Challenges and Solutions. *Journal of Documentation* 78 (2022): 464–84. <https://doi.org/10.1108/JD-06-2022-0138>.
- [6] K. Golub, J. Bergenmar, S. Humelsjö, Searching for Swedish LGBTQI Fiction: The Librarians’ Perspective. *Journal of Documentation* 79 (2023): 261–79. <https://doi.org/10.1108/JD-05-2023-0080>.
- [7] E. Drabinski, E, Queering the Catalog: Queer Theory and the Politics of Correction. *The Library Quarterly*, 83 (2013): 94-111.
- [8] M. Polk, Transdisciplinary Co-production: Designing and Testing a Transdisciplinary Research Framework for Societal Problem Solving. *Futures* 65 (2015): 110-122. <https://doi.org/10.1016/j.futures.2014.11.001>.
- [9] J. Bates, J. and J. Rowley, Social reproduction and exclusion in subject indexing: A comparison of public library OPACs and LibraryThing folksonomy. *Journal of Documentation* 67:3 (2011): 431-448. <https://doi.org/10.1108/00220411111124532>.

From *Zipf distribution* to *Universal Dependencies* – Interactive Notebooks for Swedish Text Analysis

Dimitrios Kokkinakis¹

¹ University of Gothenburg, Box 200, 405 30, Gothenburg, Sweden

Abstract

Notebook-based environments are powerful (web-based) interactive development resources for conducting exploratory (textual) data analysis (EDA). These environments allow the embedding of code (code snippets in ‘code cells’) which can be easily executed with the results immediately presented into the user’s window. This paper introduces some basic exploratory tools and techniques using *JupyterLab* notebooks, applied to Swedish using a subcorpus that address various topics related to the COVID-19 pandemic published during January-December 2021.

Keywords

[interactive] notebooks, Swedish, R, JupyterLab, text analysis, enhanced and active learning

1. Introduction

Notebook-based environments, such as *JupyterLab*² (*Jupyter*), *Google Codelab*³ (*Colab*) or the *Kaggle*⁴ *Notebooks* are powerful (web-based) interactive development resources for conducting exploratory (textual) data analysis (EDA). The purpose of EDA is to find valuable insights in the data. Notebooks facilitate in-depth EDA by allowing collaborative research while promoting transparency and reproducibility in scholarly work by easily creating and sharing computational documents such as code, and data. These environments allow the embedding of code snippets (‘code cells’) which can be easily executed with the results immediately presented into the user’s window. Formatted text or ‘markdown cells’ are used to supplement and explain the code. Moreover, code cells can be independently executed in an arbitrary order, edited between runs and iterations, share variables and functions and allow the experimentation with different methods, models, and tools. In addition, code cell outputs, which may include charts, maps, tables, and plots, are integrated within the notebook document, or saved locally as high-quality digital format (e.g., *.jpeg* or *.png* images).

Notebooks can and have been deployed in a variety of scientific contexts, ranging from educational, economic, engineering, data science and digital humanities [1-5].

2. Textual Data and Associated Resources

This paper introduces some basic exploratory tools and techniques using *JupyterLab* notebooks (v. 3.5.3), applied to Swedish. Jupyter is often used with the Python programming language or R scripting language, but other languages are also available. Here, the R language (v. 4.3.0) is used, and as the textual corpora in all the experiments we use a dataset of roughly 1,600 documents published on-line during 2021. This dataset is part of the *sv-COVID-19 corpus*, which contains published articles in Swedish assembled from the internet that address various topics related to the COVID-19 pandemic. The corpus is further divided into 8 stylistic genres depending on their original publication forum (AuTHoRiTieS; BLOG; MeDiCaL; NEWS; PuBLicMeDia; PeRiodiCaL; ReSeaRCH and SoCiaLMeDia) and can be searched and queried in SpråkBankenText’s word research platform Korp⁵.

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.



dimitrios.kokkinakis@svenska.gu.se (D. Kokkinakis)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

² <https://jupyter.org/>.

³ <https://codelabs.developers.google.com/>.

⁴ <https://www.kaggle.com/docs/notebooks>.

⁵ <https://spraakbanken.gu.se/korp/#?corpus=sv-covid-19>.

3 Components and Tools

We present here an outline of various components that are implemented in the notebooks. These components not only serve as a comprehensive guide to the implemented methodologies but also offer a dynamic showcase of the diverse analyses conducted using R-scripts in JupyterLab. This transparency ensures the reproducibility of the research, allowing others to access and verify the data sources used. All resources, i.e., textual (raw text or URL's to the original textual document subcorpus), lexical or programmatic are available in GitHub⁶. The outline in this section presents a snapshot of some of the output of various R-scripts in the JupyterLab notebooks, i.e., from basic *frequency analysis* to more advanced techniques such as *topic modeling* and the application of *universal dependencies'* Swedish models. These sophisticated analyses demonstrate the notebooks' capacity to accommodate intricate research methodologies, providing a valuable resource for scholars seeking to explore not only the breadth but also the depth of analytical possibilities within the JupyterLab environment.

3.1 A *smörgåsbord* of scripting results from Jupyter

Within this section, a number of diverse components implemented in the notebooks unfold, presenting users with a curated selection of outputs generated by various R-scripts in the environment. From basic frequency analysis to more sophisticated techniques such as topic modeling and the application of universal dependencies' Swedish models, the scripting results encapsulate a spectrum of analytical depths. This *smörgåsbord* of scripting outcomes not only showcases the flexibility of JupyterLab but also serves as an interactive guide for researchers navigating through the intricacies of data exploration.

3.1.1. Word distributions and frequencies

The first three figures below, show different ways to depict word distributions and frequencies in the examined data. The first plot of the left image 'Rank frequency', shows the Zipfian distribution of the word frequencies in the data; where few words occur very often, and many words occur very rare. The 'green' range, second plot of the left image 'log-Rank-frequency', marks the meaningful terms in the data. Here, stopwords and low frequent words (≤ 10) are removed. The middle image shows the frequencies of 8 selected words over a monthly period; while the most frequent keywords in two of the genres are shown at the far right image.

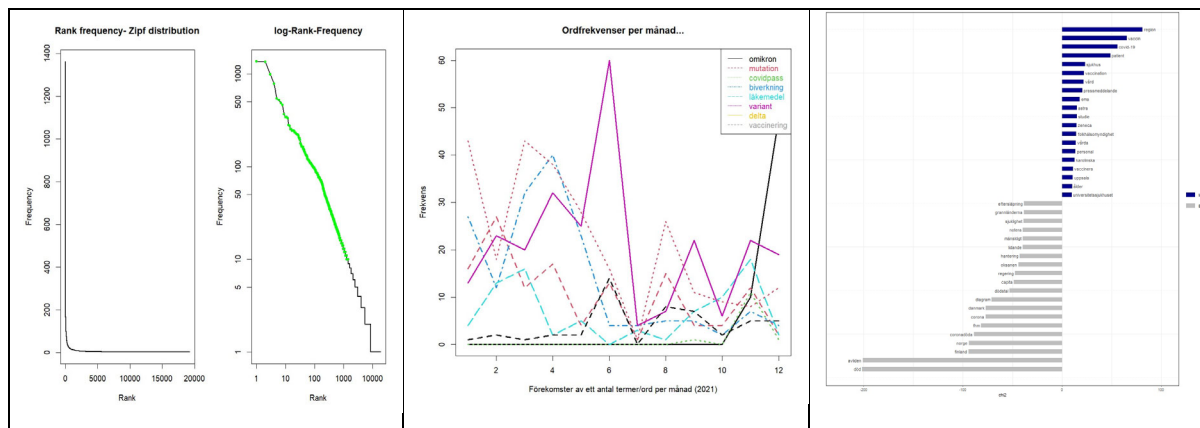


Figure 1: Word distributions and frequency analysis.

3.1.2. Significant word associations and heatmaps

The images below show different views for depicting significant words associated with a single term, here we chose the word 'omikron'. The left image shows terms strongly collocating with 'omikron'. While the middle image shows a network graph with words associated with 'omikron' in the dataset.

⁶ The URL links of the data can be found here: <https://github.com/Research-at-SBXtext/sv-JupyterLab-examples/blob/main/textual-resources/url-links-swedish-dataset.txt>.

The image to the right shows a heatmap which depicts values for a variable (here pronouns) across the genres in the dataset. Each cell is colored in a way that darker colorings imply a more frequent occurrences of a specific pronoun per genre; e.g., the personal pronoun *jag* ('I') has much higher frequency (marked in dark red) in texts that belong to the genre marked as "SCLMD", i.e., texts from social media.

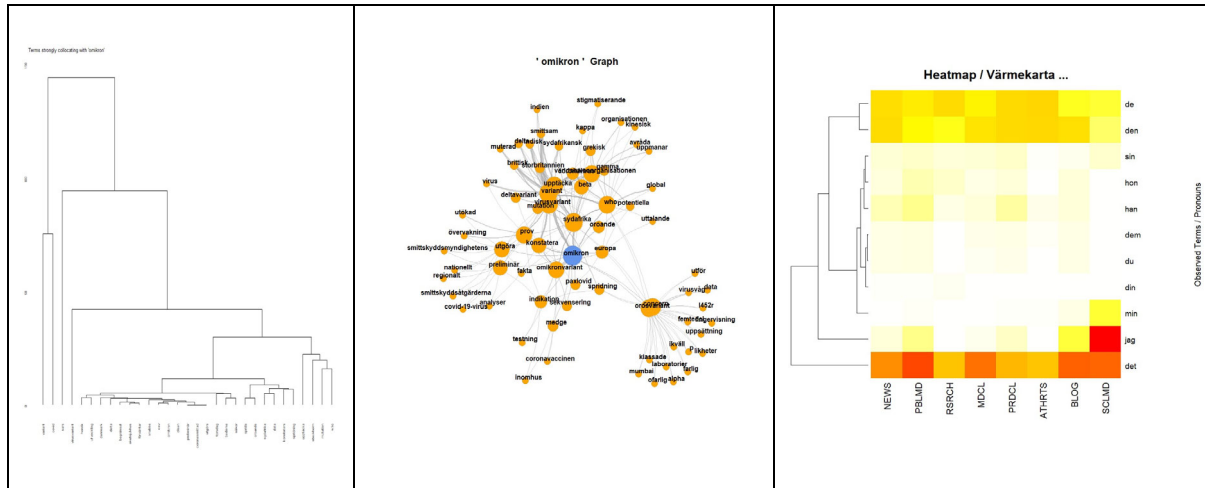


Figure 2: Significant words and words over genre distribution.

3.1.3. Sentiment analysis and universal dependencies

The left image below shows the aggregated results for sentiment analysis per corpus genre. Here we use a lexical-based approach to sentiment analysis by incorporating a large list of words with a pre-assigned sentiment value. The first bar of the plot shows that texts of category "BLOG" are much more negative than any other text genre. The image to the right shows the distribution of the Universal part-of-speech tags in the dataset (here 'NOUNS' are the most frequent type of part-of-speech). The counts originate from the application of the universal dependencies model 'swedish-talbanken-ud-2.5-191206.udpipe'.

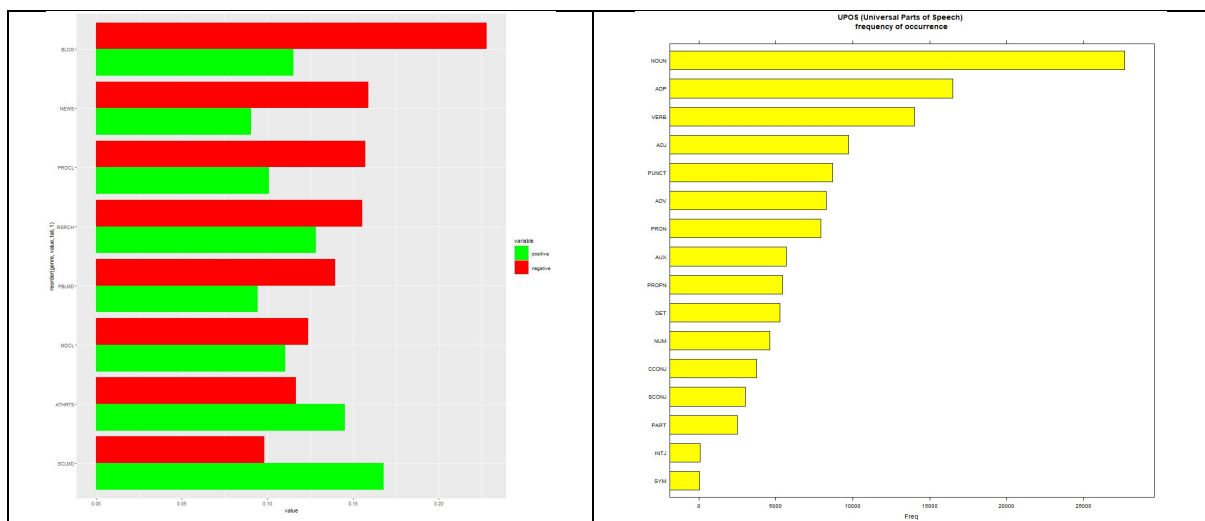


Figure 3: Sentiments and part-of-speech distribution.

3.1.4. (Flavours of) Topic modelling

Topic modeling can be used to automatically cluster and organize large document collections based on their content. There exist various 'flavours' of topic modelling techniques. The image to the left, uses a vanilla *Latent Dirichlet Allocation* (LDA), and the 40 most frequent words in one of the generated topics are shown as a word cloud. The image to the right shows the topic distribution per month (with the number of topics set to 9).

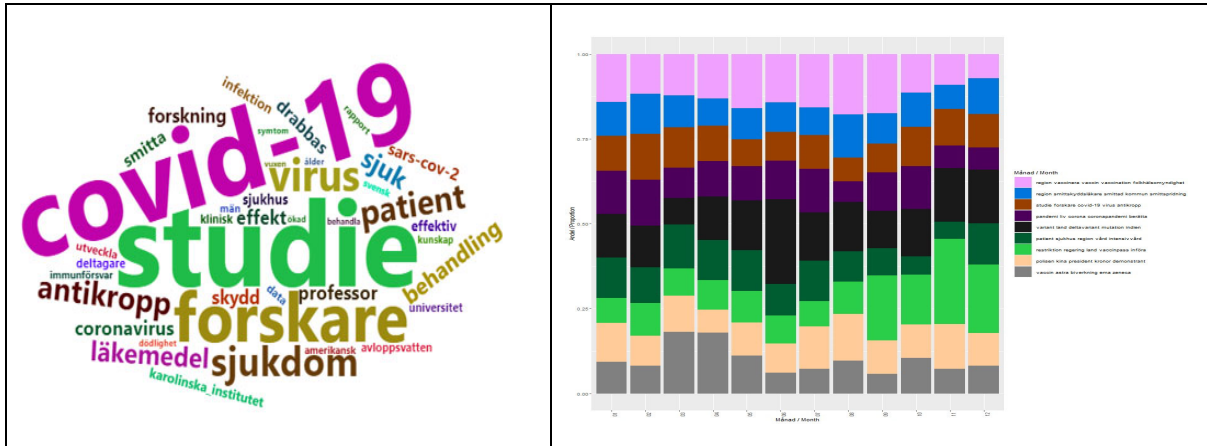


Figure 4: Topic modelling using vanilla LDA.

The two images below show some visualizations of *Structural Topic Modeling*, an approach which allows to incorporate document-metadata into the model; for instance, you can calculate the extent to which topics are more or less prevalent over *time* by incorporating the publication date of each document in the dataset. The left image shows the model diagnostics, *exclusivity*, *heldout likelihood* and *semantic coherence* (as before, 9 topics have been chosen); the right image, shows the words with the highest probabilities for each topic in the data.

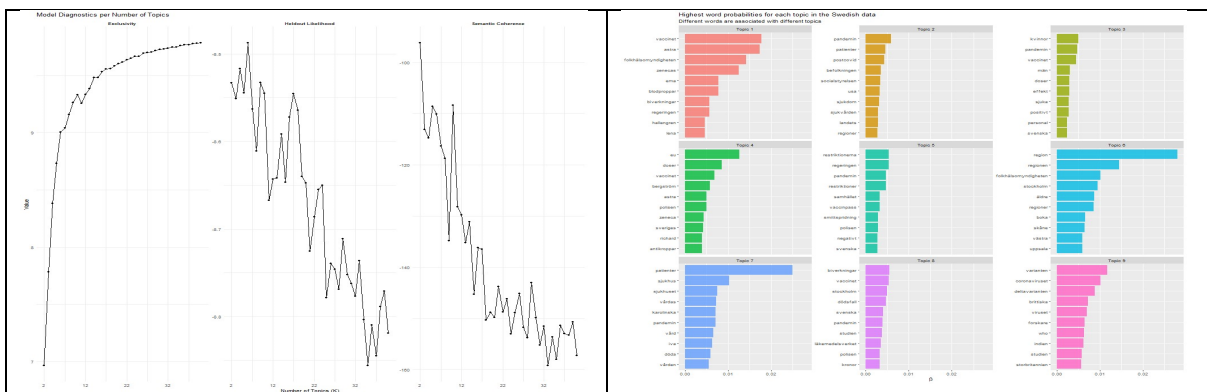


Figure 5: Structural topic modelling (diagnostics and word distributions per topic).

4 Conclusions

Interactive notebooks can be a powerful tool for researchers in e.g., digital humanities, both as a pedagogical, analytical, and scholarly tool, offering a flexible and efficient environment with rich textual documentation alongside the code, ease of collaboration and code interactivity and data visualization capabilities to help convey findings and insights. With the ability to execute code independently, researchers can engage in iterative experimentation, adjusting parameters and visualizing results instantly. This interactivity proves especially beneficial in digital humanities, where the exploration of datasets and analytical methods demands a responsive and iterative approach. Beyond the research phase, the notebooks serve as a versatile tool for report preparation, allowing scholars to generate polished documents in various formats such as PDF or HTML directly from their notebooks.

In essence, interactive notebooks emerge as a flexible, collaborative, and efficient solution, providing researchers in digital humanities with a robust platform for analysis, documentation, and communication of their scholarly endeavors. Moreover, the interactive nature of notebooks fosters a dynamic learning environment in educational settings, enabling instructors and students to actively engage with course material. This real-time interaction facilitates a deeper understanding of complex concepts and encourages hands-on exploration, making it an invaluable resource for not only teaching digital humanities but also fostering creativity, and preparing students for the rapidly evolving landscape of technology and information in the digital age in all humanistic disciplines [6].

References

- [1] David L. Alderson. 2021. Interactive Computing for Accelerated Learning in Computation and Data Science. *INFORMS Transactions on Education* Vol. 22:2. <https://doi.org/10.1287/ited.2021.0261>
- [2] Lorena A. Barba, Lecia J. Barker, Douglas S. Blank, et al. 2019. Teaching and Learning with Jupyter. URL: <https://jupyter4edu.github.io/jupyter-edu-book/index.html> and <https://jupyter4edu.github.io/jupyter-edu-book/>
- [3] Brian E. Granger, and Fernando Pérez. 2021. Jupyter: Thinking and Storytelling With Code and Data. *Computing in Science Eng.* 23(2), 7–14. <https://doi.org/10.1109/MCSE.2021.3059263>
- [4] Cécile Hardebolle. 2023. Online interactive textbooks with Jupyter Notebooks. URL: <https://www.epfl.ch/education/educational-initiatives/jupyter-notebooks-for-education/teaching-and-learning-with-jupyter-notebooks/online-interactive-textbooks-with-jupyter-notebooks/>
- [5] Quinn Dombrowski, Tassie Gniady, and David Kloster. 2019. Introduction to Jupyter Notebooks. *Programming Historian* 8. <https://doi.org/10.46430/phen0087>
- [6] Jon Chun and Katherine Elkins. 2023. The Crisis of Artificial Intelligence: A New Digital Humanities Curriculum for Human-Centred AI. *Journal of Humanities and Arts Computing*, Volume 17:2, Page 147-167. <https://doi.org/10.3366/ijhac.2023.0310>

R Packages

- quanteda:** *Quantitative Analysis of Textual Data*, <https://quanteda.io/> (v. 3.3.1)
- quanteda.textstats:** *Textual Statistics for the Quantitative Analysis of Textual Data*, <https://cran.r-project.org/web/packages/quanteda.textstats/index.html> (v. 0.96.4)
- quanteda.textplots:** *Plots for the Quantitative Analysis of Textual Data*, <https://cran.r-project.org/web/packages/quanteda.textplots/index.html> (v. 0.94.3)
- udpipe:** *Universal Dependencies pipeline*, <https://lindat.mff.cuni.cz/services/udpipe/> (v. 0.8.11)
- topicmodels:** *Topic Models*, <https://cran.r-project.org/web/packages/topicmodels/index.html> (v. 0.2.14)
- stm:** *Estimation of the Structural Topic Model*, <https://cran.r-project.org/web/packages/stm/index.html> (v. 1.3.6.1)
- Matrix:** *Sparse & Dense Matrix Classes*, <https://cran.r-project.org/web/packages/Matrix/index.html> (v. 1.5.4)
- FactoMineR:** *Multivariate Exploratory Data Analysis and Data Mining*, <https://cran.r-project.org/web/packages/FactoMineR/index.html> (v. 2.9)
- factoextra:** *Extract and Visualize the Results of Multivariate Data Analyses*, <https://cran.r-project.org/web/packages/factoextra/index.html>, (v. 1.0.7)
- dplyr:** *A Grammar of Data Manipulation*, <https://cran.r-project.org/web/packages/dplyr/index.html> (v. 1.1.3)
- ggplot2:** *Data Visualisations Using the Grammar of Graphics*, <https://cran.r-project.org/web/packages/ggplot2/index.html> (v. 3.4.4)
- ggdendro:** *Create Dendrogr. & Trees*, <https://cran.r-project.org/web/packages/ggdendro/index.html> (v. 0.1.23)
- ggraph:** *Graphics for Graphs & Networks*, <https://cran.r-project.org/web/packages/ggraph/index.html> (v. 2.1.0)
- igraph:** *Network Analysis*, <https://cran.r-project.org/web/packages/igraph/index.html> (v. 1.5.1)
- wordcloud2:** *Create Word Clouds*, <https://cran.r-project.org/web/packages/wordcloud2/index.html> (v. 0.2.1)
- reshape2:** *Reshape Data*, <https://cran.r-project.org/web/packages/reshape2/index.html> (v. 1.4.4)
- pals:** *Color palettes and colormaps*, <https://cran.r-project.org/web/packages/pals/index.html> (v. 1.7)

Tradita innovare, innovata tradere^{*}

The Gothenburg approach to computational lexicography

Lars Borin^{1,*,\dagger}, Louise Holmer^{1,*,\dagger}

¹*Department of Swedish, Multilingualism, Language Technology, University of Gothenburg, Box 200, SE-405 30, Gothenburg, Sweden*

Abstract

Swedish computational lexicography has a long history at the University of Gothenburg, both in its primary role as a central aspect of the scientific study of vocabulary and also as an infrastructural component for conducting research based on language data. Starting in the 1960s, the Språkdata research group pioneered corpus-supported lexicography for Swedish, forming the basis for successive editions of the two main descriptive dictionaries of contemporary Swedish, SAOL and SO. Language technological lexical resources for Swedish have been developed by the research unit/research infrastructure Språkbanken Text since the turn of the millennium, most recently in the framework of the *Swedish FrameNet++* initiative. After two decades of separation, these two largely mutually independently developed strands of computational lexicography have now joined forces under the umbrella of *Språkbanken's lexical research infrastructure* to advance the field technically, methodologically, and scientifically.

Keywords

Saldo, SAOL, SO, Språkbanken Text, lexicon, lexical resource, lexical infrastructure

1. Introduction

The combination of computers and lexicography has a long and distinguished history at the University of Gothenburg. Almost 60 years ago, in 1965, Sture Allén initiated the collection of digital texts for what was to become the first Swedish text corpus – the one-million word *Press 65*¹ – in order to be able to address research questions and aims such as “In a broad sense, what are the lexical units of Swedish as represented by a large corpus? How common are they, and how are they distributed over different text types? The results were primarily to be published in a frequency dictionary.” [2, p. 61, our translation].

Språkdata, the research unit founded by Allén, pursued corpus-supported lexicography for many years, concurrently with activities aimed at promoting and developing computational lin-

Huminfra Conference 2024, Gothenburg, 10–11 January 2024.


[†]For the main title of the present paper we have borrowed the motto of our university *Tradita innovare, innovata tradere* ‘Renew [our] heritage, [and] pass [it] on renewed’, a description that we feel fits our approach to construction of a research infrastructure for computational lexicography like a glove.

*Corresponding author.

[†]These authors contributed equally.

✉ lars.borin@svenska.gu.se (L. Borin); louise.holmer@svenska.gu.se (L. Holmer)

id 0000-0001-5434-9329 (L. Borin); 0009-0009-6763-6672 (L. Holmer)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Thus, *Press-65* appeared only a few years after what is usually considered the first modern text corpus, the one-million word *Brown Corpus* of American English [1].

guistics as a discipline in Gothenburg and Sweden. For various reasons, these two strands started to diverge in the early years of this millennium, but since 2021, they are again organizationally unified, with an expressed synergistic aim. The two strands have developed separately over the years, being somewhat non-communicating vessels considering researchers as well as databases and research output (graphically illustrated in Figure 1, where the two parallel strands cover roughly the twenty-year period 2002–2021).

An important historical milestone was the establishment in 1975 of *Språkbanken* (‘the Swedish Language Bank’) as a dedicated research infrastructure operated by Språkdata in support of Swedish linguistic research in general and the local lexicographical activities in particular. Språkbanken has grown considerably over the years from its humble beginnings almost 50 years ago. The main focus of its present incarnation – *Språkbanken Text* – is on language technological research rather than corpus linguistics as in the beginning of its existence. The lexicographical element has been very much present throughout its history, as described in more detail below.

The aim of this paper is to describe the background and current state of our computational lexicography infrastructure, and to point to some possible future directions for its development.

2. The *tradita*

The lexicographical projects in Gothenburg, aiming at publishing dictionaries, resulted in two print dictionaries in the 1980s, namely the first edition of the Swedish monolingual *Svensk ordbok* (‘Swedish Dictionary’ 1986 [3]; in short SOB) [4] and the 11th ed. of the *Swedish Academy Glossary* (1986 [5]; henceforth SAOL). These two datasets (SAOL and SOB, as well as their revised, later editions, respectively: [6, 7, 8]; [9], referred to as SO) have so far been treated as separate entities during the years of development. They have been revised and refined, one dictionary at a time, often by more or less the same lexicographers. SAOL and SO are financed by the Swedish Academy, and their editorial staff is employed by the University of Gothenburg [10].

2.1. SAOL: main features and functions

In this article, we mainly focus on SAOL. The first edition was published in 1874, and SAOL is therefore celebrating its 150th anniversary in 2024. SAOL is a contemporary, monolingual dictionary, aiming at providing information on orthography, inflectional patterns, and, to some extent, word formation and pronunciation [11, 12, 13].

SAOL has a unique position in Sweden – it is considered an unofficial norm, mostly regarding orthography and inflection.² Even so, many users tend to turn to SAOL for semantic information as well, although semantics is not one of the Glossary’s main features [12]. Further, a common opinion among users is that SAOL contains (only) accepted Swedish words [11]. This is of course a misconception, since a dictionary by necessity at best contains a rich selection of words rather than “the definite vocabulary”.

A ground principle for the editors of SAOL is to include new words in every new edition, preferably conventionalized ones with sufficient frequency in text, mainly contemporary newspaper text.

²There is no official normative dictionary of Swedish at the present time.

The word formation rules of Swedish allow an almost infinite number of solid compounds and derivatives, and one major editorial task in the revisional process is to decide which new words to include and which to exclude [14]. Although SAOL comprises about 126,000 headwords, it is inevitable that it will have lacunas with regards to neologisms and regularly formed derivatives [15]. It will also most certainly contain obsolete words, that ought to be excluded for various reasons [10].

The preliminary manuscript of SAOL 12 was transferred to the University of Gothenburg in 1984 [14, 4], previously being revised in Lund by the editorial staff of the historical *Swedish Academy Dictionary* (SAOB [16]). Since 2017, SAOL and SO, together with SAOB, can be accessed in the same dictionary web portal, Svenska.se. Users tend to consult SAOL mainly via the app version or the web version on Svenska.se nowadays. In the web version, search results show three (often) different lexicographical analyses for the same word: one with focus on orthography and inflection, and a more normative perspective (SAOL), one with focus on semantics, constructions and etymology and a more descriptive perspective (SO), and one providing exhaustive semantic information on a word and its other linguistic properties and their development over time (SAOB). This side-by-side presentation has made it evident to the lexicographers that coordination and harmonization between SAOL and SO are desirable. Classification in terms of parts of speech, lemma variants, lemma order between homographs etc. varies between the two dictionaries, and users tend to (with good reason) question the differences [17, 13].

Despite the advantages of Svenska.se, SAOL is one of the few contemporary dictionaries that is planned to be published in print, with the next (15th) edition slated to appear in late 2025. This should be regarded as quite exceptional since the major publishing houses in Sweden have by and large discontinued their lexicographic activities for commercial reasons over the last two decades [15].

2.2. SMDB: a Swedish morphological database

Starting with the 12th edition of SAOL (1998 [6]), a morphological database was created, SMDB, containing the 120,000 headwords and all their inflectional forms. In the print dictionaries, inflectional suffixes are presented in abbreviated form, but the SMDB allowed the inflectional paradigms to be presented in full with morphosyntactic descriptors [18]. The full inflectional paradigms of SMDB have, among other things, formed the basis of inflectional information in the e-versions of SAOL, from CD-ROM (2007), over smart phone apps (2011 and onwards) to the web version (2017).³ SMDB has also been used as a lexicographical tool for corpus investigations; by comparing SAOL's headwords and their inflectional forms to modern texts, it is possible to tease out which forms (and lemmas) are in use in texts, and which have become obsolete and are no longer in use, hence being potential candidates for exclusion from the dictionary [14]. There was also an outspoken aim for the SMDB to be continuously updated and connected to the corpora of Språkbanken [18].

³There are earlier e-versions as well, such as SAOL 11 on floppy disks etc.

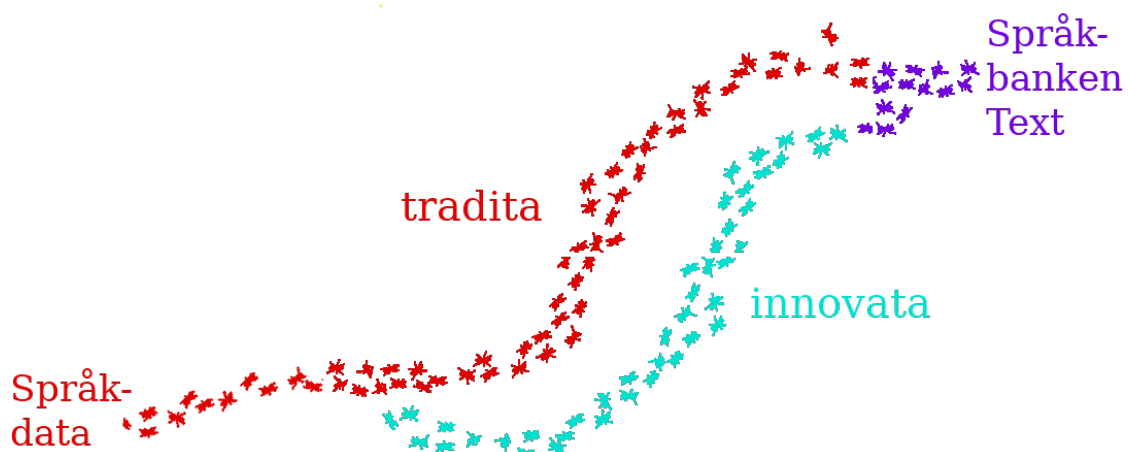


Figure 1: Computational lexicography at Gothenburg from Språkdata to Språkbanken Text. The *tradita* represents the continued development of corpus-supported lexicography as initiated by Språkdata, and the *innovata* the language technological approach to lexical resource building

3. The *innovata*

In the early years of this millennium, Språkbanken’s focus started to shift from traditional corpus linguistics in the direction of mainstream language technology. One central aim in this connection was to develop computational tools for automatic linguistic annotation of the considerable amounts of text collected in Språkbanken’s corpora,⁴ in order to make them available for research in language technology as training and testing data. Following general practice in the field, the software and language resources used should be open and freely available, ideally for all purposes, not least in order to ensure reproducibility of research. At the time, there was no freely available full-sized Swedish digital lexical resource that could be used as the basis of morphological analysis and lemmatization of the corpora, and the content of the Swedish Academy dictionary databases developed in-house could not be made openly available because of commercial commitments.

3.1. Enter Saldo

Instead, Språkbanken in practice came to initiate a parallel computational lexicographic project, which took its point of departure in SAL (*Svenskt associationslexikon* ‘Swedish Associative Thesaurus’), an onomasiological (semantic) Swedish dictionary developed by Lennart Lönngrén at Uppsala University between 1987 and 1992 [19, 20]. There was still a clear connection to our local lexicographical projects, since Lennart Lönngrén had purchased the lemma list of the recently published first edition of SO (SOB 1986 [3]) from the Språkdata group and used this as the backbone of SAL.

⁴With one exception (the part-of-speech tagged Parole corpus; <https://spraakbanken.gu.se/en/resources/parole>), the quite extensive text corpora of Språkbanken had no annotations at the time, thus allowing only for various forms of text-word and string searches.

The original SAL took the semantic-lexicon approach to an extreme, providing no information at all about linguistic form, except for using lemmas (with numerical indices in cases of colexification) as convenient labels of word senses. Thus, there was no information about part of speech, let alone inflection. In order to use it as the basis for linguistic analysis of arbitrary Swedish text, this formal linguistic information had to be added. For the inflectional information, we used a morphological processing application developed as a PhD project at nearby Chalmers University of Technology, Markus Forsberg’s *functional morphology* (FM; [21, 22]). In practice, the FM processor is used to generate a full-form lexicon from Saldo after each update of its contents. This lexicon is made available as a separate lexical resource, *Saldom* (short for ‘Saldo’s morphology’). At corpus import time, text words are matched against the full-form list and their morphosyntactic description(s) retrieved from Saldom, rather than morphologically decomposed on the fly.⁵ The only rule-based processing taking place at corpus import time is compound analysis, since the set of Swedish compounds is completely open-ended as mentioned above in Section 2.1.

The final product comprised a full-sized – containing slightly over 72,000 entries in its first release – semantic dictionary of Swedish with complete morphological specifications (inflectional paradigms plus compounding forms) provided for all entries [23], released under an open (CC-BY) license allowing all kinds of use, including for commercial purposes [24, 25]. The present development version of Saldo contains 147,650 entries, and the latest official release (Saldo 2.3, from 2015) holds slightly over 131,000 entries, i.e., Saldo is approximately comparable in size to SAOL (126,000 headwords).

Saldo differed from the print-dictionary project datasets in at least two important ways, viz. by its organization as an onomasiological lexicon and by its data model, which was explicitly designed as a formal language inspired by knowledge representation languages such as those used in the semantic web, to cater to the needs of both humans (mainly by having meaningful identifiers instead of e.g. numbers) and machines (by having an explicit syntax and compositional semantics). In the case of Saldo, the “database” is a deliberately designed intrinsic part of the lexical resource, whereas the databases used in the print-dictionary projects have always in practice been extrinsic to the lexical data: purely technical storage solutions, as it were.

There are also more subtle differences having to do with the differing aims of the two lexicographical undertakings, primarily concerning the role of easily inferrable (to a human native speaker of Swedish) information. Even though the current working version of Saldo contains more entries than the most recent edition of SAOL, the two sets of entries are not commensurable, since Saldo explicitly lists many items which are implicit in SAOL, e.g. participles – formally (deverbal) adjectives in Saldo, but inflected forms of verbs in SAOL – and verbal nouns in *-nde*. In both cases these are provided by SAOL only exceptionally. In order to serve the practical purpose of high-accuracy automatic text analysis, Saldo also includes a number of non-normative spellings and inflectional forms which frequently occur in real-world texts, but which normally are not listed in conventional reference dictionaries such as SAOL and SO.

⁵Saldom is thus broadly comparable to SMDB, mentioned in Section 2.2.

3.2. Towards Swedish FrameNet++

Even if there was no free Swedish computational lexicon available before Saldo which was both large and general enough for the intended purposes, the long history of lexicographical activity in Språkdata and Språkbanken had left behind a number of smaller and more specialized computational lexicons, resulting from various projects carried out through the years, to which could be added initiatives started elsewhere, such as the (partial) *Swedish WordNet* compiled at Lund University [26], or the crowdsourced *People’s Synonym Lexicon* created and maintained at the Royal Institute of Technology in Stockholm [27]. The *Swedish FrameNet++* project was initiated around 2010, with two complementary and interlocking aims. One aim was to combine the rich, painstakingly compiled linguistic information hidden in these both formally and content-wise quite heterogeneous resources into one unified lexical macroresource, SweFN++. The other aim was to create a computational infrastructure facilitating development of the resources themselves as well as research based on their content [25, 28].

4. The present: *innovata tradere*

So, has the *tradita* been renewed? We like to think so, in many respects. Starting in 2021, the lexicographical projects formerly organized under the Centre for Lexicology and Lexicography at our department were formally made a part of Språkbanken Text, and merged with the SweFN++ activities under a new umbrella designation: *Språkbanken’s lexical research infrastructure*. In a way, this move signalled a return to the pre-2000 organization, but at a considerably higher level of technical and methodological maturity, the former originating primarily in Språkbanken Text and the latter contributed in equal and complementary parts by the two strands of lexicographical R&D that have now joined forces.

Furthermore, the underlying databases of SAOL and SO have been migrated into the Karp lexical platform, which has been under active development for over a decade as a tool for working with formally structured language data [29], notably the computational lexical databases making up SweFN++ (in particular the Swedish FrameNet [30] and the Swedish constructicon [31]). The migration has also resulted in a long sought-after union, and to some extent harmonization, of the two sibling print-dictionary database structures (SAOL and SO).

5. The future: *tradita innovare?*

We see a bright future for computational lexicography in Gothenburg. With the recent developments described in the previous section, the strengths of the two strands that were pursued separately for two decades are synergistically combined. The result is a vibrant and multifaceted research environment intertwined with and supported by a closely integrated cutting-edge computational infrastructure for working with lexical data. This will advance Swedish computational lexicography technically, methodologically, and scientifically, and serve a broad range of R&D purposes, in particular in the humanities and social sciences.

We will now be able to draw both on highly information-rich Swedish lexical databases compiled and enriched over several decades by highly trained lexicographers and on the most

recent language technologies built on deep learning and AI. Some promising directions for the short and medium term future are development of new or improved sophisticated computational tools for mining very large text corpora in order to

- find evidence for new words and word usages, as well as obsolescing word usages [32];
- investigate phraseology and multi-word expressions [33, 34];
- track the historical development of the Swedish lexicon [35, 36, 37, 38];
- contribute to the state of the art of lexical typology [39, 40]; and, of course
- make better dictionaries (for human consumption) and lexical resources (for computer processing of Swedish text).

Acknowledgments

The work reported on here has received funding from a number of sources. Språkbanken Text is the coordinating partner of *Nationella språkbanken*, a national research infrastructure funded jointly by the Swedish Research Council and the 10 partner institutions (grant no. 2017-00626, 2018–2024). The print dictionary projects have for many years been financially supported by the Swedish Academy. In its various guises over the years, Språkbanken Text has received additional funding centrally from the University of Gothenburg, from its Faculty of Humanities, and from the Department of Swedish, Multilingualism, Language Technology, which hosts the infrastructure. For the extensive financial support that the SweFN++ initiative has received over the years, see [41, p. 8]

References

- [1] W. N. Francis, H. Kučera, *Computational Analysis of Present-Day American English*, Brown University Press, Providence, 1967.
- [2] S. Allén, Språkvetenskaplig databehandling, in: *Personliga tillbakablickar över ämnesområden vid Göteborgs universitet: Seniorakademien Dokumentationsserie Del 2, Seniorakademien vid Göteborgs universitet*, Gothenburg, 2014, pp. 61–66. <http://hdl.handle.net/2077/51552>.
- [3] SOB, *Svensk ordbok utgiven av Svenska Akademien*, 1 ed., Esselte studium, Solna, 1986.
- [4] S.-G. Malmgren, E. Sköldberg, The lexicography of Swedish and other Scandinavian languages, *International Journal of Lexicography* 26 (2013) 117–134. doi:10.1093/ijl/ect008.
- [5] SAOL 11, *Svenska Akademiens ordlista*, 11 ed., Norstedts, Stockholm, 1986.
- [6] SAOL 12, *Svenska Akademiens ordlista*, 12 ed., Norstedts, Stockholm, 1998.
- [7] SAOL 13, *Svenska Akademiens ordlista*, 13 ed., Norstedts, Stockholm, 2006.
- [8] SAOL 14, *Svenska Akademiens ordlista*, 14 ed., Norstedts, Stockholm, 2015.
- [9] SO 1, *Svensk ordbok utgiven av Svenska Akademien*, 2 ed., Svenska Akademien, Stockholm, 2021.
- [10] E. Sköldberg, L. Holmer, E. Volodina, I. Pilán, State-of-the-art of monolingual lexicography for Sweden, *Slovenščina 2.0* 7 (2019) 13–24. doi:10.4312/sl02.0.2019.1.13–24.

- [11] M. Gellerstam, Vad är Svenska Akademiens ordlista?, in: M. Gellerstam (Ed.), SAOL och tidens flykt: Några nedslag i ordlistans historia, Norstedts, Stockholm, 2009, pp. 11–30.
- [12] S.-G. Malmgren, Svenska Akademiens ordlista genom 140 år: Mot fjortonde upplagan, *LexicoNordica* 21 (2014) 81–98.
- [13] K. Blensenius, L. Holmer, E. Sköldberg, SAOL 14 som rättesnöre – diskussion om den senaste upplagan, *LexicoNordica* 28 (2021) 39–58.
- [14] S. Berg, L. Holmer, E. Sköldberg, Time to say goodbye? On the exclusion of solid compounds from the Swedish Academy Glossary (SAOL), in: A. Dykstra, T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress*, Fryske Akademy, Leeuwarden/Ljouwert, 2010, pp. 567–576.
- [15] L. Holmer, Neutrala substantiv på -ande i text och ordbok, 47, Meijerbergs arkiv för svensk ordforskning, Gothenburg, 2022.
- [16] SAOB, Svenska Akademiens ordbok, Gleerups, Lund, 1898–. URL: <https://www.saob.se/>.
- [17] E. Bäckerud, P. Nilsson, E. Sköldberg, Så används Svenska Akademiens ordböcker på nätet. Implicit och explicit feedback från användarna, in: C. Sandström, U.-M. Forsberg, C. af Hällström-Reijonen, M. Lehtonen, K. Ruppel (Eds.), *Nordiska studier i lexikografi*, volume 15, Nordisk Forening for Leksikografi, Helsinki, 2020, pp. 91–101.
- [18] S. Berg, Y. Cederholm, Att hålla på formerna: Om framväxten av Svensk morfologisk databas, in: S. Allén, S. Berg, S.-G. Malmgren, K. Norén, B. Ralph (Eds.), *Gäller stam, suffix och ord: Festskrift till Martin Gellerstam den 15 oktober 2001*, 29, Meijerbergs arkiv för svensk ordforskning, Gothenburg, 2001, pp. 58–69.
- [19] L. Lönnngren, Lexika, baserade på semantiska relationer, in: *Nordiske Datalingvistikdage og Symposium for datamatstøttet leksikografi og terminologi 1987. Proceedings*, Handelshøjskolen i København, Institut for Datalingvistik, Copenhagen, 1988, pp. 229–236.
- [20] L. Lönnngren, A Swedish associative thesaurus, in: *Euralex '98 proceedings*, Vol. 2, Euralex, Liège, 1998, pp. 467–474.
- [21] M. Forsberg, A. Ranta, Functional morphology, in: *ICFP'04. Proceedings of the ninth ACM SIGPLAN international conference of functional programming*, ACM, Snowbird, 2004, pp. 213–223. doi:10.1145/1016848.1016879.
- [22] M. Forsberg, Three Tools for Language Processing: BNF Converter, Functional Morphology, and Extract, Ph.D. thesis, Göteborg University and Chalmers University of Technology, 2007.
- [23] L. Borin, M. Forsberg, L. Lönnngren, The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology, in: J. Nivre, M. Dahllöf, B. Megyesi (Eds.), *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, number 7 in *Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia*, Uppsala University, Department of Linguistics and Philology, Uppsala, 2008, pp. 21–32.
- [24] L. Borin, M. Forsberg, L. Lönnngren, SALDO: A touch of yin to WordNet's yang, *Language Resources and Evaluation* 47 (2013) 1191–1211. doi:10.1007/s10579-013-9233-4.
- [25] L. Borin, M. Forsberg, L. Lönnngren, N. Zechner, Swedish FrameNet++: Lexical samsara, in: D. Dannélls, L. Borin, K. Friberg Heppin (Eds.), *The Swedish FrameNet++: Harmonization, Integration, Method Development and Practical Language Technology Applications*, John Benjamins, Amsterdam, 2021, pp. 69–95. doi:10.1075/nlp.14.
- [26] Å. Viberg, K. Lindmark, A. Lindvall, I. Mellenius, The Swedish WordNet project, in:

- Proceedings of EURALEX 2002, University of Copenhagen, Copenhagen, 2002, pp. 407–412. URL: <https://euralex.org/publications/the-swedish-wordnet-project/>.
- [27] V. Kann, M. Rosell, Free construction of a free Swedish dictionary of synonyms, in: Proceedings of Nodalida 2005, University of Joensuu, Joensuu, 2006, pp. 105–110.
- [28] D. Dannélls, L. Borin, K. Friberg Heppin (Eds.), The Swedish FrameNet++: Harmonization, Integration, Method Development and Practical Language Technology Applications, John Benjamins, Amsterdam, 2021. doi:10.1075/nlp.14.
- [29] L. Borin, M. Forsberg, L.-J. Olsson, J. Uppström, The open lexical infrastructure of Språkbanken, in: Proceedings of LREC 2012, ELRA, Istanbul, 2012, pp. 3598–3602.
- [30] D. Dannélls, L. Borin, M. Forsberg, K. Friberg Heppin, M. T. Gronostaj, Swedish FrameNet, in: D. Dannélls, L. Borin, K. Friberg Heppin (Eds.), The Swedish FrameNet++: Harmonization, Integration, Method Development and Practical Language Technology Applications, John Benjamins, Amsterdam, 2021, pp. 37–65. doi:10.1075/nlp.14.
- [31] B. Lyngfelt, L. Bäckström, L. Borin, A. Ehrlemark, R. Rydstedt, Constructicography at work: Theory meets practice in the Swedish constructicon, in: B. Lyngfelt, L. Borin, K. Ohara, T. T. Torrent (Eds.), Constructicography: Constructicon Development Across Languages, John Benjamins, Amsterdam, 2018, pp. 41–106. doi:10.1075/cal.22.031yn.
- [32] M. Forsberg, J. Sikora, E. Sköldberg, Words unboxed: Discovering new words with Kubord, 2023. KBLab blog post: <https://kb-labb.github.io/posts/2023-08-29-kubord/>.
- [33] L. Borin, Multiword expressions: A tough typological nut for Swedish FrameNet++, in: D. Dannélls, L. Borin, K. Friberg Heppin (Eds.), The Swedish FrameNet++: Harmonization, Integration, Method Development and Practical Language Technology Applications, John Benjamins, Amsterdam, 2021, pp. 221–259. doi:10.1075/nlp.14.
- [34] E. Sköldberg, Phraseological theory, evidence in corpora and lexicographical practice: On collocations in a monolingual dictionary of Swedish, in: K. Blensenius (Ed.), Valency and constructions: Perspectives on combining words, 46, Meijerbergs arkiv för svensk ordforskning, Gothenburg, 2022, pp. 155–182.
- [35] J. Viklund, L. Borin, How can big data help us study rhetorical history?, in: Selected Papers from the CLARIN Annual Conference 2015, LiUEP, Linköping, 2016, pp. 79–93.
- [36] E. Sköldberg, L. Holmer, Ordböcker som språkhistoriska källor, *Svenskläraren: Tidskrift för svenskundervisning* 61 (2017) 20–21.
- [37] S. Petersson, E. Sköldberg, Semantic change in Swedish – from a lexicographic perspective, in: N. Tahmasebi, L. Borin, A. Jatowt, Y. Xu, S. Hengchen (Eds.), Computational approaches to semantic change, Language Science Press, Berlin, 2013, pp. 149–167. doi:10.5281/zenodo.5040308.
- [38] Y. Adesam, P. Andersson, L. Borin, G. Bouma, A lexical resource for computational historical linguistics, in: D. Dannélls, L. Borin, K. Friberg Heppin (Eds.), The Swedish FrameNet++: Harmonization, Integration, Method Development and Practical Language Technology Applications, John Benjamins, Amsterdam, 2021, p. 98–121. doi:10.1075/nlp.14.
- [39] L. Borin, Core vocabulary: A useful but mystical concept in some kinds of linguistics, in: D. Santos, K. Lindén, W. Ng’ang’a (Eds.), Shall we play the Festschrift game? Essays on the occasion of Lauri Carlson’s 60th birthday, Springer, Berlin, 2012, p. 53–65.

- [40] L. Borin, B. Comrie, A. Saxena, The Intercontinental Dictionary Series – a rich and principled database for language comparison, in: L. Borin, A. Saxena (Eds.), *Approaches to Measuring Linguistic Differences*, De Gruyter Mouton, Berlin, 2013, pp. 285–302.
- [41] L. Borin, D. Dannélls, K. Friberg Heppin, Introduction, in: D. Dannélls, L. Borin, K. Friberg Heppin (Eds.), *The Swedish FrameNet++: Harmonization, Integration, Method Development and Practical Language Technology Applications*, John Benjamins, Amsterdam, 2021, pp. 3–35. doi:10.1075/nlp.14.

A. Online Resources

- Svenska.se
- Språkbanken Text
- Språkbanken Text: Press 65
- Språkbanken Text: Saldo
- Språkbanken Text: Saldom
- Swedish FrameNet++
- Språkbanken Text: Karp (lexical platform)
- Språkbanken Text: Korp (corpus platform)
- Språkbanken Text: lexical datasets

Collectio: a software especially designed for creating dynamic libraries for fluid and multilingual text traditions

Britt Dahlman¹

¹ Lund University, Centre for Theology and Religious Studies, LUX, Box 192, 221 00 Lund, Sweden

Abstract

This contribution presents a new software, Collectio, which can be used for creating highly complex relational MySQL databases, or more accurately, dynamic libraries. These libraries prove particularly well-suited for texts where the material has been organized in different ways and thus represents a ‘fluid’ textual tradition, or in traditions transmitted in many languages. So far, two libraries have been created using Collectio: APDB (the *Apophthegmata Patrum Database*) and HIPPO, which contains pre-modern hippiatric material. The sources included in the libraries are mainly in the form of manuscripts, editions and modern translations. Collectio employs a unique input model, built upon .txt and .csv files stored in an archive in the folder of the library. The contents of the database tables in the master database are generated from these documents. Since not only texts are registered but also the detailed structure and parallel text segments in other sources, both texts and structures can be systematically compared and analysed within and across language boundaries. In addition to the advanced research tools for comparing texts and structures, the application contains search options, indexes of names, places and concepts, metadata on the sources, pre-written SQL commands and more. A new way of encoding text, which can be converted into TEI/XML, is also introduced.

Keywords

relational database, dynamic library, TEI/XML, fluid text, multilingual text tradition, fixed-content miscellanies

1. Background

The history of the software called ‘Collectio’ begins with the research program called ‘Early Monasticism and Classical Paideia’ (MOPAI). This research program started in 2009 and was headed by Professor Samuel Rubenson at the Centre for Theology and Religious Studies at Lund University.[1] An important part of the research consisted of editing and studying monastic wisdom literature, especially the collections of *Apophthegmata Patrum* (AP), i.e. the sayings of the desert fathers and mothers. They are preserved in multiple languages and the material is both large and complex. The sayings are compiled in numerous collections that have a complicated textual transmission. Since different text redactions and stages in many languages (in particular Greek, Latin, Syriac, Armenian, Arabic and Ethiopic) were to be studied and compared, the idea of a digital tool was soon hatched. Scholars studying the AP tradition have made concordance tables to facilitate a comparative study of the sayings in the various collections.[2] Therefore, a relational MySQL database was considered the best solution, as it seemed especially suitable for comparing ‘fluid’ texts preserved in different types of organisations and in multiple languages. For the creation of this database, IT architect Kenneth Berg was consulted. Another IT technician was Leif Trulsson, who particularly helped in the creation of the graphical user interface. To secure its longevity, it was decided that the database should be based on .txt and .csv files, since it could be assumed that applications supporting those formats would exist for a long time to come. The tool was called APDB, short for the *Apophthegmata Patrum Database*, and was written in ooRexx. One of the reasons for this choice of programming language was that ooRexx includes some very powerful functions to handle character strings. The aim of the tool was not in the first place to store an archive of digitized editions or to produce digital critical editions (although this

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

 britt.dahlman@ctr.lu.se (Britt Dahlman)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

was also possible), but to provide data and advanced research tools for comparing and visualizing texts and structures.[3] The graphical interface of APDB was released for internal use in the research team in 2012.

Since APDB proved to be very useful for comparing texts consisting of clearly defined smaller narratives manifested in different types of organisations, it was decided to use the same software technology for another database called HIPPO as part of the project ‘Knowledge, Magic and Horse Medicine in Late Antiquity’ (2021–2025). The project, funded by the Swedish Research Council, is a collaboration between Lund University and the Swedish Institute in Rome. One of the main goals of this project is to make previously relatively unknown source material on horse medicine from Antiquity and the early Middle Ages available on an open access digital platform. Thus, HIPPO will be complemented by a web application called ‘Hippiatrica – a dynamic library and research tool’. This platform will be hosted by the Swedish Institute in Rome and accessible via the Institute’s website and the URBIS Library Network.[4]

In 2022 it was decided to make this useful tool available for a wider public. Thus, the software needed a common name, so that it could be used for creating other dynamic libraries. The name agreed upon was ‘Collectio’. The application is installed on the user’s own computer together with some other applications needed as prerequisites, such as ooRexx and the relational database management systems MySQL or MariaDB. The plan is to publish the source code on GitHub under the MIT license in 2024. Another expectation is to make a version of APDB (the Public DB) with a selection of the most important output functions accessible online at <https://apdb.collectio.se> in the beginning of 2024. For input of data, you will need to have the program installed on your computer, although there are plans for making it possible in the future for a small group of registered scholars to access the input files of APDB using the web interface.[5]

2. Two dynamic libraries created from Collectio: APDB and HIPPO

As for now, two dynamic libraries have been created using the Collectio software: APDB and HIPPO. Regarding the first one, APDB, work has been ongoing since 2012 to constantly improve its functions and the material. The aim was and still is to include as many sources as possible that contain dossiers of AP from editions, manuscripts and modern translations. Descriptions of the many collections of AP and their different organisations have been published in several books and articles and will not be repeated here.[6] However, soon the ambition was widened as it turns out that apophthegmata often appear in larger monastic compilations mixed with other monastic texts, and that hagiographic works often either quote from the apophthegmata or are used in collections of apophthegmata. Therefore, also works like the *Historia Lausiaca* by Palladius, the *Historia Monachorum in Aegypto*, and the *Pratum Spirituale* by John Moschos are registered, and new sources are constantly added.

HIPPO was created in 2021 and is under construction. The material to be included is much more limited compared to the large amount of monastic material that is and could still be included in APDB. Ancient hippiatric literature consists of predominantly Greek and Latin texts on the care and medical treatment of horses written from about the 4th century BC onwards. One of the most prominent veterinarians, who wrote down his advice in Latin, was Pelagonius. Greek veterinary works are generally not preserved in their entirety, but only as excerpts collected in the extensive Greek anthology *Corpus Hippiatricorum Graecorum* (or *Hippiatrica*), probably in the 5th and 6th centuries AD. This collection is transmitted in several different redactions where the material is organized in different ways, mainly alphabetically and thematically. Several works as well as redactions of the *Hippiatrica* were so popular that they were translated into many languages, including Arabic, Syriac, Armenian, and vernacular languages such as Medieval Italian and Spanish.

As in the case with the collections of AP, the texts of the *Hippiatrica* thus represent a ‘fluid’ transmission of structure and text with constant adaptation to new contexts, new audiences or settings, such as geographical, cultural, social or didactic ones. In a way, each manuscript is a unique ‘edition’, where the order and appearance of the texts may differ more or less. The scribes, or rather the compilers, may delete or include new material, reorder it, sometimes according to a recognizable principle, modify the text and sometimes reattribute the pieces. This is characteristic of ‘encyclopedical’ collections and other compilations of e.g. hagiographic, liturgical, monastic or ‘scientific’ material.

Many of them are so-called ‘fixed-content miscellanies’, i.e. manuscripts or compilations containing texts that belong to a more or less fixed genre, but where the occurrence and order of the texts vary.[7] Texts belonging to such genres are problematic to edit according to a stemmatic-genealogic method and present in traditional text-critical editions, because of the risk of contamination and of making new compilations of the material not respecting the variable tradition to which such collections of texts belong.

3. Why use Collectio and which functions should it have?

When relational databases are used in text editing projects they are often used as tools for collecting metadata, i.e. data that provides information about the sources, such as bibliographical data for publications or codicological data for manuscripts.[8] One example is *The Digital Victorian Periodical Poetry* project.[9] In this project a relational database was created for collecting metadata on more than 15,000 poems from 19th-century periodicals. Transcriptions of the poems were encoded in TEI/XML files. The data from the relational database was then integrated into an already existing TEI file or, if no transcription of the poem existed, a new TEI file consisting of only metadata from the relational database was created. According to Martin Holmes, the plan was to eliminate the relational database by the end of 2022.[10] However, some projects use relational databases for recording both metadata and text. One example is the *Database of Byzantine Book Epigrams* (DBBE) project.[11] It records both text transcriptions and metadata of book epigrams (‘metrical paratexts’) found in medieval Greek manuscripts dating up to the fifteenth century in a relational (PostgreSQL) database.[12]

A dynamic library created from Collectio is primarily intended to contain material in the form of texts and structural lists (with parallels) of manuscripts, editions and modern translations that are hard to map and analyse with traditional methods due to its ‘fluid’ nature. The scholarly debate about digital editions often concerns problems with how to handle the information found in the apparatuses of critical editions and how to present it.[13] The focus is often on textual variation and not on structural variation.

One of the advantages of relational databases is that not only the texts can be registered but also the detailed structure and parallel text segments in other sources (*loci paralleli*). Collectio allows texts and structures to be displayed in various kinds of output. Visualizations (statistics, diagrams etc.) demonstrating relations and relational distance can be created in other applications (e.g. for spreadsheets) using output from Collectio. Consequently, both texts and structures can be compared and studied systematically within a language as well as across language boundaries.

The prerequisites of Collectio would be the following:

- It should constitute a digital archive of editions, modern translations and manuscript transcriptions, but above all a research tool and laboratory for searches, analyses, comparisons and visualizations.
- It should have output functions for comparisons of both the occurrence and order of text units **structurally** within and across language boundaries and thus be an aid in the reconstruction of stages in the development of compilations/redactions and their relationship to each other.
- It should have output functions for comparisons of **texts** within and across language boundaries and thus be an aid in the reconstruction of stages in the development of compilations/redactions and their relationship to each other, as well as for text-critical investigations.
- It should have export functions that enable exchange with other databases, e.g. through marked-up modified diplomatic transcriptions.
- It should be open access; the source code should be published under the MIT license. For the existing libraries, all data in the Public DBs should be licensed under Creative Commons BY-SA.

Collectio provides an archive of structural tables and texts, indices of names, places, concepts and other term types, metadata on the sources, as well as scholarly annotations on the material, as is standard in text databases. The research tools enable searches for words, names, places and terms, as well as analysis and comparison of relations between separate text entities and collections (text and structure). Visualization in the form of graphs and diagrams can be created in other applications (e.g. for

spreadsheets) using output from Collectio. SQL queries can easily be performed, both through a list of pre-written SQL commands and through free text queries.

The files in Collectio are encoded with a new markup language, the Collectio code, consisting of a word (or letters) beginning with a colon (e.g., :codex). The creation of this tag design was inspired by GML (Generalized Markup Language). In addition, texts can be encoded with specific embedded tags consisting of curly brackets, an asterisk, and a number (such as {*9 f.10r} for indicating the beginning of folio 10 recto). They are easy to use and can be converted into TEI/XML and exported to text documents. However, the exported text documents do not contain the full TEI data – there are no `teiHeader` elements for example – and usually only a limited amount of tags are used for the text entities.

4. Database model

The model used for registering input in Collectio is unique and its analytic tools constitute a further innovative step compared to other kinds of digitization projects for text corpora using XML-based platforms. The basis of the database, i.e. the library, is .txt and .csv files using Unicode UTF-8, which currently are maintained using common applications for text documents and spreadsheets in .odt and .ods formats. From these documents, which must be correctly encoded, the contents of the database tables in the master database are created. The files are stored in an archive, which can reside in a local folder, if used in a project involving one single person, or on a file share or an FTP server, if used in a project involving several persons. For changes in the files or new files to be included in the master database, it must be reloaded in a process where it is deleted and then recreated. The reason for this is that changes in one source can affect other sources, which then must be reloaded. These sources may in turn have an impact on even more sources. The most rational way to deal with this is to reload the entire database. For a collaborative project, the process should be initiated regularly by a person in charge. Through an advanced control system, all input is checked, and if there are any technical errors, the loading is interrupted, and a list of errors is exported. The errors often depend on incorrect markup (such as if the same ID has been registered for more than one entity). Then, after the errors have been identified and corrected, the master database can be reloaded. Thus, new data or changes in the data are not immediately included in the database tables, unlike the common procedure in relational databases. This ensures that the database is consistent and technically correct, which is crucial for complex relational databases.

Why not use web applications for input? They are often used in administrative IT systems and in crowdsourced platforms when a tool will be used by many people. Web applications can be of great help in guiding the contributor to make technically correct input if the applications are well designed. However, the more complex a relational database is, the more difficult it is to create web applications for input in every situation which can occur and that you cannot anticipate. This should especially be considered if the input goes directly into the database tables without any control system, as is usually the case. In a relational database it is especially important that the connections are consistent, and no breaks occur in a chain of relations. If severe technical errors, such as inconsistency of relations, occur without the users' attention, it may take a long time before they are detected, and in the meantime, other material could have been inserted. Then it could be difficult to restore the database without information being lost. Even if it is theoretically possible to create web applications for such complex connections as are found in Collectio, and especially for making changes in already existing connections between segments, to make them easier to use for the contributor than what can be done in common applications for spreadsheets is a challenge that few programmers would be willing to take on.

A selection of the contents of the master database, where copyrighted material and work for other reasons are excluded, constitutes the public database, which thus can be published online. A copy of all contents of the master database constitutes the personal and the common databases. To get access to the personal and common databases registration might be required. The difference between the latter two databases is that the personal database allows the user to contribute to the database by uploading files through the interface and updating his/her own personal database.

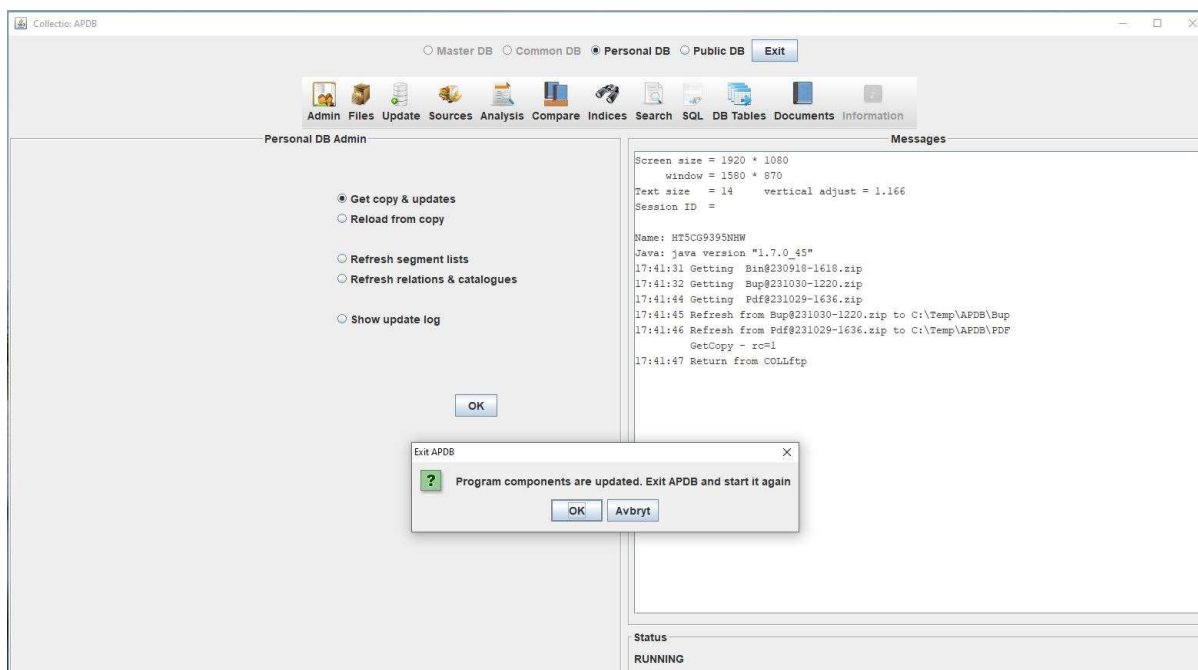


Figure 1: The internal interface of APDB. At the top, the buttons for the output functions are placed. In this case, the ‘Admin’ button and the ‘Get copy & updates’ option have been selected.

In the internal interface, the individual user of a personal database can download a new copy of the database to replace the old one when this is available (i.e. when a new master database has been created) by clicking on the ‘Admin’ button and choosing ‘Get copy and updates’. The user is also alerted if new program components exist (see Fig. 1).

In addition, the user can also update his/her own personal database. A contributor working on a file can test it by clicking on the ‘Update’ button and selecting the file in question. This can be done by both new files not yet transferred to the local folder or server and by files outchecked from it. If the file contains any technical errors due to e.g. incorrect tagging, the updating is interrupted, and a list of errors is exported. After a revision of the file, the process can be repeated and, if successful, the data is included in the personal database. In this way the contributor has a ‘personalized’ database with new data which is available before it is included in the master database. Thus, this new data can be used for various kinds of research (searches, comparisons, and visualizations) together with the other material in the analytic tools. Output from all database versions can be an HTML page that is opened in a web browser, or a file that is opened in either a spreadsheet application or a word processor. Fig. 2 gives a simplified overview of the database model.

The concept of using text files and spreadsheets for input allows the contributor to use the many functions found in common applications, e.g. search and replace letters. Private notes that the contributor does not want to share with the public user of the database can easily be inserted and left untagged, for view and revision later by the contributor. The downside is that it requires much training to acquire the knowledge needed for correctly marking up the documents, in particular the spreadsheets.

As soon as a file is saved with new information and transferred to the local folder or the server, it is marked with a timestamp and stored in an archive. All current active files as well as old files are accessible through the ‘Files’ button in the interface. This makes it easy to go back and follow the evolution of a source, to track changes, and to see who has done them and when.

5. How to use Collectio

Collectio is designed especially for comparing structures and texts consisting of small entities manifested in different types of organisations. Therefore, to facilitate comparisons, all base entities are divided into smaller entities called ‘segments’, which are given unique IDs. Base entities are ‘Unit’, ‘Title’, ‘Explicit’, and ‘Item’. They consist of one or more segments, which are labelled ‘a’, ‘b’, ‘c’, and so on. These are central terms for defining the structure of a source. Through the segments, it is

possible to connect these structural entities between sources (and within a source). A unit normally corresponds to a paragraph in a text (e.g. a saying in the AP collections or a text excerpt in the *Hippiatrica* redactions). In printed editions they are often numbered in sequence. An example of an ID for a manuscript source would be ‘Athos_Prot_86 V.2a’, meaning chapter ‘V’, unit (i.e. saying) ‘2’, segment ‘a’ in the source (i.e. the manuscript) labelled ‘Athos_Prot_86’.

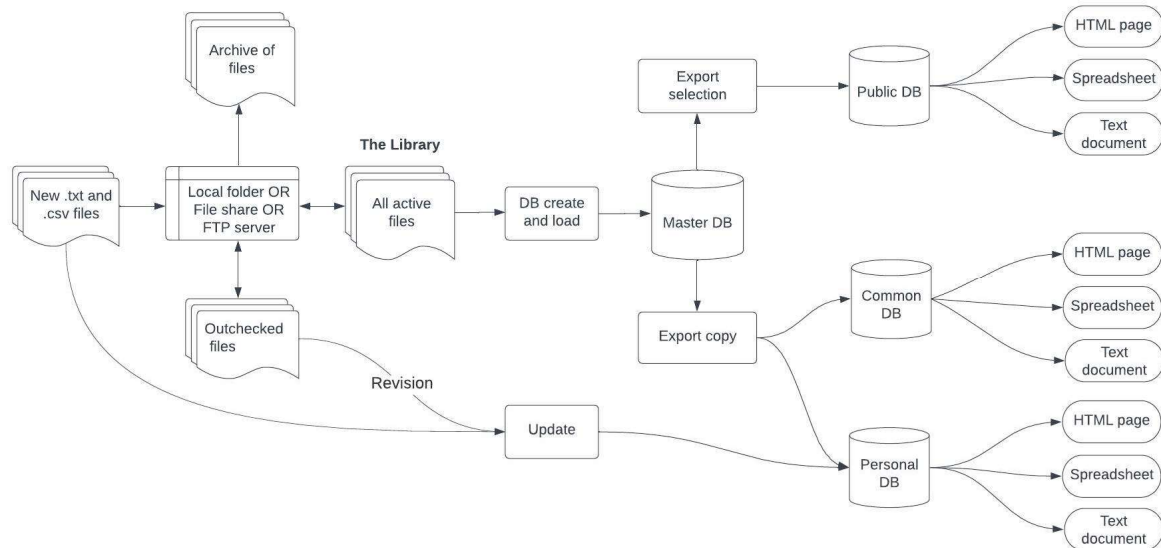


Figure 2: Overview of the database model for Collectio.

The connections between the segments are made through so-called reference series, which are lists of unique IDs based on the base entities in one or several manuscripts, editions or translations. Segments having the same registered reference within a source or between sources are thus automatically linked to each other. The connection between two or more segments can be registered in two ways: as being equivalent (marked as ‘=’), or as being similar (marked as ‘~’). To decide which text segments should be registered as (only) similar to other text segments is not always easy, but usually, the content should be approximately the same, while the wordings of the segments should have substantial differences, such as being longer or shorter. However, there is also the possibility to register only a relation (marked as ‘-’), which could be described as a more general connection. A relation cannot be registered on the segment level – it is always a relation to a base entity or a higher-level entity.[14] The connections between text segments being equivalent (=) are managed by a number of reference series. If we, for example, look at the connections for Athos_Prot_86 V.2, we find these registered connections (together with many more):

Athos_Prot_86 V.2a = Vat_lat_600 V.6 ~ SyrEn-Bedjan I.587a

Athos_Prot_86 V.2b = Vat_lat_600 V.7 = SyrEn-Bedjan I.587b

It means that the ‘a’ segment in the Athos manuscript corresponds to unit ‘V.6’ in the Latin manuscript ‘Vat_lat_600’, but that they are only similar to unit ‘I.587a’ in the Syriac edition ‘SyrEn-Bedjan’. The ‘b’ segment corresponds to unit ‘V.7’ in Vat_lat_600’ and to unit ‘I.587b’ in ‘SyrEn-Bedjan’.

Two common ways of organizing material in compilations are alphabetically and thematically. Among the collections of AP, alphabetic collections organize the material according to the names of the desert fathers and mothers, and systematic ones according to themes, usually of virtues and vices. The connections through unique segment numbers make it possible to compare different sources and to use visualization tools to demonstrate relations and relational distance between them. Fig. 3 shows how the apophthegmata in the Delta (‘D’) section in the Greek alphabetic collection (G) are distributed among the thematic chapters in the Greek systematic collection (GS). In APDB the correspondences between the reference series G and GS can be displayed through a table, as in Fig. 4. The connections can also be visualized in different ways, for example as a diagram demonstrating relational distance created in a spreadsheet application from a table of sequence numbers as output from APDB.

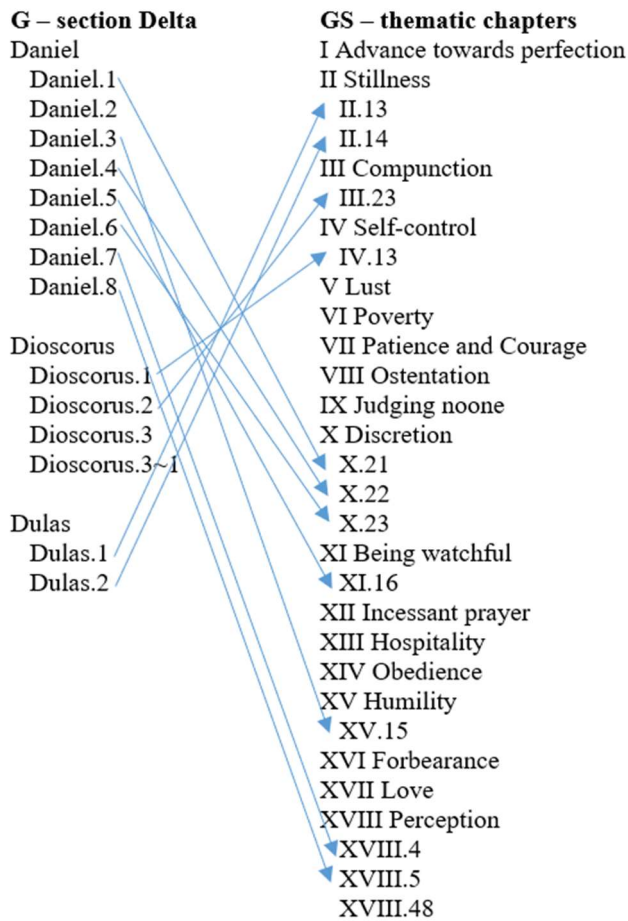


Figure 3: Connctions between parts of G (Greek alphabetic AP) and GS (Greek systematic AP).

G	GS
Daniel.1	=X.21
Daniel.2	
Daniel.3	=XV.15
Daniel.4	=X.22
Daniel.5a	=XI.16a
Daniel.5b	=XI.16b
Daniel.6a	=X.23a
Daniel.6b	=X.23b
Daniel.6c (A1)	=X.23c (A1)
Daniel.7 (A1)	=XVIII.4 (A1) ~XVIII.48 (B1)
Daniel.8	=XVIII.5
Dioscorus.1	=IV.13
Dioscorus.2	=III.23a
Dioscorus.3	
Dioscorus.3~1a	
Dioscorus.3~1b	
Dulas.1a	=II.13a
Dulas.1b	=II.13b
Dulas.2 (A1)	=II.14 (A1)

Figure 4: Table demonstrating the connections between parts of G and GS as output from APDB.

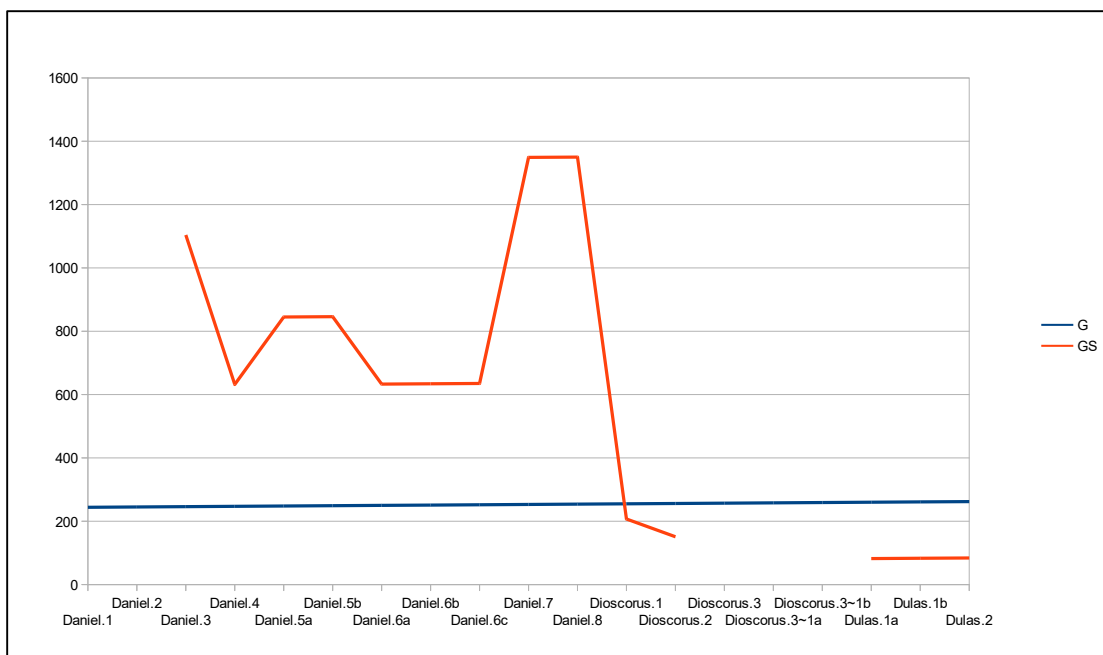


Figure 5: Comparison between G and GS visualized in a diagram created from output from APDB.

In Fig. 5, the Delta section in G has been selected as starting point and GS as a source for comparison. In the diagram, the horizontal axis displays the unit IDs in G and the vertical axis the sequence numbers in G and GS. The line of the source selected as starting point, G, is accordingly a straight line, and the other line representing GS illustrates how the segments in GS deviate from the order or are missing compared to G.

6. Concluding remarks

‘There are no shortcuts in creating a database. The code used in the design of the database sets the limits of the functions. Likewise, the data you put into a database is the data you can process. There is no magic involved in making the data better than it was when it was inserted.’ These thoughtful words come from Kenneth Berg, the creator of Collectio, a software written in ooRexx for creating highly complex relational MySQL databases or, to put it more accurately, dynamic libraries. So far, two libraries have been created using Collectio: APDB (the *Apophthegmata Patrum Database*) and HIPPO, which contains pre-modern hippiatric material.

The model used for input is distinguished from other models used in relational databases. The basis of the library is .txt and .csv files, which are stored in an archive in the folder of the library. This folder may reside on a local drive, a shared file system, or an FTP server. From these documents, the contents of the database tables in the master database are created. The master database is regularly reloaded and recreated, and through an advanced control system, all input is checked ensuring that the database is consistent and technically correct. Through a number of advanced research tools Collectio enables texts and structures to be compared and studied systematically on different levels both within a language and across languages. Files in Collectio are encoded with special markup codes that can be converted into TEI/XML and exported to text documents.

Another strength of Collectio is its independence from public (and other) funding. From 2009 to 2015 parts of the development costs were funded by Stiftelsen Riksbankens Jubileumsfond, and for the years 2021–2025, a small portion is funded by the Swedish Research Council (Vetenskapsrådet). However, for the main part, the development of the software is carried out through the voluntary work of dedicated IT technicians – in particular Kenneth Berg – and scholars and students who have given rise to many improvements through their contributions to APDB and HIPPO.

For any public digital library, the question of sustainability is crucial, both in terms of the format of the data sets and regarding a long-lived infrastructure guaranteeing the maintenance. However, to make it possible to publish a library created from such an advanced application as Collectio online, a secured long-term basic maintenance is not enough, if we would like to be able to add new material or correct errors, in particular concerning connections and relations. For what is gained if the content cannot be trusted? Reliable functions for inserting and correcting the connections of segments (such as connection tables or equivalent systems) and control systems are needed to ensure that the relations are consistent, and no technical errors occur. Collectio offers this and creates long-lived scholarly data sets. Working with original documents, spreadsheets and text documents, is a win-win situation – it combines the benefits of a database with those of a printed edition.

Acknowledgements

This article was supported by funding from the Swedish Research Council (Vetenskapsrådet) for the ‘Knowledge, Magic and Horse Medicine in Late Antiquity’ project (2020-01788), headed by Elisabet Göransson at Lund University. I am most grateful to Kenneth Berg for his tireless efforts in improving Collectio and his constant readiness to share his competence and support.

Notes

- [1] It was supported by funding from Stiftelsen Riksbankens Jubileumsfond from 2009 to 2015. For a survey of the individual projects that were parts of it, see S. Rubenson, *Det tidiga klosterväsendet och den antika bildningen. Slutrapport från ett forskningsprogram*, vol. 9 of RJ:s skriftserie, Makadam, Göteborg, 2016.

- [2] The most known and used general tables are probably those by W. Bousset, *Apophthegmata. Studien zur Geschichte des ältesten Mönchtums*, Tübingen, 1923, and L. Regnault, *Les sentences des pères du désert, troisième recueil & tables*, Solesmes, Sablé-sur-Sarthe, 1976. However, many editors and scholars have made tables for collections and manuscripts within specific languages, such as those by J.-C. Guy, *Recherches sur la tradition grecque des Apophthegmata Patrum*, vol. 36 of *Subsidia hagiographica*, 2^e édition avec des compléments, Société des Bollandistes, Bruxelles, 1984, for the manuscripts of the Greek systematic collection, and C. M. Batlle, *Die "Adhortationes sanctorum Patrum" ("Verba seniorum") im lateinischen Mittelalter: Überlieferung, Fortleben und Wirkung*, Aschendorff, Münster, 1972, for the manuscripts of the Latin systematic collection attributed to Pelagius and John (PJ).
- [3] S. Rubenson, "A Database of the Apophthegmata Patrum", in: T. Andrews, C. Macé (Eds.), *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*, vol. 1 of *Lectio: Studies in the Transmission of Texts & Ideas*, Brepols, Turnhout, 2014, pp. 207.
- [4] For more on this project, see <https://projekt.ht.lu.se/hippo/>.
- [5] Alongside the development of APDB, from 2013 another tool and digital platform has been created as well: *Monastica* – a dynamic library and research tool, <https://monastica.ht.lu.se/>. Up to August 2022, *Monastica* was dependent on APDB. *Monastica* imported all data (with some exceptions) from the master database dump of APDB on the FTP server and, thus, had basically the same contents as APDB. Since then, both APDB and *Monastica* have developed further. As for now (December 2023), *Monastica* has limited possibilities for input and corrections of connections and relations.
- [6] For a recent presentation of the collections of AP, see S. Rubenson, "Apophthegmata Patrum", in: D. G. Hunter, P. J. J. van Geest, B. J. Lietaert Peerbolte (Eds.), *Brill Encyclopedia of Early Christianity Online*, 2018. URL: http://dx.doi.org/10.1163/2589-7993_EECO_SIM_00000239. Further descriptions and references are given in S. Rubenson, "A Database of the Apophthegmata Patrum", in: T. Andrews, C. Macé (Eds.), *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*, vol. 1 of *Lectio: Studies in the Transmission of Texts & Ideas*, Brepols, Turnhout, 2014, pp. 203–205, and B. Dahlman, E. Göransson, K. Åkerman Sarkisian, "A crosslinguistic approach to the study of the Apophthegmata Patrum. A case study of textual variation in the Greek, Latin and Slavonic traditions". To appear.
- [7] For the concept of 'fixed-content miscellanies', see D. Birnbaum, "Computer-Assisted Analysis and Study of the Structure of Mixed Content Miscellanies", *Scripta & e-Scripta* 1 (2003): 15–64, who refers to several articles by A. Miltenova.
- [8] For an introduction to how relational databases can be designed within digital humanities projects, see S. Ramsey, "Databases", in: S. Schreibman, R. Siemens, J. Unsworth (Eds.), *A Companion to Digital Humanities*, Blackwell, Oxford, 2004, pp. 177–197.
- [9] For more on this project, see <https://dvpp.uvic.ca/>.
- [10] M. Holmes, "Getting Along with Relational Databases", *Journal of the Text Encoding Initiative* 14 (2021). URL: <https://journals.openedition.org/jtei/3874>. doi:10.4000/jtei.3874.
- [11] For more on this project, see <https://www.dbbe.ugent.be/>.
- [12] R. Ricceri, K. Bentein, F. Bernard, A. Bronselaer, E. De Paermentier, et al., "The Database of Byzantine Book Epigrams Project: Principles, Challenges, Opportunities", *Journal of Data Mining and Digital Humanities*, Volume: *On the Way to the Future of Digital Manuscript Studies* (2023). URL: <https://hal.science/hal-03833929v3>. doi:10.46298/jdmdh.10244.
- [13] T. Keeline, "The Apparatus Criticus in the Digital Age", *The Classical Journal* 112:3 (2017): 342–363; S. Douglas Olson, "Further Notes on the Apparatus Criticus in the Digital Age", *The Classical Journal* 114:3 (2019): 330–344; F. Fischer, "Representing the critical text", in: Ph. Roelli (Ed.), *Handbook of Stemmatology: History, Methodology, Digital Approaches*, Berlin: De Gruyter 2020, pp. 405–427 (416–427). See also T. Andrews, "Publication of digitally prepared editions", in: Ph. Roelli (Ed.), *Handbook of Stemmatology: History, Methodology, Digital Approaches*, Berlin: De Gruyter 2020, pp. 427–436.
- [14] A relation is typically registered to a Bible verse or a catalogue number, but it can also be registered to base entities in a related source, where the scholar cannot (or does not have the time to) make the correct analysis required for dividing the entities into segments.

AI, Data Curation and the Data Readiness of Heritage Collections: Exploring the Swedish Newspaper Archive at KBLab

Justyna Sikora^{1,*,\dagger}, Chris Haffenden^{2,\dagger}

¹*KBLab, Kungliga biblioteket, Karlavägen 100, 115 26 Stockholm, Sweden*

²*KBLab, as above*

Abstract

The increasing availability of digital material and tools for large-scale computational analysis has produced a growing interest in big data approaches in the humanities and social sciences. However, the vital role of *data curation* as a precondition for such projects remains underappreciated. This paper details the work of KBLab at the National Library of Sweden in testing AI tools to help curate the digitized newspaper archive and make it more amenable to quantitative, machine learning-based research. It provides a description of the library’s newspaper data to offer orientation to researchers interested in the material, before turning to recount the results of our exploration with automated data curation. It concludes by sketching possible next steps for these exploratory efforts, as well as situating this project within a broader recent turn to conceptualize and prioritize the notion of *data readiness*. Its principal argument is in drawing attention to data curation as an essential part of any digital research project, not something prior to or external from the research process.

Keywords

Data curation, data readiness, digitized newspaper archives, document AI, digital research infrastructure

1. Introduction

Digital research presumes digital data. This is a platitude, but bringing it into focus helps illuminate the critical role of infrastructural questions within the research process. While the increasing availability of large-scale digitized corpora and tools for computational analysis has produced a growing interest in operationalizing big data in the humanities and social sciences, significant blindspots remain about the complexities involved in such projects. A specific challenge is dealing with the varying gap between i) the output of the digitization process and ii) the attainment of machine-readable data of sufficient quality to pursue credible research. In short, there persists a lack of recognition of *data curation* as an enabling condition for digital research [1, 2]. This is a problem since, as Lisa Gitelman has argued, “raw data is an oxymoron” [3]; there is no such thing as ready-made data. Instead, data needs to be prepared and curated according

Huminfra Conference 2024, Gothenburg, 10–11 January 2024.

*Corresponding author.

\dagger These authors contributed equally.

✉ justyna.sikora@kb.se (J. Sikora); chris.haffenden@kb.se (C. Haffenden)

🆔 0000-0002-5561-5163 (C. Haffenden)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

to the particular specifications of a given project, making it a necessary practice that invariably demands resources and shapes results.

Data curation "encompasses gathering material, making it discoverable by describing and organizing it, placing it in a context of related information, supporting its use for diverse intellectual purposes, and ensuring its long-term survival" [4]. In the context of digital research, this touches upon a key question of multidisciplinary work focused on big data: how quantitative approaches best interact with more qualitative humanities scholarship [5]? Delineating what she terms the "dangerous art of text mining", Jo Guldi warns that data scientists risk producing "analyses that are empty, biased, or simply false", if they lack awareness of the particular history and context that has formed humanities data [6]. Though a pertinent warning, it tends towards a one-sided version of myopia: emphasizing the shortsightedness of data science while underplaying the occasional data naivety of the humanities. Here we can recall Neil Lawrence's observation that technical project members are "often treated like magicians who are expected to wave a model across a disparate and carelessly collated set of data and with a cry of 'sortitouticus' a magical conclusion is drawn". In outlining the unrealistic expectations placed upon data scientists, he suggested that "[j]ust as extracting drinkable water from the real ocean requires the expensive process of desalination, extracting usable data from the data-ocean requires a significant amount of processing" [7]. Linking questions of digital labour and project efficacy, this highlights the pivotal role of data curation in any effective collaboration between humanities scholars and data scientists.

This paper uses the National Library of Sweden's (*Kungliga biblioteket*, hereafter KB) digitized newspaper archive to discuss data curation as it pertains to making heritage collections available for large-scale research. More specifically, we show how AI tools can be used to automate elements of the curating process and thus cope with large volumes of data. The first part provides a contextual account of KB's data, explaining the what, how and why of the newspaper archive, including problems resulting from the digitization process. The second part details our exploratory work at KBLab, the library's data lab for digital research [8], in testing AI models to curate the digitized newspaper material. We have two key aims with such an account: providing orientation for researchers interested in using KB's newspaper data; and contributing towards a recent trend foregrounding an active approach to data readiness [7, 9, 10].

2. KB's collections as data

As a national library, KB collects and hosts a wide range of data sources: from postcards, radio broadcasts and computer games to more conventional print material like books and ephemera. According to Swedish legal deposit legislation, a copy of every published item is to be stored at the library for the benefit of future users. This makes KB's archives an invaluable resource for researchers, both for the historical depth and breadth of the collections.

Thanks to recent digitization efforts, a growing volume of this material is accessible in digital form. One of the services providing access to part of the digitized collection is tidningar.kb.se. Via this website, users can search through more than 300 years of newspaper material. However, for researchers seeking to conduct larger-scale, programmatic analysis, another on-site service is available at the library's premises via KBLab: the datalab.kb.se platform that makes it pos-

sible to engage with the collections as data, including the extensive newspaper archive (See: github.com/Kungbib/kblab/).

2.1. Opportunities and challenges with digital heritage data

Access to the digitized materials via KBLab enables a new type of research focused on quantitative methods *at scale*. While undoubtedly providing new opportunities, this can create challenges for more qualitatively-inclined scholars, since identifying relevant data for a project can be difficult given the extent of the collections. Moreover, finding pertinent data may only partially solve the problem of locating suitable material. Considering the content of the data is obviously the first step when investigating potential research material. However, another crucial part is determining the data readiness of these sources - i.e. is the data ready to be used as a dataset for research? Are there gaps or duplicates? What sort of data wrangling might be required before it is fit for purpose [7, 10]? Being able to use the data is thus as important as finding it; beyond locating sources, researchers also need to consider any problems these might pose in terms of data handling.

It might be thought that coming to the library premises and accessing the digital collections would be the final step in defining a research topic and seeking suitable resources for a project. Yet taking care of these issues does not necessarily mean the most challenging part of the work is complete. As the data is not always available in a user-friendly format for qualitative scholars, more preparation may be needed before any data analysis can be started.

If we consider text data, it is easy to assume that searching it would be as straightforward as skimming a newspaper to find relevant articles about the topic of interest. Instead, a researcher arriving at the lab to work with the digital collections will encounter a potentially daunting view of rather complex data structures, for parts of the collection that might not previously have been explored. While familiar ground for a data scientist or statistician, dealing with structured data might be a new experience for a humanities scholar with a qualitative background. This reinforces the point above, that finding appropriate research data also requires thinking about the data readiness of the available sources.

In sum, the turn towards quantitative methods from the data sciences can provide a means of dealing with the issue of scaling up research, creating new pathways for humanities research. But it also introduces pressing new challenges and skill requirements not previously part of the humanities toolbox, such as assessing material in terms of data readiness.

2.1.1. What does the newspaper data look like at KBLab?

One of the most requested parts of KB's digital collections is the newspaper data. The library's holdings consist of over 1,900 titles, spanning from 1645 to the present. The newspapers available via datalab.kb.se are scanned and processed using optical character recognition (OCR). As a result of this process, the pages are broken down into bounding boxes and the corresponding text. This also means the data is not structured into clear units such as articles. Instead, it is available as consecutive blocks of text along with their corresponding scanned images and the coordinates pointing to the OCR boxes containing the texts. In other words, an effect of the digitization process - beyond the introduction of OCR errors [11] - is the loss of various metadata we take for granted, i.e. which parts of a page form part of the same article, and which parts of the newspaper

comprise the same section, e.g. "sport" or "culture". This format currently requires a certain level of data literacy to work with. Initial efforts have therefore been made towards preparing the materials into a structured dataset.

2.1.2. Newspapers' structure and the complex issue of layout

Without the constraints of finite resources and a huge volume of complex data, the ideal scenario would be to provide researchers with structured datasets, with the possibility to search for individual articles about certain topics or with specific keywords. However, at present the OCR processing and the resulting structure of the newspaper data pose a substantial challenge to this goal.

A significant part of this challenge is about the stripping of metadata that results from digitizing the physical newspaper, especially reassembling the OCR boxes into a coherent order. From the perspective of machine learning this is about layout analysis, which is a complex task and where newspaper data is more challenging than other document types. When we consider material like receipts or contracts, one page typically contains coherent information belonging to a single document. A newspaper page, by contrast, may consist of multiple articles scattered over the page, and articles spread across multiple pages.

For layout analysis, working with historical newspapers is more straightforward than handling modern newspapers. With the layout of historical newspapers, the articles tend to be organized in long columns on a page. When we look at a page of a modern newspaper, identifying the boundaries of an article might appear easy at first glance, but this is not always a trivial task even for a human. Page layouts often subvert the left-to-right and up-and-down reading logic. Moreover, article paragraphs may describe different topics, making it more challenging to group them.

The most significant aspect determining the complexity of layout analysis of modern newspapers, though, is the evolution of layouts over time and their variation from one newspaper to another; it is far from a standardized, static problem and there is no one-size-fits-all solution. Simple heuristics cannot be applied to extract the text and reconstruct the articles. A more robust approach is needed to dynamically handle the diverse page design.

3. Can AI help with data curation?

With the growing number of documents produced in every aspect of life, the need for automated processing and information extraction is increasingly pressing. Looking at each and every individual document is not feasible anymore. A growing branch of AI that addresses these challenges, and one focused specifically on developing tools for processing various OCR-ed materials, is Document AI [12]. While the greater part of this research has targeted data such as receipts or contracts, any document in PDF format can be a subject for Document AI, including newspapers.

The current state of the newspaper data may, as mentioned, seem chaotic or overwhelming, particularly to those unfamiliar with structured data. To alleviate this, we can test leveraging recent advancements in Document AI. Given the complexity of newspaper layouts, performing

a full-scale layout analysis on this data is overly ambitious. However, one step towards tidying up the collection to present it in a more approachable way is to separate i) body texts from ii) headlines, captions and all other text not considered part of the main text. Insofar as this allows us to obtain the main content of the newspapers, this constitutes an experiment in automating part of the data curation process.

3.1. Image transformer and training of body text model

Inspired by the training objectives of language models such as BERT [13], the transformer architecture has been successfully applied to image processing. Several models such as LayoutLM and BEIT have been pre-trained on numerous images and can be further fine-tuned to solve tasks like document image classification or semantic segmentation.

The aim of integrating image transformers in processing the newspaper data is to distinguish the main body of articles from the surrounding page content. This is a first step towards creating a comprehensive dataset based on the newspaper materials accessible to researchers on-site at KBLab's premises.

3.2. How was the model trained?

To provide a good variation of newspaper layouts, issues from four Swedish newspapers – *Svenska Dagbladet*, *Aftonbladet*, *Dagens Nyheter* and *Expressen* between the years 2010-2021 – were sampled as training data for the model. Afterwards, the newspapers were annotated for two classes: i) boxes containing body text and ii) the rest of the contents on a page, including headers, captions, images etc. In total 64,837 boxes were annotated. The data was then divided into training and test sets, which were subsequently used to fine-tune a Document Image Transformer (DiT) model [14].

The DiT model was pre-trained using a masked image modeling task, meaning a number of inputs were randomly replaced with a [MASK] token. To conduct the pre-training, the input images were divided into non-overlapping patches and converted into visual tokens. The tokens were obtained from a custom discrete variational auto-encoder (dVAE). In contrast to other image transformers, the DiT model was pre-trained on a dataset consisting of 42 million document images to enhance the performance on scanned data. This approach makes the model especially well-suited for processing the OCR-ed newspaper data. The objective of the pre-training was then to predict the masked discrete visual tokens with the output representation.

In this work, we have fine-tuned the base version of the DiT model, which consists of 12 layers of transformers block with a hidden size of 768, to solve the image classification task. As in pre-training, images are split into patches and tokenized before the fine-tuning phase. The sequences of tokenized patches are then used as an input to the model, which outputs probabilities for an image to contain body text or non-body text. Only image features are taken into consideration, i.e. no additional information about textual content is provided to the model.

3.3. Results

After the fine-tuning for 5 epochs, the model achieved 95,5 % accuracy on the test set, which means it correctly assigned body text and non-body text labels to the test examples in 95,5 % of cases. The exemplary output is shown in the appendix below, where examples categorized as body text are marked in green, while the red boxes show the negative predictions (e.g. non-body text). As the accuracy score suggests, the body text was largely correctly recognized, with the model performing particularly well on articles with typical layout, i.e. those consisting of several consecutive paragraphs of text. The model also handles well cases where OCR-boxes resemble the main text but actually contain additional information, such as details about the authors. This applies to the byline in `mathptmx [scaled=.90]helvet courier 2`, for instance: taking into account the graphic features, the model correctly assigns the “body-text” label to all boxes but the last one.

An important factor that influences the results and makes the classification task more difficult is the specificity of the bounding boxes created by the OCR process. Certain of the body text paragraphs are split into multiple boxes, some of which may only contain one word or a sentence. However, these cases are rather rare. An example can be observed in Figure 2 where the lead has been divided into two parts – a longer paragraph and one sentence. The main part of the lead was correctly categorized, while the last sentence was mislabeled. This occurred most likely because a wide but short box containing one or a couple of sentences resembles more closely an image caption than body text.

Despite the model sometimes categorizing parts of the pages incorrectly, we suggest this is an important first step towards preparing the data for research. As the examples in the appendix suggest, text blocks can easily be separated from the headlines, images and other parts of the pages not considered as the main text based on the model’s prediction. For a researcher wishing to pursue an analysis of the newspaper data at scale, and who might thus want this investigation to be based solely on the main body of the newspaper text, this represents a significant move towards greater data readiness for the project. Document AI thus appears a promising tool for helping with data curation.

4. What next?

In a perfect world, researchers working on the newspaper archive at KBLab would access structured datasets that are cleaned and adapted to use as is, with minimal need for processing. While we are far from living in such a world, the latest advancements in AI can help boost data curation and bring us closer to this goal. By testing the latest models within Document AI, our exploratory efforts suggest automated curating is a promising way to improve the data readiness of heritage collections for large-scale research.

Possible future work involves creating a pipeline with various models for processing the data. The architecture could include both a module for recognizing the body texts and one for filtering out adverts. We have already experimented with ad classification: multiple text, image and combined models have been trained to classify adverts, with the best model achieving 97,6% accuracy [15]. Since adverts often contain boxes resembling the main body, excluding boxes marked as body text in adverts would help in isolating the actual articles (even if re-assembling

the OCR boxes into these articles remains to be solved). A further option is enriching the image transformer model with additional information, such as the contents of the boxes and geometric information about the OCR-ed boxes, to turn the classification task into a multimodal problem.

More broadly, this paper has suggested the essential role of data curation within digital research projects. Rather than something prior to or external from a project, we suggest these activities should be treated as integral to the research process – both since they constitute a necessary stage in enabling the research, and since they have a concrete effect on the outcomes, i.e. how data is curated shapes the results, as doing it differently produces different outputs. In highlighting such matters, our case study forms part of a broader recent turn to conceptualize and prioritize the notion of *data readiness*. Thinking more about the readiness of data for large-scale digital research is an excellent starting point for sustainable future collaboration between data science and the humanities.

References

- [1] G. Henry, Data curation for the humanities, in: J. M. Ray (Ed.), *Research data management: Practical strategies for information professionals*, Purdue University Press, West Lafayette, IN, 2014, pp. 347–374.
- [2] A. H. Poole, “a greatly unexplored area”: Digital curation and innovation in digital humanities, *Journal of the Association for Information Science and Technology* 68 (2017) 1772–1781. URL: <https://doi.org/10.1002/asi.23743>.
- [3] L. Gitelman (ed.), *Raw data is an oxymoron*, MIT press, Cambridge, Mass., 2013.
- [4] T. Mu noz, A. H. Renear, *Issues in humanities data curation* (2011). URL: <http://hdl.handle.net/2142/30852>.
- [5] M. Kemman, *Trading Zones of Digital History*, De Gruyter, 2021. URL: <https://www.degruyter.com/document/doi/10.1515/9783110682106/html>.
- [6] J. Guldi, *The Dangerous Art of Text Mining: A Methodology for Digital History*, Cambridge University Press, Cambridge, 2023.
- [7] N. D. Lawrence, Data readiness levels, arXiv preprint arXiv:1705.02245 (2017). URL: <https://doi.org/10.48550/arXiv.1705.02245>.
- [8] L. Börjeson, C. Haffenden, M. Malmsten, F. Klingwall, E. Rende, R. Kurtz, F. Rekathati, H. Hägglöf, J. Sikora, Transfiguring the library as digital research infrastructure: Making kblab at the national library of sweden, *SocArXiv* (2023). URL: <https://osf.io/preprints/socarxiv/w48rf>.
- [9] F. Olsson, M. Sahlgren, We need to talk about data: The importance of data readiness in natural language processing, arXiv preprint arXiv:2110.05464 (2021). URL: <https://doi.org/10.48550/arXiv.2110.05464>.
- [10] M. Hurtado Bodell, M. Magnusson, S. Mützel, From documents to data: A framework for total corpus quality, *Socius* 8 (2022) 23780231221135523. URL: <https://doi.org/10.1177/23780231221135523>.
- [11] J. Jarlbrink, P. Snickars, Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive, *Journal of Documentation* 73 (2017) 1228–1243. URL: <https://doi.org/10.1108/JD-09-2016-0106>.

- [12] L. Cui, Y. Xu, T. Lv, F. Wei, Document ai: Benchmarks, models and applications, arXiv preprint arXiv:2111.08609 (2021). URL: <https://doi.org/10.48550/arXiv.2111.08609>.
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017). URL: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- [14] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, F. Wei, Dit: Self-supervised pre-training for document image transformer, in: *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3530–3539. URL: <https://doi.org/10.1145/3503161.3547911>.
- [15] F. Rekathati, The kblab blog: A multimodal approach to advertisement classification in digitized newspapers, 2021. URL: <https://kb-labb.github.io/posts/2021-03-28-ad-classification/>.

A. Example predictions from the model



Figure 1: Predictions on a page of *Aftonbladet*, where the model has correctly identified body text (green boxes) as opposed to other texts such as headers or adverts (red boxes).



Figure 2: Predictions on a page of *Svenska Dagbladet*, where the model struggled to separate all body from non-body text, i.e. in the advert.

SAOL och svensk språkvetenskaplig infrastruktur – nu och i framtiden

Louise Holmer^{1,*}, Ann Lillieström^{1,*}, Emma Sköldberg^{1,*} and Jonatan Uppström^{1,*}

¹Göteborgs universitet, Institutionen för svenska, flerspråkighet och språkteknologi, Box 200, 40530, Göteborg, Sverige

Abstract

Svenska Akademiens ordlista (SAOL 14, 2015 [1]) spelar en viktig roll inom svensk språkvetenskaplig infrastruktur, något som framkommer i denna artikel. Vidare presenteras preliminära resultat av en undersökning av hur frekventa uppslagsorden i SAOL egentligen är i olika delkorpusar med modern allmänspråklig svenska. För att ordlistan även fortsättningsvis ska kunna användas inom svensk ordforskning, vid språkstudier m.m., men också bli mer central inom språkteknologiska sammanhang, är det avgörande att SAOL:s uppslagsord vilar på vetenskaplig grund, moderna språkteknologiska metoder och uppdaterade korpusmaterial. Fokus i artikeln ligger på de uppslagsord som inte finns belagda i korpusmaterialet, och som därmed kan tänkas mönstras ut inför den kommande femtonde upplagan.

Keywords

SAOL, lexikografi, Kubord 1–2, lemmaurval, Språkbanken Texts forskningsinfrastruktur

1. Inledning

SAOL, *Svenska Akademiens ordlista* [1], har givits ut sedan 1874, dvs. under närmare 150 år. Verket har spelat och spelar fortfarande en mycket viktig roll inom svensk språkvetenskaplig infrastruktur.

En ny upplaga av ordlistan är under utarbetande och i samband med detta ägnas, naturligt nog, uppslagsorden (lemmana) i verket en hel del uppmärksamhet [2], [3]. För att ordlistan även fortsättningsvis ska upplevas som modern och fylla de funktioner som den hittills har gjort, är det viktigt att uppslagsorden är uppdaterade och att de kan betraktas som (någorlunda) representativa för samtida svenskt allmänspråk.

En lexikografisk metod som används för att undersöka hur väl uppslagsorden i en ordbok, dvs. lemmalistan, sammanfaller med det ordförråd som den aktuella ordboken ska täcka, är att jämföra lemmalistan med de ord som förekommer i en eller flera utvalda korpusar. Denna metod har SAOL-redaktionen använt tidigare för att bilda sig en uppfattning om lemmalistas innehåll och skapa listor med möjliga nyord och s.k. utmönstringskandidater inför nästa upplaga [3].

Huminfra Conference 2024, Gothenburg, 10–11 January 2024.

*Corresponding author.

†These authors contributed equally.

✉ louise.holmer@svenska.gu.se (L. Holmer); ann.lilliestrom@gu.se (A. Lillieström);

emma.skoldberg@svenska.gu.se (E. Sköldberg); jonatan.uppstrom@svenska.gu.se (J. Uppström)

ORCID 0009-0009-6763-6672 (L. Holmer); 0000-0002-0146-4697 (A. Lillieström); 0009-0004-9885-5847 (E. Sköldberg)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Arbetet med SAOL bedrivs inom projektet Svenska Akademiens samtidsordböcker och finansieras av Svenska Akademien. De lexicografer och forskningsingenjörer som arbetar inom projektet är anställda vid Göteborgs universitet (Institutionen för svenska, flerspråkighet och språkteknologi) och verksamma inom forskningsenheten Språkbanken Text [4]. Inom projektet finns goda möjligheter att på vetenskaplig grund, med nya språkteknologiska metoder och uppdaterade korpusmaterial dels sälla fram lämpliga kandidater bland nyorden [5], dels mönstra ut uppslagsord som inte längre används i svenskt skriftspråk.

Det övergripande syftet med föreliggande artikel är dels att kortfattat presentera lemmalistan i den senaste upplagan av ordlistan, SAOL 14 (2015), dels att undersöka hur innehållet i lemmalistan står sig i förhållande till det ordförråd som kännetecknar svenskt skriftspråk av idag, representerat av ett utvalt korpusmaterial. Fokus ligger på utmönstringskandidater, dvs. uppslagsord som av frekvensskäl kan komma att strykas ur ordlistan. I artikeln diskuterar vi bl.a. vad dessa utmönstringskandidater har för gemensamma drag.

2. Lemmalistan i SAOL 14

SAOL är som nämnts en ordlista med mångåriga anor. Den utvecklades som en sorts ram till den större och mer historiskt inriktade *Svenska Akademiens ordbok* (SAOB, [6]) och mellan 1874 och 2015 har den kommit ut i sammanlagt 14 upplagor. Den första upplagan innehöll ungefär 34 000 lemman medan den senaste är uppe i drygt 126 000 lemman [3], [7].

Sedan 1980-talet utarbetas ordlistan vid Göteborgs universitet. Från att vara en mer manuell framställd tryckt bok, har den utvecklats till att bli en korpusbaserad datamängd publicerad som tryckt bok, i appform och på webben via ordboksportalen Svenska.se [8]. På SAOLhist.se [9] är de tidigare upplagornas uppslagsord sökbara samtidigt.

Enligt Gellerstam [10] råder det en föreställning bland många av ordlistans användare att SAOL innehåller alla svenska ord. Gellerstam konstaterar dock att även om verket innehåller många uppslagsord rör det sig ändå framför allt om ett urval. Svensén [2] lyfter fram att SAOL fungerar som inofficiell norm för stavning och böjning av svenska ord. SAOL är alltså en s.k. produktionsordbok ägnad att användas vid produktion av text. Ordlistan ger också upplysningar om ordbildning, t.ex. genom att stor respektive liten ordledsgräns markeras i uppslagsorden, och genom att ta med sammansättningar som visar s.k. fogemorfem i tveksamma fall. Enligt inledningen till senaste upplagan bygger de rekommendationer som ges i SAOL 14 på å ena sidan den rådande språkvårdsideologin, å andra sidan det rådande språkbruket [1].

Via Svenska Akademiens lexikala databas, Salex, som ingår i den lexikala plattformen Karp [11] inom Språkbanken Texts forskningsinfrastruktur, går det att få mer specifika uppgifter om lemmauppsättningen i SAOL 14. De största ordklasserna fördelar sig enligt följande bland de ca 126 000 uppslagsorden: drygt 93 000 substantiv, drygt 17 000 adjektiv, drygt 10 500 verb (partikelverb och reflexiva verb inräknade) och drygt 1 200 adverb. För att illustrera vilka typer av ord som ingår i ordlistan listas i figur 1 några av de uppslagsord som inleds med *hå*-.

Svenskan är ett språk som uppvisar nästan obegränsade möjligheter till sammansättning och avledning. Det märks på en ordlista som SAOL 14 i och med att nästan 90 000 av uppslagsorden är sammansatta. Ungefär 34 000 lemman är simplex (osammansatta), och bland dem återfinns ett stort antal avledningar. Exempel på sammansättningar och avledningar återfinns

hågad, hågkomst, håglös, håglöshet, hågsa, håhå, håhåjaja, håken, håkåring,
håll, håla1, håla2, hålfot, hålfotsinlägg, hålig, hålighet, håljärn, hålkaka, hålkant,
hålkindad, hålkort, hålkrus, hålkäl, hålkälad, hålkälshyvel, håll, hålla, hållande,
hållare, hållas, hållbar, hållbarhet, hållbarhetsprov, hållbarhetstid, hålldam, hållen,
hållfast, hållfasthet, hållfasthetsberäkning, hållfasthetslära, hållfasthetsprov,
hållhake, hålligång, hållning, hållningsfel, hållningslös, hållningslöshet, hållplats,
hållpunkt, hålltid, hålremsa, hålrot, hålrum, hålslag, hålslev, hålslå, hålstans,
hålstansning, hålsöm, håltagning, håltegel, håltimme, håltolk, håltång, hålven,
hålväg, hålögd, hålögdhet

Figur 1: Exempel på uppslagsord på *hå-* i SAOL 14 (2015) [1].

bland uppslagsorden i figur 1 (se t.ex. *hållfast*, *hållfasthet*, *hållfasthetsberäkning*, *hållfasthetslära* och *hållfasthetsprov* respektive *hållning*, *hållningsfel*, *hållningslös* och *hållningslöshet*). Bland exempelorden ovan finns också ett fåtal enkla ord såsom *håll*, *håla* (1, 2), *håll* och *hållas*.

Som redan påpekats fyller SAOL en viktig funktion inom flera områden. Dåvarande huvudredaktören för SAOL 14, professor Sven-Göran Malmgren, skriver: ”Svenska Akademiens ordlista (...) är inte en stor ordbok i fysisk mening. Men den är en svensk klassiker med många upplagor på nacken. (...) En ny upplaga av SAOL väcker alltid ett betydande intresse, större än t.o.m. förstaupplagor av alla andra svenska ordböcker kan glädja sig åt.” [12]. Josephson [13] påpekar att SAOL utgör det enskilt starkaste normeringsinstrumentet för svenskans del medan Vikør [14] konstaterar att ordlistan har ”stor autoritet i praxis”. Språkrådet hänvisar också ofta till SAOL när det gäller frågor om stavning och böjning (se t.ex. Frågelådan i svenska 2023 [15]). Därutöver ingår SAOL i kurslitteraturen inom såväl språkvetenskapliga ämnen som mer praktiska utbildningar (t.ex. språkkonsultprogram), men innehållet i ordlistan åberopas också inom juridiska utbildningar, receptarieprogram etc. [16]. SAOL har många olika användare som ofta konsulterar ordlistan via svenska.se. Detta framgår bland annat av den statistik och de loggfiler över sökningar som redaktionen har tillgång till internt. Däremot är användningen av SAOL:s innehåll, såvitt vi vet, fortfarande begränsad i språkteknologiska sammanhang.

3. Metodologiska överväganden och val av jämförelsekorpusar

För att SAOL ska kunna göra anspråk på att vara en samtidsordbok inkluderas ett antal nya ord i varje upplaga medan ett antal mönstras ut [3], [17]. Också ur digitala ordböcker behöver föråldrade ord mönstras ut, även om det inte är lika nödvändigt som i tryckta ordböcker. Med tanke på att den kommande upplagan av SAOL kommer att ges ut i tryckt form, förutom att publiceras digitalt, är det traditionella lexikografiska tankesättet att man inför en ny upplaga av platsskäl ofta måste ta bort motsvarande mängd som man lägger till – ett slags nollsummespel – fortfarande relativt giltigt (se vidare t.ex. [18], [7]).

Ur SAOL-perspektiv är det viktigt att såväl tillägg av nya uppslagsord som utmönstring av

befintliga sådana sker på vetenskaplig grund, så långt det är möjligt. Kilgarriff [19] konstaterar dock följande: "building a headword list for a new dictionary (or revising one for an existing title) has never been an exact science" (se även [20] om lemmaselektion i förhållande till korpusinnehåll). Med hjälp av de språkteknologiska verktyg som numera står till buds inom Språkbanken Texts forskningsinfrastruktur, är förutsättningarna för ett vetenskapligt grundat arbetssätt med förankring i språkteknologi mycket goda. En tidigare undersökning [3] visade att en relativt hög andel ord i SAOL 13 (2006) [21] trots allt inte användes i moderna texter, inte ens med vad som för den tiden ansågs vara ett större korpusmaterial. Några av de ord som inte användes mönstrades ut inför upplaga 14, men ordlistan torde alltså, trots frekvensanspråk, fortfarande innehålla en viss andel lågfrekventa uppslagsord.

Den metod som ordboksprojektet använder går således ut på att jämföra innehållet i lemmalistan med innehållet i en korpus, en korpus vars innehåll så långt det är möjligt ska vara representativt för svenskan av idag. Vid utarbetande (eller val) av korpus behöver man bl.a. beakta korpusens storlek, dess innehåll (texttyper/genrer, förhållande mellan tal- och skriftspråk m.m.) och under vilken tidsperiod de ingående texterna ska vara publicerade (se vidare [22]). Det är eftersträvansvärt att korpusen är så balanserad som möjligt. Det krävs dessutom att innehållet i korpusarna uppdateras löpande med nyare textmaterial för att lexikograferna (och förstas andra språkvetare) ska kunna studera språkets utveckling.

Att sätta samman den ultimata korpusen är knappast en enkel uppgift. Kilgarriff [19] menar t.ex. att "the corpus is never good enough" och att en korpus alltid kommer att innehålla "noises and biases". Av upphovsrättsliga skäl har tillgången till uppdaterade svenska korpusar inte varit optimal på senare år. Genom ett pågående samarbete mellan KB-Labb vid Kungliga biblioteket och Språkbanken Text har emellertid svenska ordforskare nu fått tillgång till moderna pressmaterial från flera olika tidningar och detta utan att upphovsrätten kränks. Materialen går under beteckningarna Kubord 1 och Kubord 2 och de är bl.a. tillgängliga via ordforskningsplattformen Korp [23]. Kubord 1 och 2 ger ordfrekvenser men däremot inte kontexter, och det senare är en klar nackdel i flera lexikografiska sammanhang. Vid revidering av en lemmalista räcker emellertid frekvensangivelserna långt [24].

I föreliggande studie har vi valt att jämföra lemmalistan i SAOL 14 med innehållet i en korpus bestående av tidningstexter (såväl morgon- som kvällstidningar från Kubord 1), romanliknande material (texter som skribenterna själva har publicerat på sajten Poeter.se), innehållet i Svenska Wikipedia och texter publicerade på SVT Nyheter. De olika delmaterialens respektive storlekar och publiceringsår framgår av tabell 1.

Det totala korpusmaterialet uppgår till närmare 3 miljarder token (se tabell 1). Det är avsevärt mycket mer än antalet token i de korpusar som lemmalistorna i tidigare upplagor av SAOL har jämförts med (se [3] där lemmalistan i SAOL 13 jämförs med en korpus på 250 miljoner ord). Ur ordlistans perspektiv är det en fördel att de ingående materialen är moderna och av olika slag eftersom det bl.a. gynnar spridningen (med avseende på innehåll och stil) när det gäller orden i de aktuella texterna. Men givetvis vore det en fördel om exempelvis fler skönlitterära alster ingick i materialet.

Tabell 1

Innehåll i de delkorpusar i Korp som lemmalistan i SAOL 14 har jämförts med

Källa	Data från (år)	Antal token
Aftonbladet	2010-2021	510 091 208
DN	2010-2021	410 961 454
Expressen	2010-2021	514 255 014
GP	2010-2021	283 219 014
Svenska Dagbladet	2010-2021	375 012 173
Sydsvenskan	2010-2021	329 707 101
poeter.se	uppdaterad 2019	106 196 502
Svenska Wikipedia	uppdaterad 2023	106 196 502
SVT Nyheter	2010-2021	188 875 029
Totalt		2 908 466 992

4. Resultat: SAOL:s lemmalista versus ett modernt korpusmaterial

En preliminär jämförelse mellan innehållet i lemmalistan i SAOL 14 och de ord som förekommer i de aktuella delkorpusarna visar att ca 121 000 av de 126 000 uppslagsorden uppträder i åtminstone någon av sina böjningsformer i textmaterialen. Det gäller exempelvis en absolut majoritet av de närmare 70 uppslagsord som visas i figur 1 ovan (inklusive *hållfast*, *hållfasthet*, *hållfasthetsberäkning*, *hållfasthetslära* och *hållfasthetsprov* respektive *hållning*, *hållningsfel*, *hållningslös* och *hållningslöshet*). Resterande uppslagsord, ca 4 100, används däremot inte. Detta motsvarar drygt 3,5 av de aktuella lemmarna. Detta resultat kan jämföras med [3], dvs. att cirka 10 % av uppslagsorden i SAOL 13 inte användes i den då aktuella jämförelsekorpusen. Det är dock viktigt att komma ihåg att den korpus som vi använt här är mer än 10 gånger större än den då aktuella korpusen, och att de textmaterial som använts i denna studie är av mer varierande slag. I figur 2 visas ett antal ord på bokstaven *h*- som inte används i de delkorpusar som presenteras i tabell 1.

Bland de ord som listas i figur 2 återfinns som synes såväl substantiv som adjektiv och verb. Bland dem ingår t.ex. redan nämnda *hålkälad* och *håltolk*. Båda dessa ord har funnits med sedan den nionde upplagan (från 1950), vilken haft det största ordförrådet hittills (155 000 uppslagsord, jfr [10]).

I samband med en ny ordboksupplaga får nyorden som lagts in vanligen mycket uppmärksamhet: lexikografiskt på så sätt att de lyfts fram i verkets kringtexter, men också medialt och bland ordboksanvändarna [19], [7]. De ord som mönstras ut uppmärksammas också, men vanligtvis inte i motsvarande omfattning. I inledningen till SAOL 14 noteras det att drygt 9 000 uppslagsord har mönstrats ut i förhållande till den tidigare upplagan. Kännetecknande för dessa ord är att de är ”gamla”, ”har bedömts som mindre relevanta”, anses ”föråldrade” eller ”mindre värdefulla” [1].

En kvalitativ granskning av hela listan med ord som inte används i de korpusar vi jämfört med (s.k. icke-träffar) ger vid handen att det rör sig om lite olika typer av ord. I förteckningen med icke-träffar finns det t.ex. en ansenlig mängd sammansättningar (se bl.a. *högprislinje* och *hörbarhetsgrad* i figur 2). I synnerhet finns det många flerledade sammansättningar bland de ord som inte återfinns i delkorpusarna. Några fler exempel är *arbetsavtalsförhållande*, *bofinksbo*, *boupp-*

hypostasera, hypotisering, hytteknik, hålkälad, håltolk, hårdkokning, hårdskalig, hårdsnö, hårdstoppad, hårdvallshö, hårdvindsbåt, hårkärl, hårsbredd, håvgång, häftespris, häftförband häftmedel, häftremsa, häftsträck, hägnadsvirke, häktningsdom, hållespring, hälsingländsk, hälsningstala, hälsningstalare, hämtköp, hämtrabatt, hängefjäll, hängsjuka, hängslestropp, härdmedel, härvtråd, hästkrubba, hävlig, hävrörelse, häxmästeri, högbenthet, högerspiral, högervridande, höghöjdsraket, högindustrialiserad, högiva, högkarmad, högkyrka, högmodas, högprislinje, högrationaliserad, högryggig, högskattepolitiker, högstdensamma, högtals, högtgående, högviktig, höjdvind, hönkyckling, hönskyckling, hönsminne, hönsning, hörbarhetsgrad, hörselsvag, hörselvårdsassistent, hövlighetsbetygelse, hövålmsfrisyr

Figur 2: Exempel på uppslagsord på *h*- i SAOL 14 (2015) [1] som inte finns i de aktuella korpusarna.

teckningsinstrument, förskottsinnehållning, handräckningsmanskap, kontrollstationsförhandling och *riksdagsmannaupdrag*. Av dessa ord skulle de tre sistnämnda utan vidare kunna mönstras ut, medan de andra kan kräva ytterligare lexikografiska utredningar. Orden *arbetsavtalsförhållande, bouppteckningsinstrument* och *förskottsinnehållning* är t.ex. vanligare i finlandssvenska texter, och SAOL har också inkluderat ett antal finlandssvenska ord i de senaste upplagorna. Ordet *bofinksbo* ska troligen illustrera att sammansättningar med *bofink-* använder *foge-s*.

Vidare finns det många avledningar bland icke-träffarna, trots att det enligt [17] mönstrades ut en hel del ”intuitivt lite udda” avledningar inför SAOL 14. Bland dem ingår många relativt sökta ord med avledningssuffixet *-het*, t.ex. *skvalleraktighet, skäggighet, snubblighet* och *överkloket*. Där finns också ett antal exempel på feminina beteckningar på *-inna* och *-erska*. Gellerstam [10] poängterar att ett stort antal sådana ord redan hade strukits i samband med tidigare revideringar av SAOL som en konsekvens av att svenskan blivit mer könsneutral (jfr [25], [13]). Kanske är det nu dags att uppslagsorden *sexmästarinna, givarinna* respektive *bondfångerska, folkdancerska, fördancerska, giftblanderska, plockerska* och *tjuserska* tas bort. Hädanefter kommer de i sådana fall att bli sökbara i SAOLhist [9] på samma sätt som tidigare utmönstrade ord.

Utöver dessa exempelord finns det många icke-träffar som antingen utgörs av äldre beteckningar som kommit att ersättas av nya eller beteckningar på föremål, företeelser, aktiviteter m.m. som inte längre är vanliga i det omgivande samhället. Exempel är *hörselvårdsassistent* och *hövålmsfrisyr* men också *skråtobak, smärtingsko* och *sparkasseräkning*. När användningen av de aktuella föremålen minskar, modet ändras etc., går förstås också bruket i text av de aktuella beteckningarna tillbaka.

Trots att ordlistan ska innehålla allmänspråkliga svenska ord finns det en hel del fackord i lemmalistan, bl.a. *slussventil, spanjolettstång, splintborre* och *trogloodytisk*. Det faktum att dessa ord inte ens är med i den delkorpus som består av Wikipedia är ett tecken på att de kanske inte ska ingå i det urval av ord som SAOL trots allt utgör.

Slutligen torde användningen av vissa ord av orden, som *eskimåtröja* och *fruntimmerssysla*,

ha gått tillbaka för att de av olika anledningar anses mindre lämpliga eller har fått en negativ klang. Den nuvarande redaktionen kan fortsätta på den redan inslagna vägen med att bidra till ett mer tolerant och jämställt samhälle (se vidare [25] som lyfter fram denna aspekt i sin recension av SAOL 14). Det bör dock noteras att det i förteckningen med s.k. icke-träffar också finns svenska ord som troligtvis används, men kanske främst i andra sammanhang än dem som täcks av de aktuella delkorpusarna (t.ex. *krassefrö*, *stöldskyddsmärkt* och *vitlökskapsel*). Givetvis kommer bruket av orden i förteckningen med icke-träffar att genomgå ytterligare lexikografiska utredningar innan en eventuell utmönstring av obsoleta ord sker inför nästa upplaga.

5. Slutord

Svenska Akademiens ordlista, SAOL, är en ordbok som många svenskar känner till. Den har en mycket väletablerad position bland svenska lexikografiska verk. Ordlistan används i undervisning, konsulteras av personer som skriver i tjänsten, språkforskare etc., vilket utgör några av skälen till att just SAOL är så viktig inom svensk språkvetenskaplig infrastruktur.

Gellerstam [10] menar att uttjänta ord ofta får ”stå kvar i ordböcker som stumma vittnen från en annan tid, eventuellt med en markering som ’mindre brukligt’”. Han fastslår också att benägenheten att rensa ut gamla ord i tidigare upplagor har varit mindre än önskan att introducera nya. Det är dock av största vikt att SAOL:s lemmalista hålls uppdaterad för att ordlistans kvaliteter ska bibehållas (jfr [26]). Vår undersökning visar att drygt 3,5 % av uppslagsorden i SAOL inte används i en korpus på närmare 3 miljarder ord. Många av dessa uppslagsord passar bättre i mer historiskt inriktade resurser som SAOB eller SAOLhist. Tillgången till Språkbanken Texts forskningsinfrastruktur säkrar och effektiviserar SAOL-redaktionens arbete med lemmaurvalet.

Förhoppningsvis kan SAOL användas mer i språkteknologiska sammanhang i framtiden, men en viktig förutsättning för detta är bland annat att innehållet i ordlistan blir fritt tillgängligt, dvs. under en öppen licens.

Acknowledgments

Detta arbete har möjliggjorts tack vare Nationella språkbanken och Huminfra, båda finansierade av Vetenskapsrådet (20182024, kontrakt 2017-00626; 20222024, kontrakt 2021-00176) och deras ingående samarbetsorganisationer, samt av projektet Svenska Akademiens samtidsordböcker, finansierat av Svenska Akademien.

Referenser

- [1] SAOL 14, Svenska Akademiens ordlista, 14 ed., Norstedts, Stockholm, 2015.
- [2] B. Svensén, Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik, 2 upplagan, Norstedts, Stockholm, 2004.
- [3] S. Berg, L. Holmer, E. Sköldberg, Time to say goodbye? On the exclusion of solid compounds from the Swedish Academy Glossary (SAOL), in: A. Dykstra, T. Schoonheim

- (Eds.), Proceedings of the 14th EURALEX International Congress, Fryske Akademy, Leeuwarden/Ljouwert, 2010, pp. 567–576.
- [4] Språkbanken Text, 2023. URL: <https://spraakbanken.gu.se/>.
- [5] E. Sköldberg, Hur fångar vi upp svenskans nya ord med hjälp av Kubord?, 2022. Språkbanken Text bloggpost: <https://spraakbanken.gu.se/blogg/20221128-hur-fangar-vi-upp-svenskans-nya-ord-med-hjalp-av-kubord/>.
- [6] SAOB, Svenska Akademiens ordbok, Gleerups, Lund, 1898–. URL: <https://www.saob.se/>.
- [7] L. Holmer, Neutrala substantiv på -ande i text och ordbok, Meijerbergs arkiv för svensk ordforskning, 47, Göteborg, 2022.
- [8] Svenska.se, Svenska Akademiens ordboksportal, 2023. URL: <https://svenska.se/>.
- [9] SAOLhist, 2023. URL: <https://spraakbanken.gu.se/saolhist/>.
- [10] M. Gellerstam, SAOL i många upplagor, in: M. Gellerstam (Ed.), SAOL och tidens flykt: Några nedslag i ordlistans historia, Norstedts, Stockholm, 2009, pp. 53–83.
- [11] Karp, Språkbankens lexikala plattform, 2023. URL: <https://spraakbanken.gu.se/karp/>.
- [12] S.-G. Malmgren, Svenska Akademiens ordlista genom 140 år: mot fjortonde upplagan, LexicoNordica 21 (2014) 81–98.
- [13] O. Josephson, Språkpolitik, 2 uppl., Morfem förlag, Stockholm, 2022.
- [14] L. Vikør, Rettskrivingsordbøker som normeringsreiskapar i Norden, in: L.-G. Andersson, O. Josephson, I. Lindberg, M. Thelander (Eds.), Språkvård och språkpolitik. Svenska språknämndens forskningskonferens i Saltsjöbaden 2008, Norstedts, Stockholm, 2010, pp. 304–322.
- [15] Språkrådet, Frågelådan i svenska 2023: Hur kan jag själv söka svar på språkfrågor?, 2023. URL: <https://frageladan.isof.se/visasvar.py?svar=79917>.
- [16] A. Blücker, Juridiska – ett nytt språk? En studie av juridikstudenters språkliga inskolning, Ph.D. thesis, Uppsala universitet, Uppsala, Sverige, 2010.
- [17] S. Berg, S.-G. Malmgren, Varför travintresse ska ut och fotbollsintresse ska in, Språkbruk 21 (2010). URL: <https://sprakbruk.fi/artiklar/varfor-travintresse-ska-ut-och-fotbollsintresse-in/>.
- [18] M. Rundell, From print to digital: Implications for dictionary policy and lexicographic conventions, Lexikos 25 (2015) 301–322.
- [19] A. Kilgarriff, Using corpora as data sources for dictionaries, in: H. Jackson (Ed.), Bloomsbury Companion to Lexicography, Bloomsbury, London/New York, 2013, pp. 77–96.
- [20] L. Trap-Jensen, Researching lexicographical practice, in: H. Jackson (Ed.), Bloomsbury Companion to Lexicography, Bloomsbury, London/New York, 2013, pp. 19–30.
- [21] SAOL 13, Svenska Akademiens ordlista, 13 ed., Norstedts, Stockholm, 2006.
- [22] B. Atkins, M. Rundell, The Oxford Guide to Practical Lexicography, Oxford University Press, Oxford, 2008.
- [23] Korp, Språkbankens ordforskningsplattform, 2023. URL: <https://spraakbanken.gu.se/korp/>.
- [24] M. Forsberg, J. Sikora, E. Sköldberg, Words unboxed: Discovering new words with Kubord, 2023. KB-Labb bloggpost: <https://kb-labb.github.io/posts/2023-08-29-kubord/>.
- [25] B. Silén, Den fjortonde upplagan av SAOL, LexicoNordica 23 (2016) 283–299.
- [26] K. Blensenius, L. Holmer, E. Sköldberg, SAOL 14 som rättesnöre – diskussion om den senaste upplagan, LexicoNordica 28 (2021) 39–58.

Curating a historical source corpus of 20th century patient organization periodicals

Gijs Aangenendt^{1,*}, Maria Skeppstedt² and Ylva Söderfeldt¹

¹Uppsala University, Department of History of Science and Ideas, Uppsala, Sweden

²Uppsala University, Centre for Digital Humanities and Social Sciences, Department of ALM, Uppsala, Sweden

Abstract

Acting out Disease: How Patient Organizations Shaped Modern Medicine (ActDisease) explores the history of patient organizations in 20th century Europe. By combining traditional historiographic methods with text mining techniques, the project aims to shed light on how patient organizations co-constructed concepts of and management of disease. Part of the project is to digitize print sources and build a digital corpus for historical text mining. The corpus consists of periodical publications from selected British, French, German and Swedish patient organizations, a type of material that poses a number of challenges in scan quality, layout, and lack of consistency. This paper discusses the technical process of building the ActDisease corpus from digitizing patient organization periodicals to OCR post-processing. It touches upon the methodological questions and challenges of curating a corpus of fragmented and heterogeneous historical source material tailored to a specific project.

Keywords

Corpus curation, historical text digitization, OCR processing, patient organizations

1. Introduction

A well-known challenge in writing the history of medicine is that the sources available regarding past experiences of illness, management of disease, and ideas about health overwhelmingly derive from other people than those who were actually sick. Medical literature, health administration files, and patient records all share fundamental limitations as historical sources to how most people felt, thought, and acted around health and illness. Since the 1980s, historians of medicine have debated how to overcome this obstacle, whether it is possible to capture such a thing as “the patient’s view” or if a historical patient-as-subject can be said to exist at all [1, 2, 3]. Considering this crucial problem in the field, it is surprising that a large body of printed media produced by and for people living with disease has so far been mostly neglected by historians: Beginning in the late 19th century, clubs and organizations for people suffering from particular illnesses began forming in the US and Europe and by the 1940s, there was a considerable number of such patient organizations with remarkable resources and influence. Contrary to the belief held in most research on patient advocacy, the patient organization as a historical phenomenon hence far predates the 1960s and -70s “new social movements” [4].

huminfra Conference 2024, Gothenburg, 10–11 January 2024.

*Corresponding author.

✉ gijs.aangenendt@idehist.uu.se (G. Aangenendt); maria.skeppstedt@abm.uu.se (M. Skeppstedt);

ylva.soderfeldt@idehist.uu.se (Y. Söderfeldt)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Since patient organizations from early on issued printed publications, excellent source material is available for studying their aims, activities, and ideas. Not only do these sources reflect life with a particular disease – for example diabetes – from a viewpoint closer to the actual everyday realities of patients, they also profoundly shaped the experience of living with that illness. In our project, ActDisease, we aim to study how patient organizations co-constructed concepts of and management of disease in the 20th century.

The aim of the project is not to write the history of any particular disease, nor of a specific healthcare system, but of the patient organization as a historical actor. Traditional historiographic methods based on close reading alone therefore do not suffice. Our hypothesis is that illness characteristics are expressed in the aggregated vocabularies of the patient organization periodicals, and hence that their variance across lay and professional contexts, between countries and specific illnesses, as well as their transformations over time can be tracked through text mining techniques.

Other than professional medical literature, however, these sources have hitherto not been available in digital form. Rather than confining ourselves to already digitized material, we are digitizing the periodicals of a selected number of patient organizations. Digitizing the sources within the project enables a higher level of transparency in the methodology. Any corpus curation, including ours, involves a myriad of decisions and steps that shape the material in ways that have consequences for the analysis. By curating our own corpus in collaboration with in-house computer scientists, the historians in the project do not need to rely on selections and technical decisions made by other entities like libraries, publishers, or tech companies, which otherwise threaten to affect the analysis in unknown ways [5]. Nevertheless, our process also involves potential sources of selection bias and unknown distortion, as we will discuss below.

2. Project team

The ActDisease project assembles an interdisciplinary research team which in its current form consists of the PI, a research engineer and a research assistant. In 2024, two history postdoctoral researchers, focusing on Great Britain and France respectively, are joining the project as well as one Digital Humanities postdoctoral researcher focusing on methodological development in the field of historical text mining.

The project is conducted in close collaboration with the Centre for Medical Humanities (CHM) and the Centre for Digital Humanities and Social Sciences (CDHU) at Uppsala University. The main support provided by CDHU is 1) access to a pool of qualified research engineers, with expertise in relevant areas such as Optical Character Recognition (OCR), Natural Language Processing (NLP), webscraping, and front- and backend development, and 2) access to computational resources for data processing, data storage, and data backup. Additionally, the collaboration provides the opportunity to create synergies with other ongoing research projects that CDHU is involved in. For example, the research infrastructure project Communicating Medicine: Digitalisation of Swedish Medical Periodicals, 1781–2011 (SweMPer). Both projects concern medical periodicals and have overlapping technical and infrastructural needs which CDHU can provide. Considering the similarities, the aim is to integrate the ActDisease and SweMPer materials into a shared database. Lastly, the collaboration allows us to make tools and models for text processing and exploration developed within the project available to a wider audience as Huminfra resources.

Table 1

Overview of the ActDisease corpus for Swedish and German patient organization periodicals

<i>Patient organization (Disease)</i>	Periodical	Period (Number of pages)
Sweden:		
<i>De lungsjukas Riksförbund</i> (Lung disease, later also heart disease)	Status	1938-1991 (16 790)
<i>Riksförbundet för sockersjuka</i> (Diabetes)	Diabetes	1949-1990 (8 891)
<i>Riksförbundet för mot astma och allergi</i> (Allergies and asthma)	Allergia	1957-1990 (4 054)
Germany:		
<i>Deutscher Diabetiker Bund</i> (Diabetes)	Diabetiker Journal	1951-1990 (19 324)
<i>Deutsche Multipel Sklerose Gesellschaft</i> (Multiple sclerosis)	MS Gesellschaft	1954-1990 (5 646)
<i>Deutscher Allergiker und Asthmatiker Bund</i> (Allergies and asthma)	Der Allergiker Jahresberichte	1959-1985 (2 397) 1901-1972 (8 529)

3. Constructing a corpus

The criteria for selecting the patient organizations aimed at covering as much of the 20th century as possible and to include the main European languages (English, German, French) plus Swedish. We selected for the study the two or three oldest patient organizations from each country that had issued a periodical publication (newsletter, magazine, annual report). The earliest publication started in 1899 and the most recent in 1959. We digitized every selected series from its first preserved issue up until and including the year 1990.

Compared to other datasets typically used for historical text mining, like collections of books, parliamentary print, or newspapers, the material of the patient organizations is relatively small. For Sweden and Germany, the ActDisease corpus consists of periodical publications from three Swedish and three German patient organizations (Table 1). In the near future, the corpus will be extended with periodical publications from British and French patient organizations which are currently being digitized. In total, our corpus will comprise about 150.000 pages but the size of the individual series varies considerably. Some volumes and issues are missing, creating gaps in the series. The format and layout varies greatly between the series and also over time. Furthermore, the patient organization journals contain a very diverse range of texts and images: besides regular articles also advertisements, crossword puzzles, cartoons, charts, lists and much else. Finding ways to meaningfully compare texts of such diverse volume and type, as well as in different languages, is one of the objectives of the project.

The scanned periodicals were delivered as PDF files. A utility script developed by CDHU using ImageMagick was used to extract the image data from the PDF files into individual image files [6]. These image files formed the basis for OCR processing. To be able to integrate the ActDisease and SweMPer materials into a shared database in the future, a joint consistent filename scheme was adopted, including information such as periodical name, year, volume, and issue.

4. Performing OCR

The OCR processing stage posed several methodological questions and challenges originating from the scan quality, evolving design/layout of the periodicals, and the intended uses of the OCR output. As the decisions taken during this stage influenced how the periodicals could be used in the future, these questions and challenges had to be addressed collectively by the research team. For the output, we selected formats that would support traditional historiographic methods, computer-based analysis, as well as the creation of a database: TXT, searchable PDF, and XML.

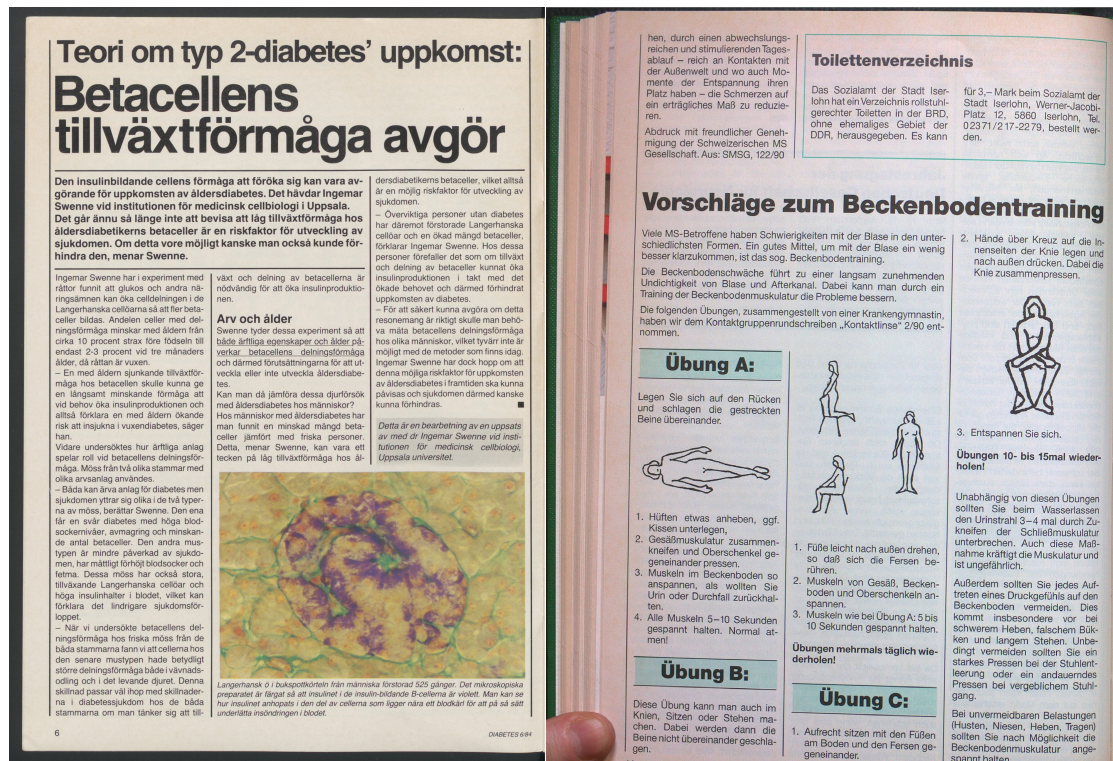


Figure 1: Example pages of the Swedish periodical Diabetes 6:1984 p. 6 and the German periodical Mitteilungsblatt, 149:1990 p. 14.

The periodicals of the Swedish patient organizations were professionally scanned by the University Library of Gothenburg University. The periodicals of the German patient organizations, on the other hand, were scanned with the intention to only be used for close reading. This resulted in a more inconsistent scan quality compared to the Swedish material and presented a challenge for achieving sufficient OCR results. Figure 1 illustrates the difference in scan quality between the Swedish (left) and the German periodicals (right) in terms of lighting, orientation of text lines and paragraphs, and warping of the page. Furthermore, the example pages showcase the complexity of the layout of the periodicals. The layout of the Swedish and German periodicals became gradually more complex over time, with multiple columns, images, illustrations, in-text advertisements, and creative cursive fonts appearing more frequently. Combined with other layout features such as colored backgrounds, fracture script, and text inside images, this added another

layer of complexity to the OCR process.

The quality of the OCR output has a direct impact on the implementation of NLP text mining methods, such as concordance analysis, named entity recognition, and topic modeling. To facilitate subsequent NLP text mining tasks and achieve reliable results, we strived for an OCR quality of 90%. Meaning that of all the words on a page, at least 90% would be correctly recognized [7, 8]. Advanced NLP methods require a good OCR quality not only in terms of correctly recognized words, but also in terms of accurate layout analysis and reading order detection [9]. Due to the complex and inconsistent layout of the periodicals, ensuring the correct reading order for each page was labor intensive and in many cases required judgment calls from the historians. Not only was it impossible to define a set rule for reading order (e.g. left to right, top to bottom), there were also many pages where a “correct” reading order was difficult to determine even for a human reader.

Similarly, making a consistent rule for what would count as a text unit (e.g. an article) and extracting individual texts from the material, similar to the repositories of many scientific journals, proved to be challenging. Texts would regularly be broken up over several nonconsecutive pages or issues, and the distinction between what constitutes, for instance, a text versus a subsection of a text was ambiguous. For that reason, we decided that the page would constitute our base text unit. This decision has the consequence of obscuring the association between words that appear in a multi-page article, and creating an association between words that appear in different texts on the same page. However, this remoteness/closeness between words also represents a material fact that is consistent with the situation when the material was being read by its intended audience. Segmenting the material into individual texts, we concluded, would have created a greater discrepancy between the digital output and the paper original.

ABBYY FineReader Server 14 was used for the OCR processing. For each periodical a separate workflow was designed. Depending on the scan quality, preprocessing steps were included in the periodical’s workflow, such as straightening the text lines or deskewing of the page. In order to ensure a good quality, a confidence character threshold was set. If the low confidence character rate of a page exceeded 5%, the image was manually reviewed in the software’s verification station. In practice, this meant that between 5 to 10% of the pages were sent to the verification station for review. Here text and image areas were added, removed, and/or redrawn, OCR errors corrected, and the reading order checked. The confidence character rate of a page was not always accurate. The software could be confident about interpreting a character correctly while it in reality did not, or the other way around, be insecure about a character while it was accurately recognized. Overall, the performance of the OCR software on the Swedish and German material was good on a character and word level. However, the layout and reading order detection was sometimes incorrect. Given the complex page layout combined with the inconsistent scan quality of the German periodicals, this was not surprising. Combining multiple preprocessing methods in one workflow seemed to negatively impact the accuracy of recognized reading order during our tests. Therefore, we opted to only perform basic preprocessing if it significantly improved the quality of the OCR.

5. OCR post-correction

To achieve an estimate of the OCR quality independent of ABBYY's own estimation — as well as to correct some of the errors still remaining in the corpus — we implemented a word-list based approach for OCR post-correction based on previous work by Thompson et al. [10]. We based the implementation on a spellchecker [11], which we extended with additional functionalities such as compound splitting. With Diabetes as a first test case, we configured the post-correction to only suggest corrections that 1) had an edit distance of one from the original, unknown word, and 2) were more frequent in the OCR text output than the original word. As reference data for the spellchecker and for the OCR quality estimation, we used word-lists gathered from corpora and from medical terminology, which we manually extended with correct words that were flagged as unknown by the spellchecker. Finally, we manually verified the spelling corrections suggested by the spellchecker before replacing the original word as it had been interpreted by the OCR software. Similar to the estimations made by the ABBYY OCR software, our word-list based quality measurements indicate a consistent and low error rate, with only 382 of 8 991 pages exceeding an error rate of 5% unknown words¹.

6. Conclusion

Tailoring source corpora to the scope and questions of a specific project opens up a wider field for using text mining in historical research. It also allows a greater transparency in the research method since the decisions made in the digitization and post-processing stage are consequential for the analysis. In our corpus of 20th century patient organization periodicals, we achieved a high OCR quality on the character and word level which could further be improved with spell checking, but layout parsing presented more challenges. The software's difficulty in determining the correct reading order derives in part from an inherent ambiguity in the layout. Our material is unusually diverse, complex, and fragmented from a text mining standpoint, but typical for historical source materials. Using digital methods more broadly for historical research will require improved methods for handling structurally inconsistent and fragmented materials.

Acknowledgements

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.

¹Our implementation of Thompson et al.'s [10] algorithm for OCR post-correction and quality estimation, as well as instructions for how to practically apply it, will be made available through CDHU as a Huminfra resource. It is available at CDHU's GitHub: <https://github.com/CDHUppsala/Customised-OCR-Correction>.

References

- [1] R. Porter, The Patient's View: Doing Medical History from below, *Theory and Society* 14 (1985) 175–198. URL: <http://www.jstor.org/stable/657089>.
- [2] L. Jordanova, The Social Construction of Medical Knowledge, *Social History of Medicine* 8 (1995) 361–381. URL: <https://academic.oup.com/shm/article-lookup/doi/10.1093/shm/8.3.361>. doi:10.1093/shm/8.3.361.
- [3] F. Condrau, The Patient's View Meets the Clinical Gaze, *Social History of Medicine* 20 (2007) 525–540. URL: <https://doi.org/10.1093/shm/hkm076>. doi:10.1093/shm/hkm076.
- [4] Y. Söderfeldt, The Truth Within: Making Medical Knowledge in the Hay Fever Association of Heligoland, 1899–1909, *Isis* 112 (2021) 531–547. URL: <https://www.journals.uchicago.edu/doi/10.1086/715653>. doi:10.1086/715653.
- [5] M. Fridlund, Digital history 1.5: A middle way between normal and paradigmatic digital historical research, in: M. Fridlund, M. Oiva, P. Paju (Eds.), *Digital Histories: Emergent Approaches within the New Digital History*, Helsinki University Press, Helsinki, 2020, pp. 69–87. doi:10.1145/90417.90738.
- [6] CDHU, *cdhu ocrscripts*, <https://github.com/CDHUppsala/cdhu-ocrscripts>, 2023.
- [7] D. van Strien, K. Beelen, M. Ardanuy, K. Hosseini, B. McGillivray, G. Colavizza, Assessing the Impact of OCR Quality on Downstream NLP Tasks:, in: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, SCITEPRESS - Science and Technology Publications, Valletta, Malta, 2020, pp. 484–496. URL: <http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009169004840496>. doi:10.5220/0009169004840496.
- [8] M. J. Hill, S. Hengchen, Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study, *Digital Scholarship in the Humanities* 34 (2019) 825–843. URL: <https://academic.oup.com/dsh/article/34/4/825/5476122>. doi:10.1093/llc/fqz024.
- [9] C. Neudecker, K. Baierer, M. Gerber, C. Clausner, A. Antonacopoulos, S. Pletschacher, A survey of OCR evaluation tools and metrics, in: *The 6th International Workshop on Historical Document Imaging and Processing*, ACM, Lausanne Switzerland, 2021, pp. 13–18. URL: <https://dl.acm.org/doi/10.1145/3476887.3476888>. doi:10.1145/3476887.3476888.
- [10] P. Thompson, J. McNaught, S. Ananiadou, Customised OCR correction for historical medical text, in: *2015 Digital Heritage*, IEEE, Granada, Spain, 2015, pp. 35–42. URL: <http://ieeexplore.ieee.org/document/7413829/>. doi:10.1109/DigitalHeritage.2015.7413829.
- [11] T. Barrus, *pyspellchecker*, <https://pyspellchecker.readthedocs.io>, 2018.

On two SweLL learner corpora – SweLL-pilot and SweLL-gold

Elena Volodina

University of Gothenburg, Sweden

Abstract

SweLL – **S**wedish **L**earner **L**anguage – is a unifying term for the infrastructure module for research on Swedish as a Second Language (L2), deployed and maintained as a part of bigger infrastructure of Språkbanken Text at the University of Gothenburg, Sweden. The SweLL infrastructure module consists of a number of learner data collections, and tools for annotation and management of learner data. As a result, many of its components contain the prefix *SweLL* in their names, which has created some confusion, especially with regards to the two corpora. In this article we shortly introduce the various SweLL-components with a special focus on the differences between the two SweLL corpora.

Keywords

SweLL, learner corpus research infrastructure, Swedish as a second language, correction annotation aka error annotation, normalization, CEFR labels

1. Introduction

Learner corpora are collections of essays written by learners of some language, where essays are used for empirical evidence in research on the development of learner language or in related fields. Some examples are ASK for L2 Norwegian [1], FALKO for L2 German [2, 3], MERLIN for L2 Czech, German and Italian [4], COPLE2 for L2 Portuguese [5], CzeSL for L2 Czech [6], LAVA for L2 Latvian [7], Icelandic L2 Error Corpus [8] and multiple learner corpora for L2 English [e.g. 9, 10, 11].

For L2 Swedish there exist several collections, such as CrossCheck with essays from different levels of schools/courses [12], ASU with L2 essays and transcribed L2 speech [13], and Uppsala Corpus of Student Writings with an extensive collection of essays from Swedish national exams [14]. These corpora are very valuable, reflecting different aspects of L2 Swedish, but are not easy to gain access to.

The SweLL initiative first emerged in 2012, when several smaller collections were offered to Språkbanken Text for processing and maintenance, namely *TISUS-texts* from 2006 and *SWI203* from 2012. In parallel, compilation of another smaller corpus, *SpIn*, was in progress at Språkbanken Text itself. Being too small to be released as three individual corpora, the three collections were unified and released under the name of a *SweLL-corpus* in 2016 [15]. This first SweLL compilation was used as a starting point for a grant application with the same name, *SweLL - research*

Huminfra Conference 2024, Gothenburg, 10–11 January 2024.

✉ elena.volodina@svenska.gu.se (E. Volodina)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

infrastructure for Swedish as a second language,¹ granted by Riksbankens Jubileumsfond for years 2017–2020. During the *SweLL project* time, a new corpus of L2 Swedish was developed within a highly cross-disciplinary group of collaborators from the University of Gothenburg, Stockholm university, Uppsala university and Umeå university representing fields of Second Language Acquisition, Learner Corpus Research and Natural Language Processing.

In hindsight, it was unfortunate, that the name of the first corpus and the project name were the same, especially since the corpus produced in the *SweLL project* was, naturally, also called *SweLL corpus*. One more term using *SweLL* prefix is the *SweLL portal* [16] - an annotation management tool developed within the *SweLL project*.

In 2021, during preparation of the *SweLL* release v.1, a decision was made to rename the first of the *SweLL*-corpora to *SweLL-pilot* – which largely reflects its nature, being a proof-of-concept and source of wisdom; and to call the project-generated corpus *SweLL-gold* – which reflects its status as a corpus with a higher standard with extensive correction annotation of high quality.

To summarize the chronological order of the use of the term (or, rather, modifier) *SweLL*:

- 2016: *SweLL corpus* [15] → since 2021: *SweLL-pilot corpus*
- 2017: *SweLL project*¹ (finished in 2021)
- 2018: *SweLL portal* [16]
- 2019: *SweLL corpus* [17] → since 2021: *SweLL-gold corpus* (NOTE! incl. *SweLL(-gold) target* and *SweLL(-gold) original* subparts)
- 2021: *SweLL infrastructure (module)*

Below, we shortly introduce the standards for metadata, annotation and file formats in the Learner Corpora Research field (section 2), and zoom into the two *SweLL*-corpora, *SweLL-pilot* and *SweLL-gold* to summarize their similarities and differences (section 3).

2. Ideal infrastructure for learner language

Despite the relatively long history of Learner Corpus Research (LCR), of at least three decades, there is still no agreement about what to consider an ideal standard [18], which reflects the dynamic nature of the fields involved. This refers to sets of metadata; which annotation to include and in which standard; data formats for release; and search tools.

Metadata for learner corpora is extremely important for pursuing different types of research and for the interoperability between corpora [19, 20]. For example, age, gender and first languages are important for identification of learning problems for different demographic groups; task metadata – for studying the impact of the task on the type of language produced by learners in the essays. However, there are many other metadata aspects that are easily overlooked by corpus compilers, although similarly important.

Work on metadata standardization in LCR was initiated by Paquot and Granger in 2017 [21], was followed up by König et al. in 2022 [22] and is still ongoing [23]. Paquot et al. [23] identify eight groups of metadata – administrative, corpus design, learner, text, task, annotation, annotator and transcriber² – with multiple subcategories divided into obligatory and optional. In both

¹<https://spraakbanken.gu.se/en/projects/swell>

²Work-in-progress document available here: <https://tinyurl.com/L2metadataV2>

	SweLL-pilot				SweLL-gold		Total
	Spln	SW1203	TISUS	total	original	normalized	
Year of collection	2012 -2016	2012 -2013	2006	2006 -2016	2017 -2020		2006 -2020
Nr tokens	46 911	52 518	60 632	160 061	147 842	(151 851)	307 903
Nr sentences	4 302	3 145	3 422	10 869	7 807	(8 137)	18 676
Nr essays	256	141	105	502	502	(502)	1 004
Nr A1 essays	59	0	0	59	Beginner: 289		N/A
Nr A2 essays	143	0	0	143			N/A
Nr B1 essays	46	40	0	86	Intermediate: 45		N/A
Nr B2 essays	2	71	32	105			N/A
Nr C1 essays	0	23	73	96	Advanced: 168		N/A
Nr C2 essays	0	7	0	7			N/A
Nr Unknown	6	0	0	6			N/A

Table 1
Overview of SweLL-pilot and SweLL-gold statistics per subcorpus

SweLL corpora, most of the obligatory metadata are considered,⁴⁵⁶⁷ however, the metadata for characterizing annotators and transcribers is not among those, which is difficult to rectify post-factum.

Annotation standards in LCR cover both manual and automatic annotation, stratified further into linguistic annotation, anonymization (vs pseudonymization), normalization, error correction (vs correction annotation), etc. [18]. Included here are also tools for annotation and annotation management. Most previous projects relied on xml schema for annotation where corrections were assigned to the original strings [e.g. 1, 5], which has recently started to be replaced by alternative approaches, such as viewing original and corrected versions of essays as independent aligned versions of a parallel corpus [e.g. 6, 7]. Unlike most predecessors, the *SweLL-gold* corpus has been *pseudonymized* (not anonymized) – i.e. personal information in texts has been substituted by alternative strings to preserve the integrity of learner texts and to conform to the requirements of the GDPR [24]; *normalized*, i.e. rewritten to an alternative independent corrected version, and corrections were *correct-annotated*³ for the nature of the difference between the original and normalized strings. All in all, the *SweLL project* has contributed (1) to increased attention to the need for structured pseudonymization of learner essays [25, 26]; (2) to an emerging new paradigm of learner corpora where the original and normalized versions are treated as parallel corpora [27]; and (3) to shifting the focus from 'errors' in learner versions to their 'corrections' since these corrections are only some of several possible hypothetical ways to interpret (errors in) learner writing [28].

The automatic linguistic annotation present in both SweLL corpora comes from Sparv annotation pipeline [29] and contains tokenization, lemmatization, word sense disambiguation, morpho-syntactic annotation, syntactic dependencies and a few others.

Formats are largely influenced by the way annotation is conceptualized, such as whether to treat the corrected version as an independent text, or to attach a corrected string directly into the original sentence. However, even the search interfaces set limitations on formats, most

³In majority of other learner corpora this is called 'error annotation'

prominently, corpus workbench depending heavily on TEI-XML. Most error-annotated corpora are, therefore, distributed in xml file formats with only a few distributed alternatively also in json format [30]. *SweLL-gold* and *SweLL-pilot* are distributed in three file formats: raw texts, linguistically annotated xml (TEI-XML) and json (in case of *SweLL-gold* containing correction and pseudonymization tags).

Search tools are critical for accessing and analyzing learner data, with multiple solutions, often adapted to a corpus in question. For both *SweLL* corpora, it was possible to use Korp [31], where the user can see each subcorpus individually under 'L2 Korp' – SpIn, SW1203, TISUS, *SweLL-origial*, *SweLL-target* (Table 1) and perform searches in any combination of those.

3. The two *SweLL* corpora

The *SweLL* (Swedish Learner Language) *infrastructure* currently contains two *SweLL* corpora (which are shown as five subcorpora in Korp search tool), collected at two different periods of time: *SweLL-pilot* between 2006–2016 [15] and *SweLL-gold* between 2017–2020 [17]. As the name suggests, *SweLL-pilot* was the initial attempt to collect learner essays; whereas *SweLL-gold* is built upon those experiences, accounting for the lessons learnt, correcting the limitations and extending the scope of annotation. Notably, during the *SweLL-gold* period a larger group of researchers and annotators was involved and richer annotation schemes and tools were developed. Table 1 provides an overview of statistics over the two essay collections.

3.1. *SweLL-pilot* - a corpus of learner essays with CEFR labels

SweLL-pilot is a corpus of essays written by adult learners of Swedish during exam settings and collected from students who have signed consents. It was collected during the period of 2006-2016, with the first release of 339 essays in 2016 [15] – transcribed from hand-written essays and anonymized. In 2018, 163 more essays were transcribed, anonymized and added to the *SweLL-pilot* collection. In 2020-2021 the *SweLL-pilot* collection was added to the *SweLL portal* [16] to ensure comparable json format [27] and harmonized metadata attributes with the *SweLL-gold* collection. Nowadays, *SweLL-pilot* contains 502 essays that have been anonymized and labeled with the CEFR levels.

The *SweLL-pilot* collection contains three subcorpora, all of which represent multiple first languages (L1) and age groups:

- SpIn⁴ - 256 essays collected from Language Introduction course (mid-term exams) for newly arrived refugees. Some of the students are recurrent.
- SW1203⁵ - 141 essays collected from university students in exam setting, most of who wrote three essays each.
- TISUS⁶ - 105 essays written as a part of a Test In Swedish for University Studies. All essays are on the same topic "Stress" and within the argumentative genre.

⁴SpIn metadata: <https://spraakbanken.github.io/swell-release-v1/Metadata-SpIn>

⁵SW1203 metadata: <https://spraakbanken.github.io/swell-release-v1/Metadata-SW1203>

⁶TISUS metadata: <https://spraakbanken.github.io/swell-release-v1/Metadata-TISUS>

	Metadata	Privacy	Normalized	Correct-annotated	CEFR labeled
SweLL-pilot	Harmonized	Anonymized	no	no	yes
SweLL-gold	Harmonized	Pseudonymized	yes	yes	no

Table 2

Major differences between the annotations present in the two SweLL subcorpora

SweLL-pilot is the first and the only CEFR-labeled learner corpus of L2 Swedish. The Inter-annotator agreement on CEFR labeling measured for the SW1203 subcorpus (141 essays) is 0.80% Krippendorff’s alpha [32] which corresponds to high annotation quality.

3.2. SweLL-gold - a corpus of learner essays with error annotation

The *SweLL-gold* corpus⁷ [17] was developed within the *SweLL project*¹ [33], the purpose of which was to set up an infrastructure for collection, digitization, normalization, and annotation of L2 Swedish adult learner written production, as well as to make available a linguistically annotated *parallel* corpus, where it would be possible to search for various types of linguistic structures, without the researcher having to guess what such a structure might look like in original essays, since there is a parallel normalized version available.

The essays were collected from several schools around Sweden where teachers assisted with consent forms, personal and task metadata forms, and essays. The type of school was used as an indication of the approximate level of learners, e.g. *upper-secondary* and *university preparatory* courses being representative of ‘Advanced’ levels (C); *SVA* (Swedish as a Second Language) courses for adults representing ‘Intermediate’ levels (B); and *SFI* courses (Swedish For Immigrants) representing ‘Beginner’ levels (A). The original essays were transcribed and normalized (i.e. rewritten in standard Swedish that conforms to grammatical norms); and all corrections were labeled as to their nature, i.e. correct-annotated (in other corpora called error-annotated). The result was an aligned parallel corpus of original and normalized essays with correction labels attached to the aligned segments of the essay.

The *SweLL infrastructure* components, such as *SweLL portal* [16], *SVALA* annotation tool [27] and multiple guidelines for annotation [34, 35, 36, 37] were developed to ensure high quality of data annotation, which resulted in Inter-Annotator Agreement of 88% by Fleiss’ kappa and 76% by Krippendorff’s alpha [38, 32] as measured on 10% of the essays (i.e. 50 essays).

SweLL-gold is the first and the only correction-annotated L2 learner corpus of Swedish.

3.3. Differences between SweLL-pilot and SweLL-gold

The general overview of the statistics for the two corpora, provided in Table 1, shows that the size of the corpora is comparable, albeit relatively modest, amounting in total to 1004 essays representing 307 903 tokens. We can also see from Table 2 that the metadata and attribute names have been harmonized between the two corpora. However, there are four critical aspects that differ between the corpora, namely, (1) the way personal data in the text was handled (‘Privacy’);

⁷SweLL-gold metadata: <https://spraakbanken.github.io/swell-release-v1/Metadata-SweLL>

(2) absence or presence of a corrected version of the original essay ('Normalized'); (3) absence or presence of the manually assigned labels for corrections ('Correct-annotated'); and (4) absence or presence of CEFR labels ('CEFR labeled').

This means that the two corpora can practically never be used for the same research questions or applied to the same development problems. For example, if you are interested in Grammatical Error Detection (GED) or Correction (GEC), only SweLL-gold is appropriate. If you want to develop an automatic essay grading (AEG) system or identify a scope of vocabulary/grammar used at particular level, only SweLL-pilot can be used.

In an ideal world, each of the corpora would be complemented for the missing annotation. However, in the real world it demands additional funding to make it happen. On the bright side, *SweLL-pilot* is currently being pseudonymized in accordance with the standards of the *SweLL-gold* corpus [35, 39] so that it can be used for work on automatic pseudonymization of research data within the 'Grandma Karl' project⁸ [26]. That version of *SweLL-pilot* will be released in the future. Normalization and correction annotation of *SweLL-pilot*, as well as CEFR-labeling of *SweLL-gold* are, however, left for future.

3.4. Access to the data

The two *SweLL corpora* contain private information - both in the form of metadata and as private mentions in texts, and are therefore under the GDPR [24] protection. This sets limitations to the openness of data, namely, that only individuals living and working in Europe can have access to the data; with a further restriction that the area of application should be connected to education (teaching, learning, research or development).

Due to that, access to the *SweLL corpora* is administered through an application form.⁹ The approved user gets access to the data in three file formats: raw text, linguistically annotated xml and json; as well as through a corpus search system Korp¹⁰ [31].

4. SweLL impact: a game changer in Swedish L2?

It is a fact that languages and research domains, that can boast rich data collections, have more empirical and data-intensive research done on them [40, 41]. This makes us believe that now, with the *SweLL data* available for research and development, the field of Swedish as a second language and related research fields will get a boost. Since the release of *SweLL-pilot* in 2016, we can see a steady increase in interest to

- (1) development of automatic tools and approaches, such as classification of essays, lexical complexity prediction, error detection and correction [42, 43, 44, 45, 46, 47, 48, 49, 50, 51];
- (2) data-driven linguistic studies on vocabulary and grammar scopes in second language learning, grammatical patterns at different levels of linguistic development, etc. [52, 46, 53, 54, 55, 56];
- (3) novel approaches to feedback generation [57, 50];

⁸<https://mormor-karl.github.io/>

⁹<https://sunet.artologik.net/gu/swell>

¹⁰<https://spraakbanken.gu.se/korp/>

(4) methodological studies, such as, pseudonymization of research data, effects of errors on the performance of automatic tools, fairness and bias in language assessment [58, 59, 26], etc.

A number of derivative resources have been developed since 2016 based on the two SweLL corpora, such as wordlists for language learners – SweLLex [52] and later Sen*Lex [45] – for studies on lexical competences of L2 learners; DaLAJ [60] for studies on linguistic acceptability [61], CoDeRooMor [62] for studying derivational morphology of Swedish, MuClaGED [63] for error classification, synthetic datasets imitating real-life errors [64] and many others. The Swedish MultiGED dataset¹¹ based on *SweLL-gold* has been used for the MultiGED shared task [48] and we plan new shared tasks based on the *SweLL corpora* in the near future.

5. Future directions

We are expecting both short-term and long-term impact from the two corpora described in this article on the fields of Swedish as a Second language, Learner Corpus Research (nationally and internationally), and NLP- and AI-based approaches to L2 Swedish.

First of all, we intend to **promote** the use of the datasets among NLP researchers through organization of multilingual *shared tasks*.¹²

Second, we will work towards **extending** *authentic* learner datasets through setting on-the-fly pseudonymization algorithms for *continuous collection of essays* directly from schools.

In parallel, we will also work on generation of *synthetic* datasets with basis in the current SweLL data, for example experimenting with GPT models to generate mock learner essays at different levels of proficiency, using real-life essays as samples, or generating error datasets using linguistic patterns observed in the SweLL-gold data.

Finally, we will search for possibilities to *harmonize the two SweLL corpora* (and potential other subcorpora that will be added to the SweLL infrastructure module) between each other through normalization and correction annotation of SweLL-pilot and CEFR-labeling of SweLL-gold. We do not exclude that these steps will be performed automatically (with subsequent manual proofreading) after we have experimented with the *automatic approaches to normalization, correction annotation and CEFR labeling*.

Acknowledgments

Work on the article has been supported by *Nationella språkbanken* and *Huminfra*, both funded by the Swedish Research Council (2018-2024, contract 2017-00626; 2022-2024, contract 2021-00176) and their participating partner institutions. Work on SweLL-pilot was supported by the *Center for Language Technology* at the University of Gothenburg (SpIn and SW1203), and by *Ebba Danelius stiftelse för sociala och kulturella ändamål* and *Granholms stiftelse*, at Stockholm University (TISUS). Work on SweLL-gold has been supported by the infrastructure grant from the Swedish Riksbankens Jubileumsfond *SweLL – research infrastructure for Swedish as a second language*, grant IN16-0464:1.

¹¹<https://github.com/spraakbanken/multiged-2023>

¹²<https://spraakbanken.gu.se/en/compsla>

References

- [1] K. Tenfjord, P. Meurer, K. Hofland, The ASK corpus: A language learner corpus of Norwegian as a second language, in: LREC'06, 2006, pp. 1821–1824.
- [2] A. Lüdeling, M. Walter, E. Kroymann, P. Adolphs, Multi-level error annotation in learner corpora, in: Proceedings of corpus linguistics, volume 1, Citeseer, 2005, pp. 14–17.
- [3] M. Reznicek, A. Lüdeling, C. Krummes, F. Schwantuschke, Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.0, Humboldt-Universität zu Berlin, Berlin, Germany, 2012.
- [4] A. Boyd, J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schöne, B. Štindlová, C. Vettori, The MERLIN corpus: Learner Language and the CEFR, in: LREC'14, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014.
- [5] A. Mendes, S. Antunes, M. Janssen, A. Gonçalves, The COPLE2 corpus: a learner corpus for Portuguese., in: LREC'16, 2016.
- [6] A. Rosen, J. Hana, B. Vidová Hladká, T. Jelínek, S. Škodová, B. Štindlová, Compiling and annotating a learner corpus for a morphologically rich language: CzeSL, a corpus of non-native Czech, Nakladatelství Karolinum, 2020.
- [7] R. Darģis, I. Auzina, K. Levāne-Petrova, I. Kaija, Detailed Error Annotation for Morphologically Rich Languages: Latvian Use Case, in: Human Language Technologies–The Baltic Perspective, IOS Press, 2020, pp. 241–244.
- [8] I. Glisic, A. K. Ingason, The nature of Icelandic as a second language: An insight from the learner error corpus for Icelandic, in: CLARIN Annual Conference, 2022, pp. 23–33.
- [9] H. Yannakoudakis, T. Briscoe, B. Medlock, A new dataset and method for automatically grading ESOL texts, in: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011, pp. 180–189.
- [10] M. Paquot, Corpora and Second Language Acquisition, in: The Routledge Handbook of Corpora and English Language Teaching and Learning, Routledge, 2022, pp. 26–40.
- [11] O. Vinogradova, O. Lyashevskaya, Review of Practices of Collecting and Annotating Texts in the Learner Corpus REALEC, in: International Conference on Text, Speech, and Dialogue, Springer, 2022, pp. 77–88.
- [12] J. Lindberg, G. Eriksson, CrossCheck-korpusen - en elektronisk L2-korpus för skriven svenska., in: B. de Geer A. Malmberg (Red.), Språk på tvärs Rapport från ASLA:s höstsymposium Södertörn, 11-12 november 2004, Svenska föreningen för tillämpad språkvetenskap. Uppsala 2005, 2004, pp. 89–98.
- [13] B. Hammarberg, Introduktion till ASU-korpusen: En longitudinell muntlig och skriftlig textkorpus av vuxna inlärares svenska med en motsvarande del från infödda svenskar., 2010.
- [14] B. Megyesi, J. Näsman, A. Palmér, The Uppsala corpus of student writings: Corpus creation, annotation, and analysis, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 3192–3199.
- [15] E. Volodina, I. Pilán, I. Enström, L. Llozhi, P. Lundkvist, G. Sundberg, M. Sandell, SweLL on the rise: Swedish learner language corpus for European reference level studies, in: Proceedings of the 10th Language Resources and Evaluation Conference (LREC), Portorož, Slovenia, 2016.
- [16] Y. A. Mohammed, A. Matsson, E. Volodina, Annotation Management Tool: A Requirement

- for Corpus Construction, in: CLARIN Annual Conference, 2022, pp. 101–108.
- [17] E. Volodina, L. Granstedt, A. Matsson, B. Megyesi, I. Pilán, J. Prentice, D. Rosén, L. Rudebeck, C.-J. Schenström, G. Sundberg, et al., The SweLL language learner corpus: From design to annotation, *Northern European Journal of Language Technology (NEJLT)* 6 (2019) 67–104.
- [18] E. W. Stemle, A. Boyd, M. Jansen, T. Lindström Tiedemann, N. Mikelić Preradović, A. Rosen, D. Rosén, E. Volodina, Working together towards an ideal infrastructure for language learner corpora, *Widening the Scope of Learner Corpus Research* (2019).
- [19] E. Volodina, M. Janssen, T. L. Tiedemann, N. M. Preradovic, S. K. Ragnhildstveit, K. Tenfjord, K. de Smedt, Interoperability of second language resources and tools, in: *Proceedings of the CLARIN Annual Conference, 2018*, pp. 90–94.
- [20] A. König, J.-C. Frey, E. W. Stemle, Exploring reusability and reproducibility for a research infrastructure for I1 and I2 learner corpora, *Information* 12 (2021) 199.
- [21] S. Granger, M. Paquot, Core Metadata [Schema] for Learner Corpora Draft 1.0, 2017.
- [22] A. König, J.-C. Frey, E. W. Stemle, A. Glaznieks, M. Paquot, Towards standardizing LCR metadata, in: *Book of Abstracts from the Learner Corpus Research Conference, Italy, 2022*.
- [23] M. Paquot, A. König, E. W. Stemle, J.-C. Frey, A core metadata schema for L2 data, in: *Book of Abstracts from the EuroSLA Conference 2023, 2023*.
- [24] E. EU Commission, General Data Protection Regulation., *Official Journal of the European Union*, 59, 1-88., 2016. URL: <https://gdpr-info.eu/>(Accessed2019-11-19).
- [25] E. Volodina, Y. A. Mohammed, S. Derbring, A. Matsson, B. Megyesi, Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays, in: *Proceedings of the 28th International Conference on Computational Linguistics, 2020*, pp. 357–369.
- [26] E. Volodina, S. Dobnik, T. L. m Tiedemann, X.-S. Vu, Grandma Karl is 27 years old—research agenda for pseudonymization of research data, in: *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService), IEEE, 2023*, pp. 229–233.
- [27] M. Wirén, A. Matsson, D. Rosén, E. Volodina, Svala: Annotation of second-language learner text based on mostly automatic alignment of parallel corpora, in: *CLARIN Annual Conference, Pisa, Italy, 8-10 October, 2018*, Linköping University Electronic Press, 2019, pp. 222–234.
- [28] L. Rudebeck, G. Sundberg, On the other side of the error tag: The nature and functions of the corrected texts, in: *Book of Abstracts from the Learner Corpus Research Conference 2022, Italy, 2022*, p. 103.
- [29] M. Hammarstedt, A. Schumacher, L. Borin, M. Forsberg, Sparv 5 User Manual, Technical Report, Göteborg, 2022.
- [30] Š. Arhar Holdt, I. Kosem, Šolar, the developmental corpus of Slovene (2023).
- [31] L. Borin, M. Forsberg, J. Roxendal, Korp – the corpus infrastructure of Språkbanken, in: *Proceedings of LREC 2012. Istanbul: ELRA, volume Accepted, 2012*, p. 474–478.
- [32] K. Krippendorff, *Content analysis: An introduction to its methodology*, Sage publications, 2018.
- [33] E. Volodina, B. Megyesi, M. Wirén, L. Granstedt, J. Prentice, M. Reichenberg, G. Sundberg, A friend in need?: Research agenda for electronic Second Language infrastructure, in:

- Swedish Language Technology Conference (SLTC) 2016, 2016.
- [34] E. Volodina, B. Megyesi, SweLL transcription guidelines, L2 essays (2021). URL: <https://gupea.ub.gu.se/handle/2077/69429>.
- [35] B. Megyesi, L. Rudebeck, E. Volodina, SweLL pseudonymization guidelines, 2021. URL: <http://hdl.handle.net/2077/69431>.
- [36] L. Rudebeck, G. Sundberg, SweLL correction annotation guidelines (2021). URL: <https://gupea.ub.gu.se/handle/2077/69434>.
- [37] L. Rudebeck, G. Sundberg, M. Wirén, SweLL normalization guidelines (2021). URL: <https://gupea.ub.gu.se/handle/2077/69432>.
- [38] R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics, *Computational linguistics* 34 (2008) 555–596.
- [39] B. Megyesi, L. Granstedt, S. Johansson, J. Prentice, D. Rosén, C.-J. Schenström, G. Sundberg, M. Wirén, E. Volodina, Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish, in: *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, LiU Electronic Press, Stockholm, Sweden, 2018, pp. 47–56. URL: <https://aclanthology.org/W18-7106>.
- [40] M. Perc, The Matthew effect in empirical data, *Journal of The Royal Society Interface* 11 (2014) 20140378. doi:<https://doi.org/10.1098/rsif.2014.0378>.
- [41] A. Søgaard, Should We Ban English NLP for a Year?, in: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 5254–5260.
- [42] I. Pilán, E. Volodina, T. Zesch, Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks, in: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 2101–2111.
- [43] I. Pilán, E. Volodina, Investigating the importance of linguistic complexity features across different datasets related to language learning, in: *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, 2018, pp. 49–58.
- [44] D. Alfter, E. Volodina, Towards single word lexical complexity prediction, in: *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, 2018, pp. 79–88.
- [45] D. Alfter, Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective, *Data Linguistica* 31, University of Gothenburg, 2021.
- [46] D. Alfter, T. L. Tiedemann, E. Volodina, Crowdsourcing relative rankings of multi-word expressions: Experts versus non-experts, *North European Journal of Language Technology* Vol. 7 (2022).
- [47] M. Nyberg, Grammatical error correction for learners of Swedish as a second language, 2022.
- [48] E. Volodina, C. Bryant, A. Caines, O. De Clercq, J.-C. Frey, E. Ershova, A. Rosen, O. Vinogradova, MultiGED-2023 shared task at NLP4CALL: Multilingual Grammatical Error Detection, in: *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, 2023, pp. 1–16.
- [49] R. Östling, K. Gillholm, M. Kurfalı, M. Mattson, M. Wirén, Evaluation of really good grammatical error correction, *arXiv preprint arXiv:2308.08982* (2023).

- [50] A. Masciolini, E. Volodina, D. Dannlls, Towards automatically extracting morphosyntactical error patterns from L1-L2 parallel dependency treebanks, in: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), 2023, pp. 585–597.
- [51] J. Ehnroth, Y. Park, Correction of Grammatical Errors in Swedish (2023).
- [52] E. Volodina, I. Pilán, L. Llozhi, B. Degryse, T. François, SweLLex: second language learners’ productive vocabulary, in: Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition, 2016, pp. 76–84.
- [53] E. Volodina, D. Alfter, T. L. Tiedemann, Crowdsourcing ratings for single lexical items: a core vocabulary perspective, *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave* 10 (2022) 5–61.
- [54] T. Lindström Tiedemann, D. Alfter, Y. Ali Mohammed, D. Piipponen, B. Silén, E. Volodina, Multi-word Expressions in Swedish as a second language – typology, annotation and initial results Submitted (2024).
- [55] G. Sundberg, J. Prentice, SweLL: En svensk inlärarkorpus, *ASLA:s skriftserie/ASLA Studies in Applied Linguistics* 30 (2023) 428–453. doi:<https://doi.org/10.17045/sthlmuni.24321526428>.
- [56] E. Volodina, Y. Ali Mohammed, T. Lindström Tiedemann, Swedish Word Family – Construction, Applicability, Strengths and first Experiments (????).
- [57] A. Masciolini, A query engine for L1-L2 parallel dependency treebanks, in: Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), University of Tartu Library, Tórshavn, Faroe Islands, 2023, pp. 574–587. URL: <https://aclanthology.org/2023.nodalida-1.57>.
- [58] E. Volodina, Y. Ali Mohammed, S. Derbring, A. Matsson, B. Megyesi, Towards Privacy by Design in Learner Corpora Research: A Case of On-the-fly Pseudonymization of Swedish Learner Essays, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 357–369. URL: <https://aclanthology.org/2020.coling-main.32>. doi:10.18653/v1/2020.coling-main.32.
- [59] E. Volodina, D. Alfter, T. Lindström Tiedemann, M. S. Lauriala, D. H. Piipponen, Reliability of automatic linguistic annotation: native vs non-native texts, in: Selected papers from the CLARINAnnual Conference 2021, Linköping University Electronic Press (LiU E-Press), 2022.
- [60] E. Volodina, Y. A. Mohammed, A. Berdičevskis, G. Bouma, J. Öhman, DaLAJ-GED-a dataset for Grammatical Error Detection tasks on Swedish, in: Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning, 2023, pp. 94–101.
- [61] J. Klezl, Y. A. Mohammed, E. Volodina, Exploring Linguistic Acceptability in Swedish Learners’ Language, in: Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning, 2022, pp. 84–94.
- [62] E. Volodina, Y. A. Mohammed, T. L. Tiedemann, CoDeRoomor: A new dataset for non-inflectional morphology studies of Swedish, in: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), 2021, pp. 178–189.
- [63] J. C. Moner, E. Volodina, Swedish MuClAGED: A new dataset for Grammatical Error Detection in Swedish, in: Proceedings of the 11th Workshop on NLP for Computer Assisted

- Language Learning, 2022, pp. 36–45.
- [64] J. C. Moner, E. Volodina, Generation of Synthetic Error Data of Verb Order Errors for Swedish, in: Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), 2022, pp. 33–38.

STUnD: ett Sökverktyg för Tvåspråkiga Universal Dependencies-trädbanker

Arianna Masciolini^{1,†}, Márton A. Tóth^{1,†}

¹*Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs Universitet, Sverige*

Abstract

Föreliggande artikel introducerar STUND, ett Sökverktyg för Tvåspråkiga Universal Dependencies-trädbanker som möjliggör parallella syntaktiska sökningar. Vi demonstrerar dess praktiska tillämpning i en fallstudie på tempusformen presens perfekt i svenska och engelska. Resultaten visar att presens perfekt används i ungefär lika stor utsträckning i båda språken, men att det förekommer viss variation som verkar bero på språkspecifika konventioner och översättningsstrategier.

Keywords

sökverktyg, Universal Dependencies, parallellkorpusar, komparativ lingvistik, tempus-aspekt

1. Introduktion

I denna artikel presenterar vi STUND (Sökverktyg för Tvåspråkig Universal Dependencies), ett nytt sökverktyg för tvåspråkiga parallella korpusar som riktar sig till studier inom komparativ lingvistik.¹ Till skillnad från de flesta andra korpussökverktyg kan man i STUND söka på två språk parallellt. I detta syfte utnyttjar STUND Universal Dependencies (UD) [1], en grammatikformalism för tvärspråkligt konsekvent morfosyntaktisk trädbanksannotering. Frågorna uttrycks i ett språkoberoende frågespråk som är särskilt lämpligt för att beskriva syntaktiska mönster. Detta språk tillåter s.k. *tvåspråkiga sökningar*, där användaren anger två olika mönster som ska matchas av två motsvarande segment i de två språken.

I det följande beskriver vi den nuvarande STUND-prototypen och dess användning (avsnitt 2). Dessutom demonstrerar vi STUnD:s praktiska tillämpning i en fallstudie på tempusformen presens perfekt i svenska och engelska (avsnitt 3). Vi avslutar med en diskussion om våra planer för STUnD:s fortsatta utveckling och några förslag på hur verktyget kan användas i framtida lingvistiska studier (avsnitt 4).

2. Verktyget STUND

2.1. Inputformat

Såsom nämns i introduktionen är STUND ett sökverktyg för tvåspråkiga parallella UD-annoterade korpusar. Dessa korpusar består av två *meningslänkade CoNLL-U filer* som var och en innehåller

Huminfra Conference 2024, Gothenburg, 10–11 January 2024.

[†]These authors contributed equally.

✉ arianna.masciolini@gu.se (A. Masciolini); marion.toth@gu.se (M. A. Tóth)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹Källkod och körbara filer för Linux och Windows kan laddas ner på github.com/harisont/STUnD.

en samling morfosyntaktiskt annoterade meningar på käll- respektive målspråket. CoNLL-U är ett språkoberoende och maskinläsbart klartextformat för dependensträdbanker som ger information om syntaktiska relationer mellan olika ord samt varje ords lemma, ordklasstagg och morfologiska egenskaper.² För många språk kan denna typ av annotering utföras automatiskt med rimlig tillförlitlighet genom lättanvända UD-parsers, såsom UDPipe [2].³ Meningarna i en CoNLL-U fil kan dessutom visualiseras som dependensträd.⁴

En *parallell* UD-trädbank består av två eller flera CoNLL-U filer, länkade på meningsnivå. När det handlar om tvåspråkiga material innebär detta att varje mening i källspråksfilen (hädanefter L1) motsvaras av exakt en mening i målspråksfilen (hädanefter L2).

2.2. Användargränssnitt och frågespråk

The screenshot shows a web interface for searching through CoNLL-U files. At the top, there are file paths for English and Swedish corpora. A search query is entered: `TREE_ (FEATS_ "VerbForm={Part->Sup}") [AND [LEMMA "{have->ha}"], FEATS_ "Tense=Pres"]]`. Below the search bar, there are radio buttons for 'text' (selected) and 'CoNLL-U', and a 'search' button. A 'save as TSV' link is also present. The search results are displayed in a table with two columns: English and Swedish. Each row shows a sentence with the verb 'have' highlighted in bold in both languages. The highlighted segments are: 'has been rising', 'har ökat', 've requested', 'har bett', 'have suggested', 'har föreslagit', 'has been certified', and 'har certifierats'.

Figur 1: En tvåspråkig fråga körs på en engelsk-svensk trädbank i textläge. Resultatet är en lista med parallella meningar där samma presens perfekt-konstruktion används både på engelska och på svenska. De frågematchande segmenten markeras med fet stil.

Den nuvarande STUND-prototypen består av en skrivbordsapplikation som körs i webbläsaren. När programmet startas anger användaren sökvägarna till två parallella CoNLL-U-filer. Därefter kan programmet utföra *enkla* eller *tvåspråkiga* sökningar.

Korpusfrågor uttrycks i ett mönstermatchningspråk som ursprungligen skapades i samband med GF-UD verktygslådan för dependensträdbanker[3, 4]⁵ och utökades för parallella L1-L2-frågor enligt beskrivningen i [5]. En uttömmande beskrivning av frågespråket ligger utanför ramen för denna artikel, men kortfattat består en enkel fråga av en av tre möjliga typer

²Annoteringsstandarden beskrivs utförligt på UD:s officiella webbsida, universaldependencies.org.

³UDPipe har ett användarvänligt webbgränssnitt som finns tillgängligt på lindat.mff.cuni.cz/services/udpipe.

⁴En webbaserad CoNLL-U-viewer finns tillgänglig på universaldependencies.org/conllu_viewer.html.

⁵Själva namnet GF-UD syftar på det att en av huvudkomponenterna i systemet är ett konverteringsverktyg mellan UD och grammatikformalismen Grammatical Framework.

av mönster:⁶

- *enkla mönster*, såsom POS "VERB", matchar (del)träd vars syntaktiska huvud är ett verb;
- *sekvensmönster* matchar sammanhängande token- eller delträdssekvenser. SEQUENCE [DEPREL "nsubj", POS "VERB", DEPREL "obj"], till exempel, är ett typiskt mönster för SVO (Subjekt-Verb-Objekt) språk;
- *trädmönster* beskriver dependensträdsstrukturer. Detta är ovanligt bland korpussöksystem, men särskilt användbart vid eftersökning av ordföljdsberoende syntaktiska mönster. TREE (POS "VERB") [DEPREL "advmod"] matchas exempelvis av alla satser vars huvudverb modifieras av ett adverb, oavsett dess placering.

En tvåspråkig (L1-L2) fråga består av två enkla frågor som ska matchas parallellt av ett L1 segment och dess L2-motpart. Detta används främst vid sökningar på översättningsdivergenser. Tvåspråkiga frågor liknar enkla frågor förutom att användaren kan genom den så kallade "{X→Y}"-syntaxen specificera att en del av frågan ska matchas i käll- respektive målspråket. Vill man söka efter meningar där svenskans enkelt presens motsvaras av engelskans present progressive (*am/are/is* + *-ing*-formen) kan man t.ex. använda söksträngen AND [FEATS_ "Tense=Pres", FEATS_ "VerbForm={Fin→Part}"]. Den här frågan matchar alla meningar där de två texterna uppvisar två syntaktiskt motsvarande verbformer i presens, där den ena är en finit form (t.ex. *sök*) och den andra ett particip (t.ex. *searching*).

Användaren kan välja att få sökresultaten i två olika format. I *textläge* (se figur 1) visar programmet fullständiga meningar, där de frågematchande segmenten markeras i fet stil. Resultaten kan sparas som TSV (Tab-Separated Values) och importerar i ett kalkylbladsprogram för vidare analys. I *CoNNL-U-läge* isoleras de relevanta delträden så att de kan sparas till två parallella CoNNL-U-filer. Dessa filer kan importerar i ett externt verktyg för t.ex. visualisering eller användas som inmatningsfiler för att utföra ytterligare sökningar i STUND. För att förenkla eller förfinas sökresultaten kan användaren även ange ett *ersättningsmönster*, som modifierar alla matchande träd på båda språken.⁷

2.3. Implementering

STUND baseras på L2-UD [5, 6], ett verktyg för studier i andraspråksinlärning. Dess ursprungliga mål är att jämföra ogrammatiska meningar med deras korrigeringar. L2-UD kombinerar CONCEPT-ALIGNMENT [7], som är ett programvarupaket för syntaxbaserad ord- och fraslänkning, med GF-UD:s mönstermatchingsspråk.

STUND är ett grafiskt användargränssnitt för en lätt anpassad version av L2-UD, optimerad för flerspråkiga trädbanker snarare än elevtexter. Den största skillnaden mellan de två versionerna är att de använder olika länkingsregler. I STUND baseras länkningen exklusivt på syntaktisk analys. L2-UD tar däremot hänsyn även till lemman, vilket är irrelevant när man undersöker tvåspråkiga material.

⁶Den fullständiga specifikationen av GF-UD:s frågespråk och dess utökning finns tillgängliga på github.com/GrammaticalFramework/gf-ud/blob/master/doc/patterns.md respektive github.com/harisont/L2-UD.

⁷Ersättningspråket är dokumenterat på github.com/GrammaticalFramework/gf-ud/blob/master/doc/patterns.md.

Frågematchning fungerar på följande sätt:

1. Länkingsmodulen inför fras- och ordlänkning. Närmare sagt försöker CONCEPT-ALIGNMENT att hitta motsvarigheter mellan delträd, vilket leder till grovkornig länkning när de två språken skiljer sig mycket från varandra, och mer finfördelad länkning (upp till ordnivå) när de två språken använder liknande konstruktioner;
2. Den användarspecificerade söksträngen transformeras till två separata frågor: en L1-fråga och en L2-fråga;⁸
3. De två resulterande frågorna matchas mot de L1-L2 korrespondenser som hittades i steg 1. Om länkningen misslyckas för ett visst meningspar matchas enspråkiga frågor endast mot L2-meningen. Detta säkerställer att inga sökträffar går förlorade, så länge UD-annoteringen är korrekt;⁹
4. Om ett ersättningsmönster har angetts, tillämpas det på alla frågematchande delträd.

3. Fallstudie

I följande visar vi STUND-verktygets tillämpning i en fallstudie. Fallstudien undersöker vilka tempus-aspektformer¹⁰ som används i engelskan där svenskan använder tempusformen presens perfekt, och tvärtom: vilka tempusformer som används i svenskan där engelskans present perfect används. Vi kommer av enkelhetsskäl att använda förkortningen *PresPf* för presens perfekt i både svenskan och engelskan. Vidare använder vi i huvudsak engelska benämningar på engelskans tempus-aspektformer för tydlighetens skull.

Det övergripande syftet med fallstudien är att visa hur verktyget kan användas för att göra iakttagelser om hur grammatiska strukturer fungerar i språk. Mer specifikt visar vi hur man kan undersöka hur språks tempus-aspektsystem fungerar. Olika språk har olika grammatiska former för att uttrycka tid, och genom att undersöka distributionen mellan dessa former tvärspråkligt kan vi dra slutsatser om likheter och skillnader mellan språkens tempus-aspektsystem. *PresPf* är intressant i sammanhanget eftersom det visar på stor tvärspråklig variation [8]. Exempelvis är det i svenskan inte ovanligt att använda *PresPf* i en fråga som *Har du sovit gott?* där det i engelskan skulle vara vanligare att använda preteritum *Did you sleep well?* i samma situation [9]. Därmed uppstår frågan till vilken utsträckning användningen av *PresPf* stämmer överens mellan dessa två språk. Genom att titta på svenskans och engelskans *PresPf* visar vi hur man kan undersöka distributionen hos en tvärspråkligt sett likartad kategori i olika språk.

3.1. Svenskans och engelskans tempus-aspektsystem

Svenskan och engelskan har en snarlik indelning i tempusformer:¹¹ enkelt presens (sv. *äter*, eng. *eats*), preteritum (sv. *ät*, eng. *ate*), presens perfekt (sv. *har ätit*, eng. *has eaten*), pluskvamperfekt

⁸Om frågan är enspråkig används wildcardt `TRUE` som L1-fråga och den fullständiga originalsträngen som L2-fråga.

⁹I så fall markerar programmet inte det relevanta segmentet i L1-meningen.

¹⁰I fortsättningen kommer vi att använda benämningen *tempusformer* när vi endast talar om svenskan och *tempus-aspektformer* när vi talar om engelskan eller båda språken på samma gång, eftersom engelskan har en grammatisk form för progressiv aspekt (se avsnitt 3.1).

¹¹För enkelhetens skull använder vi de svenska benämningarna på tempusformer i båda språken här.

(sv. *hade ätit*, eng. *had eaten*) och futurala tempusformer (sv. *ska/kommer att äta*, eng. *will/ be going to eat*). En skillnad mellan språken är att engelskan har en form för progressiv aspekt, nämligen *be + -ing*-formen, vilken även kan förekomma med PresPf, t.ex. *I have been eating* (jfr. *I have eaten*). Båda formerna kan översättas med presens perfekt i svenskan: *Jag har ätit*. Svenskan har olika konstruktioner för att uttrycka progressiv aspekt (t.ex. *hålla på att äta*), men – till skillnad från engelskan – ingen inflektionell form för progressiv aspekt [10].

Svenskans presens perfekt består av hjälpverbet *ha* i presens och supinumform av huvudverbet, t.ex. *har sett*, *har ätit*, och syftar på en förfluten aktion¹² (jfr [11] 4:152). En sats som *Jag har ätit min macka* anger att den förflutna aktionen “ätit mackan” har relevansen att “mackan är uppäten” i nuet. I svenskan kan hjälpverbet *ha* utelämnas i bisatser [12], t.ex. *Jag ser att du redan (har) ätit*. Svenskans presens perfekt motsvaras i engelskans grammatik av present perfect som består av hjälpverbet *have* i presens och ett verb i perfekt particip, t.ex. *have eaten*.

3.2. Material och metod

Materialet som användes var parallellkorpusen Parallel Universal Dependencies (PUD) [13], vilken består av 1000 meningar samlade från wikipediatexter och tidningsartiklar. 750 av dessa meningar har engelska som ursprungsspråk. Resterande 250 meningar samlades från tyska, franska, italienska och spanska texter, men dessa meningar översattes till andra språk via engelska. Materialet har översatts till 17 språk, däribland svenska, och dess UD-annotering har utförts – eller åtminstone granskats – manuellt.

Eftersom vi var intresserade av både exakta korrespondenser och översättningsdivergenser, bestämde vi att köra två separata enspråkiga sökningar: en på svenskans PresPf för att få fram träffar på engelskans motsvarigheter, och en på engelskans PresPf för att få fram träffar på svenskans motsvarigheter. Sökningen på engelskans PresPf gjordes utifrån söksträngen

```
TREE_ (FEATS_ "VerbForm=Part") [AND [LEMMA "have", FEATS_ "Tense=Pres"]]
```

och på svenskans PresPf utifrån söksträngen

```
TREE_ (FEATS_ "VerbForm=Sup") [AND [LEMMA "ha", FEATS_ "Tense=Pres"]].
```

Den engelska söksträngen gör att vi även kan hitta förekomster av present perfect progressive-konstruktioner, t.ex. *have been eating*. Söksträngarna möjliggör även att passiva verbformer såsom *har setts* och *have been seen* kan förekomma i materialet. I denna artikel tittar vi emellertid inte närmare på diates (dvs. förhållandet mellan aktiva och passiva konstruktioner), utan fokuserar på tempus-aspekt.

3.3. Resultat och diskussion

Sökningen på engelskans PresPf gav 87 träffar (se bilaga A för fullständiga sökresultat). Tempusformerna i den svenska översättningen anges i tabell 1. PresPf förekom i de allra flesta fallen (t.ex. *har föreslagit*), men i sex fall var hjälpverbet utelämnat och supinum förekom ensamt (t.ex.

¹²Med *aktionen* menar vi alla sorters företeelser som verb kan syfta på, både dynamiska t.ex. *springa* och *bli*, och statiska, t.ex. *vara*.

de tidslinjer som flutit runt i medierna, där *flutit* är ensamt supinum). I ett fall förekom preteritum (*visade*) och i tre fall förekom enkelt presens (t.ex. huvud verbet *har*, där engelskan använde PresPf *has got*).

Tabell 1

Svenska motsvarigheter till engelska present perfect

Form	Frekvens
Presens perfekt	77 (88,5%)
Supinum utan hjälpverb	6 (6,9%)
Enkelt presens	3 (3,4%)
Preteritum	1 (1,1%)
Summa	87 (100%)

Tabell 2

Engelska motsvarigheter till svenska presens perfekt

Form	Frekvens
Present perfect non-progressive	75 (94,9%)
Present perfect progressive	3 (3,8%)
Present tense	1 (1,3%)
Summa	79 (100%)

Sökningen på svenskans PresPf gav 79 träffar (se bilaga B för fullständiga sökresultat). Tempus-aspektformerna som motsvarar svenskans presens perfekt i den engelska texten anges i tabell 2. Non-progressive i tabellen står för fall där progressivformen inte användes. Present perfect non-progressive (t.ex. *have suggested*) förekom i de allra flesta fallen, men present perfect progressive (t.ex. *has been rising*) förekom tre gånger. I samtliga tre fall användes ingen konstruktion i svenskan för att ange progressivitet, utan det var endast PresPf-formen som användes, t.ex. *har ökat*. Den enda förekomsten av present tense i engelska var i en passivkonstruktion (*are referred to*), där svenskan använde PresPf i en passivkonstruktion (*har kejsarna ofta omtalats*).

Resultatet visade på stor överensstämmelse mellan svenskans och engelskans PresPf. Emellertid användes engelskans PresPf i en aningen större utsträckning än svenskans PresPf. Variationen mellan språken tycks både handla om språkspecifika konventioner och översättningsstrategier. Exempelvis översattes engelskans PresPf i [...] *I have made a full recovery* med svenskans enkelt presens och perfekt particip [...] *är jag fullt återhämtad*. Detta torde bero på att *make a full recovery* är närmast ett idiomatiskt uttryck i engelskan som inte har en direkt motsvarighet i svenskan: därför kan översättaren ha föredragit uttrycket *vara fullt återställd* som i sammanhanget låter naturligare i presens än i PresPf.

Vidare kan vårt resultat bero på genrespecifika tendenser. En pilotstudie som vi utförde i en annan UD-korpus, LinES [14, 15], som består av bland annat skönlitterära texter, visade att svenskans PresPf användes i högre utsträckning än engelskans PresPf. I framtiden vore det därför fördelaktigt att titta på ett större material från olika genrer för att undersöka distributionen mellan tempus-aspektformer i de två språken.

4. Slutsatser och framåtblick

Verktyget var användbart för att hitta motsvarande tempus-aspektformer i de två språken. En fördel visade sig vara att vi med enkelhet kunde hitta alla PresPf-former, eftersom verktyget kan hitta dependenser mellan hjälpverbet och supinum i svenskan eller perfekt particip i engelskan. Tack vare TREE-frågor kunde verktyget göra det även när det förekom flera ord mellan hjälpverbet och huvud verbet, som i t.ex. *har kejsarna ofta omtalats*. Den engelska sökningen gjorde det möjligt för oss att även hitta de fall där hjälpverbet var utelämnat. Som visas i figur 1 markerar

verktyget de frågematchande segmenten med fet stil. Detta gjorde sökresultaten lättöverskådliga. Att sökningarna returnerar motsvarande segment underlättade den visuella inspektionen av sökresultaten. I denna första studie använde vi inga tvåspråkiga frågor. I framtiden kan dock tvåspråkiga frågor vara användbara för att leta efter de översättningsdivergensmönster som vi fann i andra korpusar, såsom den ovan nämnda LinES (se 3.3).

När det gäller själva verktyget är målet att det ska bli mer användarvänligt. I detta syfte planerar vi att omimplementera det som webbapplikation, integrera det med en automatisk UD-parser för att göra det lättare att undersöka oannoterade texter samt implementera trädvisualiseringar direkt i användargränssnittet. Dessutom vill vi tillgängliggöra en lista med vanliga ersättningsmönster att välja mellan. Slutligen kommer en del optimeringsarbete att krävas för att göra verktyget praktiskt tillämpningsbart på storskaliga trädbanker.

Fallstudien visade på likheter och skillnader i distribution mellan de två språken, men vi har i föreliggande artikel inte undersökt semantiken hos tempus-aspektformerna: detta vill vi göra i framtida studier. I framtiden vill vi också använda verktyget till att bland annat undersöka vilka former språk som inte har en direkt motsvarighet till svenskans PresPf – till exempel japanskan – använder i samma situationer som svenskan använder PresPf i. En sådan studie skulle visa på hur språk som har olika språkliga kategorier använder sitt språks resurser i liknande kontexter.

En pilotstudie där vi sökte på japanskans motsvarigheter till svenskans PresPf i PUD-korpusen visade att det förekommer större variation mellan svenskan och japanskan än mellan svenskan och engelskan. Denna variation verkar delvis bero på att japanskan har många olika tempus-aspektformer som kan användas på liknande sätt som svenskans PresPf. I vissa fall fanns det emellertid ingen direkt översättningsmotsvarighet till PresPf i den japanska texten, vilket tyder på variation som beror på översättningsstrategier från det engelska källspråket. Verktyget torde därför kunna användas både för att undersöka likheter och skillnader i strukturer mellan två språk, och för att undersöka översättningsvetenskapliga frågor.

Referenser

- [1] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, *Computational Linguistics* 47 (2021) 255–308. URL: https://doi.org/10.1162/coli_a_00402. doi:10.1162/coli_a_00402.
- [2] M. Straka, UDPipe 2.0 prototype at CoNLL 2018 UD shared task, in: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 197–207. URL: <https://aclanthology.org/K18-2020>. doi:10.18653/v1/K18-2020.
- [3] P. Kolachina, A. Ranta, From abstract syntax to Universal Dependencies, *Linguistic Issues in Language Technology* 13 (2016). URL: <https://aclanthology.org/2016.lilt-13.4>.
- [4] A. Ranta, P. Kolachina, From Universal Dependencies to abstract syntax, in: *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, Association for Computational Linguistics, Gothenburg, Sweden, 2017, pp. 107–116. URL: <https://aclanthology.org/W17-0414>.
- [5] A. Masciolini, A query engine for L1-L2 parallel dependency treebanks, in: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, University of Tartu

- Library, Tórshavn, Faroe Islands, 2023, pp. 574–587. URL: <https://aclanthology.org/2023.nodalida-1.57>.
- [6] A. Masciolini, E. Volodina, D. Dannélls, Towards automatically extracting morphosyntactical error patterns from L1-L2 parallel dependency treebanks, in: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 585–597. URL: <https://aclanthology.org/2023.bea-1.50>.
- [7] A. Masciolini, A. Ranta, Grammar-based concept alignment for domain-specific Machine Translation, in: Proceedings of the Seventh International Workshop on Controlled Natural Language (CNL 2020/21), Special Interest Group on Controlled Natural Language, Amsterdam, Netherlands, 2021. URL: <https://aclanthology.org/2021.cnl-1.2>.
- [8] H. de Swart, et al., Perfect usage across languages, *Questions and answers in linguistics* 3 (2016) 57–62.
- [9] I. Larsson, B. Lyngfelt, Tempus i svenskan, in: *Tid och tidsförhållanden i olika språk*, number 2 in *Studia Interdisciplinaria, Linguistica et Litteraria (SILL)*, 2011.
- [10] K. Blensenius, En pluraktionell progressivmarkör? hålla på att jämförd med hålla på och, *Språk och stil* 23 (2013) 175–204.
- [11] U. Teleman, S. Hellberg, E. Andersson, *Svenska Akademiens grammatik*, Svenska Akademien, 1999.
- [12] L. Bäckström, *Etableringen av ha-bortfall i svenskan. Från kontaktfenomen till inhemsk konstruktion*, Institutionen för svenska språket, Göteborgs universitet, 2020.
- [13] D. Zeman, J. Hajič, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, S. Petrov, CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies, in: D. Zeman, J. Hajič (Eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1–21. URL: <https://aclanthology.org/K18-2001>. doi:10.18653/v1/K18-2001.
- [14] L. Ahrenberg, LinES: An English-Swedish parallel treebank, in: J. Nivre, H.-J. Kaalep, K. Muischnek, M. Koit (Eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, University of Tartu, Estonia, Tartu, Estonia, 2007, pp. 270–273. URL: <https://aclanthology.org/W07-2441>.
- [15] L. Ahrenberg, Converting an English-Swedish parallel treebank to Universal Dependencies, in: J. Nivre, E. Hajičová (Eds.), *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala University, Uppsala, Sweden, Uppsala, Sweden, 2015, pp. 10–19. URL: <https://aclanthology.org/W15-2103>.

A. Sökresultat för TREE_ (FEATS_ "VerbForm=Part") [AND [LEMMA "have", FEATS_ "Tense=Pres"]]

Svenska	Engelska
<p>Den andelen har ökat stadigt med åren – bara 11 procent av de samlade rösterna lades före valdagen 1996 enligt folkbokföringsbyrå – och det verkar troligt att den kommer öka ordentligt igen .</p> <p>” Vi har bett andra stater att hjälpa oss befolka djurparken med olika djurarter , inklusive en gris ” , sade Saqib .</p> <p>Hon åttalas även för att ha försökt mörda sin tvååriga dotter .</p> <p>Flera analytiker har föreslagit att Huawei har bäst position för att tjäna på Samsungs tillbakagång .</p> <p>Men det kan vara så att BA och IAG har knäckt det och kan erbjuda något som kan verka pålitligt .</p> <p>10-veckorskursen har ” certifierats ” av brittiska spionmyndigheten GCHQ .</p> <p>Genom historien har den internationella hårmarknaden alltid haft en politisk dimension , säger Tarlo .</p> <p>Shenzhens trafikpolis har valt okonventionella straff förut .</p> <p>Seagal , vars farmor kom från Vladivostok i Rysslands östligaste delar har gjort flera resor till Ryssland de senaste åren och besökte Kamtjatka och Sakhalin i september .</p> <p>Forskare har undersökt potentialen för manliga hormonella preventivmedel i ungefär 20 år .</p> <p>Ms Pugh har fått behandling vid Papworths och Addenbrookes sjukhus i Cambridgeshire .</p> <p>Men en undersökning visade att tumören i Ms Pughs högra lunga växer och hon har varit tvungen att lämna försöket .</p> <p>Men en undersökning visade att tumören i Ms Pughs högra lunga växer och hon har varit tvungen att lämna försöket .</p> <p>Studenter som Rai har träffat kuratorer på skolan för att prata om det som hände , men hon säger att den största trösten har kommit från att träffa sina vänner .</p> <p>Studenter som Rai har träffat kuratorer på skolan för att prata om det som hände , men hon säger att den största trösten har kommit från att träffa sina vänner .</p> <p>Det bör inte finnas något företräde i regeringens tillgänglighet , eller sken av företräde i tillgängligheten , som tilldelas individer eller organisationer för att de har givit finansiella bidrag till politiker och politiska partier , lyder riktlinjerna .</p> <p>Och , medgav hon , ” du måste se på var hon har medgivit att vi måste göra något annorlunda – vi kan bättre – och där hon har uttryckt ånger . ”</p> <p>Och , medgav hon , ” du måste se på var hon har medgivit att vi måste göra något annorlunda – vi kan bättre – och där hon har uttryckt ånger . ”</p> <p>Som många jag känner har jag tillbringat de senaste månaderna med att sitta uppe sent och i skräck läsa opinionsundersökningar . Christian Wolmar som har skrivit ett antal böcker om järnvägshistoria kommer att kandidera den 1 december .</p> <p>Samtidigt har den frånskilda frun till en av regeringens inpiskare lanserat sin kampanj för att ta platsen .</p> <p>Ett av deras många sörjande facebookinlägg har en rad som skulle ha gjort Chris förtjust : ” Mer afrikansk än du , har jag ej känt . ”</p>	<p>That share has been rising steadily over the years — only 11 percent of the total vote was cast before Election Day in 1996 , according to the Census Bureau -- and seems likely to jump again this year .</p> <p>“ We 've requested other nations to help us populate the zoo with different species of animals , including a pig , ” Saqib said .</p> <p>She has also been charged with trying to kill her two - year - old daughter .</p> <p>Several analysts have suggested Huawei is best placed to benefit from Samsung 's setback .</p> <p>But it may be that BA and IAG have cracked it and can offer something vaguely reliable . ”</p> <p>The 10 - week course has been "certified by UK spy agency GCHQ .</p> <p>Throughout history , the international hair market has always had a political dimension , says Tarlo .</p> <p>Shenzhen 's traffic police have opted for unconventional penalties before .</p> <p>Seagal , whose grandmother was from Vladivostok in Russia 's far east , has made frequent trips to Russia in recent years and visited Kamchatka and Sakhalin in September .</p> <p>Researchers have been investigating potential for male hormonal contraceptives for around 20 years .</p> <p>Ms Pugh has received treatment at Papworth and Addenbrooke 's Hospitals in Cambridgeshire .</p> <p>But a scan has shown the tumour in Ms Pugh 's right lung is growing , and she has had to leave the trial .</p> <p>But a scan has shown the tumour in Ms Pugh 's right lung is growing , and she has had to leave the trial .</p> <p>Students like Rai have been meeting with counsellors at the school to talk about what happened , but she said the biggest comfort has come from seeing her friends .</p> <p>Students like Rai have been meeting with counsellors at the school to talk about what happened , but she said the biggest comfort has come from seeing her friends .</p> <p>There should be no preferential access to government , or appearance of preferential access , accorded to individuals or organizations because they have made financial contributions to politicians and political parties , the guidelines read .</p> <p>And , she granted , “ you have to look at where she has acknowledged that we need to do something different — we can do better — and where she has expressed regret . ”</p> <p>And , she granted , “ you have to look at where she has acknowledged that we need to do something different — we can do better — and where she has expressed regret . ”</p> <p>Like many people I know , I 've spent recent months staying up late , reading polls in terror .</p> <p>Christian Wolmar , who has written a number of books on railway history , will stand in the contest on 1 December .</p> <p>Meanwhile , the estranged wife of a government whip has launched her campaign to take the seat .</p> <p>One of their many grieving Facebook posts has a line which would have delighted Chris : “ More African than you , I have not known . ”</p>

Han **har** den där hårda eggen i sitt spel men också de mjuka händerna .

Han **har uttalat** sig till stöd för tortyr .

RSPB:s ståndpunkt **har** även **gjort** att det kommit i konflikt med flera ledande naturvårdare , från namninsamlingens författare Mark Avery till TV-presentatören Chris Packham .

Men stora minskningar **har** redan **skissats** ned vad gäller torsk , sjötunga , spätta , glasvar och sej i Keltiska havet och Irländska sjön .

Hong Kongs regering , som leds av Beijing-vänliga lagstiftare , **har ansett** att paret inte bör tillträda posten .

Att spela lika en match du **dominerat** är lätt att skaka av sig , men att göra så tre gånger i rad tyder på en svaghet .

Vi är så besvikna eftersom vi **har tappat** sex poäng när vi spelat hemma .

Den årliga enkäten visade också att **oron** över att ta på sig nya ekonomiska bördor har skjutit i höjden .

Det har även öppnat upp VW för enorma skadeståndsanspråk , rättsliga åtgärder och **har sett** det ta mer än 16 md euro i avskrivningar .

Det **har** även **öppnat** upp VW för enorma skadeståndsanspråk , rättsliga åtgärder och har sett det ta mer än 16 md euro i avskrivningar .

Företag hade förväntat sig att minska i juli , direkt efter Brexitomröstningen , men **har** istället **kunnat fortsätta** växa stadigt .

Detta **har** inte **hindrat** investerare från att flockas för att sätta in sina pengar i fonderna .

Detta betyder att de inte **har tjänat** på den ökning som sterlingens fall har orsakat hos tillgångar utomlands .

Detta betyder att de inte har tjänat på den ökning som sterlingens fall har orsakat hos tillgångar utomlands .

Efter en del genial kirurgi och mycket väldigt krävande rehabilitering **är jag fullt återställd** .

Fyra av tio vuxna **har skadats** på grund av dåligt väder medan nio av tio underskattar hur kyligt Storbritannien kan bli vintertid .

Enligt regissörens egen beräkning **har** han till dags dato **gjort** åtta långfilmer .

Mer än 5,7 miljoner floridabor **har** redan **gått** och röstat efter cirka två veckor med personligt förtidsröstande .

En taleskvinna för Harley-Davidson sade att de tidslinjer som **flutit** runt i medierna – 2017 eller 2020 – inte stämmer .

I städerna där vi **lanserat** oss eller är under uppbyggnad kommer vårt arbete att fortsätta , sade Barratt .

I Nederländerna **har myndigheterna tagit** till ett lågteknologiskt angreppssätt för att spåra drönare .

Det finns ingen parad och det **har** aldrig **funnits** en .

Alejandra González Anaya , en av paradens kreativa chefer , förklarade för CNN varför Mexiko **har beslutat** att det är dags nu att bjuda på en sådan show .

Fastän Island stod under Danmarks politiska kontroll fram till ett mycket senare datum (1918) **har** väldigt lite inflytande och inlån från danskan **skett** i det isländska språket .

Format av globala plattetektoniska krafter som **skapat** den östafrikanska riften , är östra Afrika platsen för Kilimanjaro och Mount Kenya , de två högsta bergtopparna i Afrika .

Danevirke **har förblivit** i tysk ägo sedan dess .

Global uppvärmning **har orsakat** en förändring i regnperiodernas mönster .

Den innehåller en liten lagun som nästan **har torkat** ut .

He **has got** that hard edge to his game but also the soft hands .

He **'s spoken** in favour of torture .

The RSPB 's stance **has** also **brought** it into conflict with many leading conservationists , from the author of the petition , Mark Avery , to the TV presenter Chris Packham .

But big reductions **have** also **been pencilled** in for cod , sole , plaice , megrim and pollack in the Celtic and Irish Seas .

The Hong Kong government , which is led by pro-Beijing lawmakers , **have argued** that the pair should not take office .

To draw a game you **have dominated** is easy to brush off , but to do so three times in a row suggests a weakness .

We are so disappointed because we **have dropped** six points playing at home .

The annual survey also revealed that worries about taking on fresh financial burdens **has rocketed** .

It has also opened VW up to huge compensation claims , legal action and **has seen** it take more than € 16bn of provisions .

It **has** also **opened** VW up to huge compensation claims , legal action and has seen it take more than € 16bn of provisions .

Businesses had expected to start contracting in July , immediately after the Brexit vote , but instead **have managed** to keep growing steadily .

This **has** not **stopped** investors flocking to put their money in the funds .

This means that they **have** not **benefited** from the uplift that the fall in sterling has given to overseas assets .

This means that they have not benefited from the uplift that the fall in sterling **has given** to overseas assets .

After some genius surgery and a lot of very tough rehab , I **have made** a full recovery .

Four in 10 adults **have been injured** due to bad weather while nine in 10 underestimate how chilly Britain can get in the winter time .

According to the director 's own count , to date he **has made** eight feature films .

More than 5.7 million Floridians **have** already **hit** the polls after about two weeks of in - person early voting .

A Harley - Davidson spokeswoman said timelines that **have been floated** in the media -- 2017 or 2020 -- are n't accurate .

"In the cities where we **'ve launched** or are under construction , our work will continue , Barratt said .

In the Netherlands , authorities **have taken** a lower - tech approach to tracking drones .

There is no parade and there never **has been** .

Alejandra González Anaya , one of the parade 's creative directors , explained to CNN why Mexico **have decided** now is the time to put on such a show .

Although Iceland was under the political control of Denmark until a much later date (1918) , very little influence and borrowing from Danish **has occurred** in the Icelandic language .

Shaped by global plate tectonic forces that **have created** the East African Rift , East Africa is the site of Mount Kilimanjaro and Mount Kenya , the two tallest peaks in Africa .

The Danevirke **has remained** in German possession ever since .

Global warming **has caused** a change in the pattern of the rainy seasons .

It contains a tiny lagoon , which **has** all but **dried** up .

De senaste 50 åren **har** förstörelse av livmiljöer runt jordbruksmarker **försämrat** cirka 40 % av jordbruksarealer världen över genom erosion , försaltning , kompaktering , utarmning av näringsämnen , förorening och urbanisering .

Dammstormar börjar i öknarnas utkanter snarare än i själva öknarna där de finare materialen redan har blåsts bort .

I Thailand **har** dock urbaniseringen också **resulterat** i massiva öknningar av problem som fetma .

Tyvärr **har** en snabb befolkningstillväxt och stadsspridning **täckt** mycket av dessa ekosystem med bebyggelse .

Sedan 1960-talet **har** den sydkoreanska ekonomin **vuxit** enormt och den ekonomiska strukturen transformerades radikalt .

Aldrin **har varit gift** tre gånger .

Låg inkomst per capita **har föreslagits** som en anledning till missnöje , som leder till beväpnat uppror .

Det **har dock gjorts** omfattande statistisk analys av demografi- och befolkningsdata som inkluderar kvinnor , särskilt i sin barnafödande roll .

Radarsystemet som Alvarez är mest känd för och som **har spelat** en stor roll i luftfart , alldeles särskilt i efterkrigets luftbro till Berlin , var Ground Controlled Approach (GCA) .

Fram till augusti 2015 **har** Kesha **släppt** väldigt lite information om hennes kommande tredje studioalbum .

Med 71 mål på 137 internationella matcher **har** han **slagit** rekord i antal mål för DFB .

Bilden visar tydligt den smala pekskärmsstrimman ovanför tangentbordet som **har förutsetts** i rykten .

En svensk studie **har visat** att djur sällan , om alls , landar efter den 2 månader långa parningsperioden .

Lagen beskriver ett antal villkor vars uppfyllelse kan få överenskommelsen att gälla igen : USA skulle behöva dra tillbaka alla sina trupper från länder som gått med i NATO efter 2000 , upphäva alla sanktioner mot Ryssland samt ersätta de kostnader som har uppkommit som en följd av sanktionerna .

Det är en politisk process och jag **har beslutat** att inte vara närvarande , så var det sagt .

För tillfället **har** överenskommelsen mellan Aoun och Hariri **fört** de två fiendefaktionerna närmare tillsammans .

Det är ingen överraskning att de federala och statliga regeringarna har klassificerat det nationella naturarvet som en landsomfattande strävan av högsta prioritet och **har dokumenterat** dess början 2005 i ett koalitionskontrakt .

Det är ingen överraskning att de federala och statliga regeringarna **har klassificerat** det nationella naturarvet som en landsomfattande strävan av högsta prioritet och har dokumenterat dess början 2005 i ett koalitionskontrakt .

Till skillnad från det från 28 oktober **har** mr Comeys brev knappt **kommenterats** .

AKP:s islamistiska konservativa **har tagit** denna vägran som ett rättfärdigande för utfrågningarna .

Guiden är problematisk , för det första eftersom den **har tagits fram** ” i ett sammanhang av undantagstillstånd ” .

Det **har betonats** att vi absolut inte kan fortsätta med de som är helt emot Italien .

CGI Mestre **har uppgett** i en notering att det inte kommer förekomma någon höjning av moms , i alla fall inte under 2017 .

Corrado Passera drar tillbaka erbjudandet för Mps på grund av ” den attityd av fullkomlig slutenhet som banken har visat oss ” .

Det är vad som **skrivits** i en anteckning till Mps efter ex-ministerns beslut att lämna partiet .

Over the past 50 years , the destruction of habitat surrounding agricultural land **has degraded** approximately 40 % of agricultural land worldwide via erosion , salinization , compaction , nutrient depletion , pollution , and urbanization .

Dust storms usually start in desert margins rather than the deserts themselves where the finer materials **have** already **been blown** away .

However , in Thailand , urbanization **has also resulted** in massive increases in problems such as obesity .

Unfortunately , rapid population growth and urban sprawl **has covered** much of these ecosystems with development .

Since the 1960s , the South Korean economy **has grown** enormously and the economic structure was radically transformed .

Aldrin **has been married** three times .

Low per capita income **has been proposed** as a cause for grievance , prompting armed rebellion .

There **has** , however , **been** extensive statistical analysis of demographic and population data which includes women , especially in their childbearing roles .

The radar system for which Alvarez is best known and which **has played** a major role in aviation , most particularly in the post war Berlin airlift , was Ground Controlled Approach (GCA) .

As of August 2015 , Kesha **has released** little information about her upcoming third studio album .

With 71 goals in 137 international matches , he **has shot** a record number of goals for the DFB .

The photo clearly shows the narrow touch display bar above the keyboard that **has been anticipated** in rumors .

A Swedish study **has shown** that animals seldom , if at all , land after the 2 month breeding period .

The law delineates a number of conditions whose fulfillment could bring the agreement back into effect : the USA would have to withdraw all of its troops from countries who joined NATO after 2000 , rescind all of the sanctions against Russia as well as reimburse the costs that **have been incurred** as a result of the sanctions .

It is a political process and I **have decided** not to be present , so it was said .

For the time being , the deal between Aoun and Hariri **has brought** the two enemy factions closer together .

It is no surprise that the federal and state governments have classified the National Natural Heritage as a nation - wide endeavor of the highest priority and **have documented** it starting in 2005 in a coalition contract .

It is no surprise that the federal and state governments **have classified** the National Natural Heritage as a nation - wide endeavor of the highest priority and have documented it starting in 2005 in a coalition contract .

Unlike that of the 28th of October , Mr. Comey 's letter **has hardly been commented** upon .

The AKP 's Islamic conservatives **have taken** this refusal as justification for the questionings .

The guide is problematic , firstly because it **has been developed** "in the context of a state of emergency " .

It **has been emphasised** that we absolutely can not continue with those who are completely against Italy .

The CGI Mestre **have stated** in a note , that there will be no rise in VAT , for 2017 at least .

Corrado Passera withdraws the offer for Mps due to the attitude of total closure that the Bank **has shown** to us " .

This is what **has been written** in a note to Mps after the decision of the ex-minister to leave the party .

Svenska	Engelska
<p>Fyra studenter vid Roma Tre University har utvecklat en motorcykelhjälm som kan "läsa tankar" genom att hjälpa till att förutse förarens handlingar .</p> <p>Ett av exemplen på forskning som tillämpas under säkerhetsåtgärder har också producerat en humanoid robot för att tävla med Valentino Rossi .</p> <p>Amazon har blivit det fjärde amerikanska företaget med det högsta börsvärdet , överträffande ExxonMobil .</p> <p>För första gången under de senaste sex åren har arbetslöshetsgraden sjunkit under 20 % och det är redan 600 000 personer fler anställda än det var för ett år sedan .</p> <p>Fackförbunden har varit väldigt stridslystna och vid flera tillfällen har de krävt ett upphävande av den lagstiftning som Fátima Báñez är mest stolt över .</p> <p>Försvarsministern har tillåtit henne att stanna kvar som generalsekreterare för partiet för tillfället .</p> <p>De rådfrågade källorna har sagt att det är "100 % Cospedal" .</p> <p>May har fått stor kritik för att ha undvikit och inte svarat öppet till media efter rättsutlåtandet om Brexit .</p> <p>De islamiska medborgare som har bosatt sig sedan början av 1960-talet emigrerade i huvudsak från Turkiet .</p> <p>Sedan 1960-talet har stadens ekonomi varit i brant förfall .</p> <p>Emellertid har vänskapen fallit isär på grund av inofficiella samarbeten mellan de båda , vilka gett upphov till juridiska dispyter .</p> <p>Emellertid har vänskapen fallit isär på grund av inofficiella samarbeten mellan de båda , vilka gett upphov till juridiska dispyter .</p>	<p>Four students at Roma Tre University have developed a motorcycle helmet that can 'read thoughts' by helping to anticipate drivers' actions .</p> <p>One of the examples of research applied under safety measures has also produced a humanoid robot to compete with Valentino Rossi .</p> <p>Amazon has become the fourth American company with the largest market capitalisation , surpassing ExxonMobil .</p> <p>For the first time in the last six years the rate of unemployment has dropped below 20 % , and there are already 600,000 more people employed than there were a year ago .</p> <p>The unions have been very combative and on several occasions have called for the repeal of the legislation that Fátima Báñez is proudest of .</p> <p>The Minister of Defense has allowed her to remain General Secretary of the party for now .</p> <p>The consulted sources have said that it is "100 % Cospedal" .</p> <p>May has received great criticism for avoiding and not responding openly to the media after the judicial ruling on Brexit .</p> <p>The Islamic citizens who have settled since the beginning of the 1960s emigrated primarily from Turkey .</p> <p>Since the 1960s , the city 's economy has been in steep decline .</p> <p>However , the friendship has fallen apart due to unofficial collaborations between the two , which has given rise to legal disputes .</p> <p>However , the friendship has fallen apart due to unofficial collaborations between the two , which has given rise to legal disputes .</p>

B. Sökresultat för TREE_ (FEATS_ "VerbForm=Sup") [AND [LEMMA "ha", FEATS_ "Tense=Pres"]]

Engelska	Svenska
<p>That share has been rising steadily over the years — only 11 percent of the total vote was cast before Election Day in 1996 , according to the Census Bureau -- and seems likely to jump again this year .</p> <p>" We 've requested other nations to help us populate the zoo with different species of animals , including a pig , " Saqib said .</p> <p>Several analysts have suggested Huawei is best placed to benefit from Samsung 's setback .</p> <p>But it may be that BA and IAG have cracked it and can offer something vaguely reliable . "</p> <p>The 10 - week course has been "certified by UK spy agency GCHQ .</p> <p>Throughout history , the international hair market has always had a political dimension , says Tarlo .</p> <p>Shenzhen 's traffic police have opted for unconventional penalties before .</p> <p>Much of the debate , from the Democratic side this year , has been about white male identity .</p> <p>Seagal , whose grandmother was from Vladivostok in Russia 's far east , has made frequent trips to Russia in recent years and visited Kamchatka and Sakhalin in September .</p> <p>Researchers have been investigating potential for male hormonal contraceptives for around 20 years .</p>	<p>Den andelen har ökat stadigt med åren – bara 11 procent av de samlade rösterna lades före valdagen 1996 enligt folkbokföringsbyrån – och det verkar troligt att den kommer öka ordentligt igen .</p> <p>" Vi har bett andra stater att hjälpa oss befolka djurparken med olika djurarter , inklusive en gris " , sade Saqib .</p> <p>Flera analytiker har föreslagit att Huawei har bäst position för att tjäna på Samsungs tillbakagång .</p> <p>Men det kan vara så att BA och IAG har knäckt det och kan erbjuda något som kan verka pålitligt .</p> <p>10-veckorskursen har "certifierats" av brittiska spionmyndigheten GCHQ .</p> <p>Genom historien har den internationella hårmarknaden alltid haft en politisk dimension , säger Tarlo .</p> <p>Shenzhens trafikpolis har valt okonventionella straff förut .</p> <p>Mycket av debatten , från Demokraternas sida i år , har handlat om vit manlig identitet .</p> <p>Seagal , vars farmor kom från Vladivostok i Rysslands östligaste delar har gjort flera resor till Ryssland de senaste åren och besökt Kamtjatka och Sakhalin i september .</p> <p>Forskare har undersökt potentialen för manliga hormonella preventivmedel i ungefär 20 år .</p>

Ms Pugh **has received** treatment at Papworth and Addenbrooke 's Hospitals in Cambridgeshire .

Students like Rai have been meeting with counsellors at the school to talk about what happened , but she said the biggest comfort **has come** from seeing her friends .

Students like Rai **have been meeting** with counsellors at the school to talk about what happened , but she said the biggest comfort has come from seeing her friends .

There should be no preferential access to government , or appearance of preferential access , accorded to individuals or organizations because they have made financial contributions to politicians and political parties , the guidelines read .

And , she granted , “ you have to look at where she has acknowledged that we need to do something different — we can do better — and where she has expressed regret . ”

And , she granted , “ you have to look at where she has acknowledged that we need to do something different — we can do better — and where she has expressed regret . ”

Like many people I know , I **'ve spent** recent months staying up late , reading polls in terror .

Christian Wolmar , who **has written** a number of books on railway history , will stand in the contest on 1 December .

Meanwhile , the estranged wife of a government whip **has launched** her campaign to take the seat .

One of their many grieving Facebook posts has a line which would have delighted Chris : “ More African than you , I have not known . ”

He **'s spoken** in favour of torture .

The RSPB 's stance **has also brought** it into conflict with many leading conservationists , from the author of the petition , Mark Avery , to the TV presenter Chris Packham .

But big reductions **have also been pencilled** in for cod , sole , plaice , megrim and pollack in the Celtic and Irish Seas .

The Hong Kong government , which is led by pro-Beijing lawmakers , **have argued** that the pair should not take office .

We are so disappointed because we **have dropped** six points playing at home .

The annual survey also revealed that worries about taking on fresh financial burdens has rocketed .

It has also opened VW up to huge compensation claims , legal action and **has seen** it take more than € 16bn of provisions .

It **has also opened** VW up to huge compensation claims , legal action and has seen it take more than € 16bn of provisions .

This **has not stopped** investors flocking to put their money in the funds .

This means that they **have not benefited** from the uplift that the fall in sterling has given to overseas assets .

This means that they have not benefited from the uplift that the fall in sterling has given to overseas assets .

Four in 10 adults **have been injured** due to bad weather while nine in 10 underestimate how chilly Britain can get in the winter time . According to the director 's own count , to date he **has made** eight feature films .

Voting **has** , in the vernacular of terror , **become** the new soft target .

More than 5.7 million Floridians **have already hit** the polls after about two weeks of in - person early voting .

Ms Pugh **har fått** behandling vid Papworths och Addenbrookes sjukhus i Cambridgeshire .

Studenter som Rai har träffat kuratorer på skolan för att prata om det som hände , men hon säger att den största trösten **har kommit** från att träffa sina vänner .

Studenter som Rai **har träffat** kuratorer på skolan för att prata om det som hände , men hon säger att den största trösten har kommit från att träffa sina vänner .

Det bör inte finnas något företräde i regeringens tillgänglighet , eller sken av företräde i tillgängligheten , som tilldelas individer eller organisationer för att de **har givit** finansiella bidrag till politiker och politiska partier , lyder riktlinjerna .

Och , medgav hon , ” du måste se på var hon **har medgivit** att vi måste göra något annorlunda – vi kan bättre – och där hon har uttryckt ånger . ”

Och , medgav hon , ” du måste se på var hon har medgivit att vi måste göra något annorlunda – vi kan bättre – och där hon **har uttryckt** ånger . ”

Som många jag känner **har jag tillbringat** de senaste månaderna med att sitta uppe sent och i skräck läsa opinionsundersökningar .

Christian Wolmar som **har skrivit** ett antal böcker om järnvägshistoria kommer att kandidera den 1 december .

Samtidigt **har** den fränskilda frun till en av regeringens inpiskare **lanserat** sin kampanj för att ta platsen .

Ett av deras många sörjande facebookinlägg har en rad som skulle ha gjort Chris förtjust : ” Mer afrikansk än du , **har jag ej känt** . ”

Han **har uttalat** sig till stöd för tortyr .

RSPB:s ståndpunkt **har även gjort** att det kommit i konflikt med flera ledande naturvårdare , från namninsamlingens författare Mark Avery till TV-presentatören Chris Packham .

Men stora minskningar **har redan skissats** ned vad gäller torsk , sjötunga , spätta , glasvar och sej i Keltiska havet och Irländska sjön .

Hong Kongs regering , som leds av Beijing-vänliga lagstiftare , **har ansett** att paret inte bör tillträda posten .

Vi är så besvikna eftersom vi **har tappat** sex poäng när vi spelat hemma .

Den årliga enkäten visade också att oron över att ta på sig nya ekonomiska bördor **har skjutit** i höjden .

Det har även öppnat upp VW för enorma skadeståndsanspråk , rättsliga åtgärder och **har sett** det ta mer än 16 md euro i avskrivningar .

Det **har även öppnat** upp VW för enorma skadeståndsanspråk , rättsliga åtgärder och har sett det ta mer än 16 md euro i avskrivningar .

Detta **har inte hindrat** investerare från att flockas för att sätta in sina pengar i fonderna .

Detta betyder att de inte **har tjänat** på den ökning som sterlingens fall har orsakat hos tillgångar utomlands .

Detta betyder att de inte har tjänat på den ökning som sterlingens fall **har orsakat** hos tillgångar utomlands .

Fyra av tio vuxna **har skadats** på grund av dåligt väder medan nio av tio underskattar hur kyligt Storbritannien kan bli vintertid .

Enligt regissörens egen beräkning **har** han till dags dato **gjort** åtta långfilmer .

Röstning **har** , i terrorns jargong , **blivit** det nya mjuka målet .

Mer än 5,7 miljoner floridabor **har redan gått** och röstat efter cirka två veckor med personligt förtidsröstande .

The Dutch students **have** yet to decide if they will be commercializing their electric motorcycle .

In the Netherlands , authorities **have taken** a lower - tech approach to tracking drones .

There is no parade and there never **has been** .

Alejandra González Anaya , one of the parade 's creative directors , explained to CNN why Mexico **have decided** now is the time to put on such a show .

Although Iceland was under the political control of Denmark until a much later date (1918) , very little influence and borrowing from Danish **has occurred** in the Icelandic language .

There are also languages derived from Finnish , having evolved separately , known known as Meänkieli in Sweden and Kven in Norway .

The Danevirke **has remained** in German possession ever since .

Global warming **has caused** a change in the pattern of the rainy seasons .

It contains a tiny lagoon , which **has** all but **dried up** .

Over the past 50 years , the destruction of habitat surrounding agricultural land **has degraded** approximately 40 % of agricultural land worldwide via erosion , salinization , compaction , nutrient depletion , pollution , and urbanization .

Dust storms usually start in desert margins rather than the deserts themselves where the finer materials have already been blown away .

However , in Thailand , urbanization **has** also **resulted** in massive increases in problems such as obesity .

Unfortunately , rapid population growth and urban sprawl **has covered** much of these ecosystems with development .

Since the 1960s , the South Korean economy **has grown** enormously and the economic structure was radically transformed .

Low per capita income **has been proposed** as a cause for grievance , prompting armed rebellion .

There **has** , however , **been** extensive statistical analysis of demographic and population data which includes women , especially in their childbearing roles .

Outside Japan , beginning with Emperor Shōwa , the Emperors **are** often **referred** to by their given names , both whilst alive and posthumously .

The radar system for which Alvarez is best known and which **has played** a major role in aviation , most particularly in the post war Berlin airlift , was Ground Controlled Approach (GCA) .

As of August 2015 , Kesha **has released** little information about her upcoming third studio album .

With 71 goals in 137 international matches , he **has shot** a record number of goals for the DFB .

The photo clearly shows the narrow touch display bar above the keyboard that **has been anticipated** in rumors .

A Swedish study **has shown** that animals seldom , if at all , land after the 2 month breeding period .

The law delineates a number of conditions whose fulfillment could bring the agreement back into effect : the USA would have to withdraw all of its troops from countries who joined NATO after 2000 , rescind all of the sanctions against Russia as well as reimburse the costs that have been incurred as a result of the sanctions .

It is a political process and I **have decided** not to be present , so it was said .

For the time being , the deal between Aoun and Hariri **has brought** the two enemy factions closer together .

De nederländska studenterna **har** ännu inte **bestämt** om de kommer att kommersialisera sin elektriska motorcykel .

I Nederländerna **har** myndigheterna **tagit** till ett lågteknologiskt angreppssätt för att spåra drönare .

Det finns ingen parad och det **har** aldrig **funnits** en .

Alejandra González Anaya , en av paradens kreativa chefer , förklarade för CNN varför Mexiko **har beslutat** att det är dags nu att bjuda på en sådan show .

Fastän Island stod under Danmarks politiska kontroll fram till ett mycket senare datum (1918) **har** väldigt lite inflytande och inlån från danskan **skett** i det isländska språket .

Det finns även språk härledda från finskan , som **har utvecklats** separat , kända som meänkieli i Sverige och kända kven i Norge .

Danevirke **har förblivit** i tysk ägo sedan dess .

Global uppvärmning **har orsakat** en förändring i regnperiodernas mönster .

Den innehåller en liten lagun som nästan **har torkat** ut .

De senaste 50 åren **har** förstörelse av livmiljöer runt jordbruksmarker **försämrats** cirka 40 % av jordbruksarealer världen över genom erosion , försaltning , kompaktering , utarmning av näringsämnen , förorening och urbanisering .

Dammstormar börjar i öknarnas utkanter snarare än i själva öknarna där de finare materialen redan **har blåsts** bort .

I Thailand **har** dock urbaniseringen också **resulterat** i massiva ökningar av problem som fetma .

Tyvärr **har** en snabb befolkningstillväxt och stadsspridning **täckt** mycket av dessa ekosystem med bebyggelse .

Sedan 1960-talet **har** den sydkoreanska ekonomin **vuxit** enormt och den ekonomiska strukturen transformerades radikalt .

Låg inkomst per capita **har föreslagits** som en anledning till missnöje , som leder till beväpnat uppror .

Det **har** dock **gjorts** omfattande statistisk analys av demografi- och befolkningsdata som inkluderar kvinnor , särskilt i sin barnafödande roll .

Utanför Japan , med början med kejsar Shōwa , **har** kejsarna ofta **omtalats** med sina förnamn , både medan de lever och postumt .

Radarsystemet som Alvarez är mest känd för och som **har spelat** en stor roll i luftfart , alldeles särskilt i efterkrigets luftbro till Berlin , var Ground Controlled Approach (GCA) .

Fram till augusti 2015 **har** Kesha **släppt** väldigt lite information om hennes kommande tredje studioalbum .

Med 71 mål på 137 internationella matcher **har** han **slagit** rekord i antal mål för DFB .

Bilden visar tydligt den smala pekskärmsstrimman ovanför tangentbordet som **har förutsetts** i rykten .

En svensk studie **har visat** att djur sällan , om alls , landar efter den 2 månader långa parningsperioden .

Lagen beskriver ett antal villkor vars uppfyllelse kan få överenskommelsen att gälla igen : USA skulle behöva dra tillbaka alla sina trupper från länder som gått med i NATO efter 2000 , upphäva alla sanktioner mot Ryssland samt ersätta de kostnader som **har uppkommit** som en följd av sanktionerna .

Det är en politisk process och jag **har beslutat** att inte vara närvarande , så var det sagt .

För tillfället **har** överenskommelsen mellan Aoun och Hariri **fört** de två fiendefaktionerna närmare tillsammans .

Engelska	Svenska
<p>It is no surprise that the federal and state governments have classified the National Natural Heritage as a nation - wide endeavor of the highest priority and have documented it starting in 2005 in a coalition contract .</p>	<p>Det är ingen överraskning att de federala och statliga regeringarna har klassificerat det nationella naturarvet som en landsomfattande strävan av högsta prioritet och har dokumenterat dess början 2005 i ett koalitionskontrakt .</p>
<p>It is no surprise that the federal and state governments have classified the National Natural Heritage as a nation - wide endeavor of the highest priority and have documented it starting in 2005 in a coalition contract .</p>	<p>Det är ingen överraskning att de federala och statliga regeringarna har klassificerat det nationella naturarvet som en landsomfattande strävan av högsta prioritet och har dokumenterat dess början 2005 i ett koalitionskontrakt .</p>
<p>Unlike that of the 28th of October , Mr. Comey 's letter has hardly been commented upon .</p>	<p>Till skillnad från det från 28 oktober har mr Comeys brev knappt kommenterats .</p>
<p>The AKP 's Islamic conservatives have taken this refusal as justification for the questionings .</p>	<p>AKP:s islamistiska konservativa har tagit denna vägran som ett rättfärdigande för utfrågningarna .</p>
<p>The guide is problematic , firstly because it has been developed "in the context of a state of emergency " .</p>	<p>Guiden är problematisk , för det första eftersom den har tagits fram ” i ett sammanhang av undantagstillstånd ” .</p>
<p>It has been emphasised that we absolutely can not continue with those who are completely against Italy .</p>	<p>Det har betonats att vi absolut inte kan fortsätta med de som är helt emot Italien .</p>
<p>The CGI Mestre have stated in a note , that there will be no rise in VAT , for 2017 at least .</p>	<p>CGI Mestre har uppgett i en notering att det inte kommer förekomma någon höjning av moms , i alla fall inte under 2017 .</p>
<p>Corrado Passera withdraws the offer for Mps due to the attitude of total closure that the Bank has shown to us " .</p>	<p>Corrado Passera drar tillbaka erbjudandet för Mps på grund av ” den attityd av fullkomlig slutenhet som banken har visat oss ” .</p>
<p>Four students at Roma Tre University have developed a motorcycle helmet that can ' read thoughts ' by helping to anticipate drivers ' actions .</p>	<p>Fyra studenter vid Roma Tre University har utvecklat en motorcykelhjälm som kan ” läsa tankar ” genom att hjälpa till att förutse förarens handlingar .</p>
<p>One of the examples of research applied under safety measures has also produced a humanoid robot to compete with Valentino Rossi .</p>	<p>Ett av exemplen på forskning som tillämpas under säkerhetsåtgärder har också producerat en humanoid robot för att tävla med Valentino Rossi .</p>
<p>Amazon has become the fourth American company with the largest market capitalisation , surpassing ExxonMobil .</p>	<p>Amazon har blivit det fjärde amerikanska företaget med det högsta börsvärdet , överträffande ExxonMobil .</p>
<p>For the first time in the last six years the rate of unemployment has dropped below 20 % , and there are already 600,000 more people employed than there were a year ago .</p>	<p>För första gången under de senaste sex åren har arbetslöshetsgraden sjunkit under 20 % och det är redan 600 000 personer fler anställda än det var för ett år sedan .</p>
<p>The unions have been very combative and on several occasions have called for the repeal of the legislation that Fátima Báñez is proudest of .</p>	<p>Fackförbunden har varit väldigt stridslystna och vid flera tillfällen har de krävt ett upphävande av den lagstiftning som Fátima Báñez är mest stolt över .</p>
<p>The Minister of Defense has allowed her to remain General Secretary of the party for now .</p>	<p>Försvarsministern har tillåtit henne att stanna kvar som generalsekreterare för partiet för tillfället .</p>
<p>The consulted sources have said that it is "100 % Cospedal " .</p>	<p>De rådfrågade källorna har sagt att det är ” 100 % Cospedal ” .</p>
<p>May has received great criticism for avoiding and not responding openly to the media after the judicial ruling on Brexit .</p>	<p>May har fått stor kritik för att ha undvikit och inte svarat öppet till media efter rättsutlåtandet om Brexit .</p>
<p>The Islamic citizens who have settled since the beginning of the 1960s emigrated primarily from Turkey .</p>	<p>De islamiska medborgare som har bosatt sig sedan början av 1960-talet emigrerade i huvudsak från Turkiet .</p>
<p>However , the friendship has fallen apart due to unofficial collaborations between the two , which has given rise to legal disputes .</p>	<p>Emellertid har vänskapen fallit isär på grund av inofficiella samarbeten mellan de båda , vilka gett upphov till juridiska dispyter .</p>

DASH Swedish National Doctoral School in Digital Humanities: From Local Expertise to National Research Infrastructure

Matti La Mela¹, Daniel Brodén², Coppélie Cocq³, Anna Foka⁴, Koraljka Golub⁵, Clelia LaMonica⁴ and Jonathan Westin²

¹ Department of ALM, Uppsala University

² Gothenburg Research Infrastructure in Digital Humanities, University of Gothenburg

³ Humlab, Umeå University

⁴ Centre for Digital Humanities and Social Sciences, Department of ALM, Uppsala University

⁵ Institute, Linnaeus University

Abstract

This paper presents the Swedish National Doctoral School in Digital Humanities: Data, Culture, and Society – Critical Perspectives (DASH) that is run in 2023–2027 by Uppsala University, Umeå University, Linnaeus University, and Gothenburg University. Though Swedish universities have established PhD courses, MA programmes and training in digital humanities previously, DASH is the first encompassing educational programme in digital humanities at the doctoral level. The present paper discusses the rationale behind the DASH doctoral school, its role in the landscape of Swedish humanities infrastructures, and provides insights from the first PhD courses and seminars. The focus of DASH is to equip PhD candidates in humanities and social sciences with knowledge and skills necessary to pursue high quality, innovative and critical research in digital humanities. DASH aims to provide knowledge in relation to digital research, its methods, tools, and critical perspectives, and to build and strengthen the networks among early career scholars. DASH facilitates access and use of the resources in the national infrastructures in the humanities, but also emerges as an element in the infrastructure by providing new resources and competences.

Keywords

digital humanities, doctoral education, research infrastructures, doctoral school, Sweden

1. Introduction

DASH (*the Swedish National Doctoral School in Digital Humanities: Data, Culture, and Society – Critical Perspectives*) is a national doctoral school funded by the Swedish Research Council during 2023–2027. Emerging as a new educational programme in the national humanities infrastructure, DASH aims to provide knowledge and skills in digital humanities (DH) research – the use and critical study of computational methods and tools in humanities scholarship (see e.g. [1]) – to PhD candidates in the humanities and social sciences.

Sweden has pioneered digital research and promoted the use of computing technologies in disciplines outside of the natural sciences (see [2]). Over more than two decades, several Swedish universities have established research centres in DH, offered PhD courses, and instituted master’s level education and training in DH. DASH, however, is the first doctoral school in DH, and addresses the growing need for education and training at the doctoral level. PhD candidates in the humanities conducting digital research at different departments across Sweden rarely have the opportunity to meet and discuss methodological or ethical dimensions of their work outside of broader national or international meeting points, for example the yearly DH conference in the Nordic and Baltic Countries

Huminfra Conference 2024, Gothenburg, 10–11 January 2024.

✉ matti.lamela@abm.uu.se (M. La Mela); daniel.broden@lir.gu.se (D. Brodén); coppelie.cocq@umu.se (C. Cocq); anna.foka@abm.uu.se (A. Foka); koraljka.golub@lnu.se (K. Golub); clelia.lamonica@abm.uu.se (C. LaMonica); jonathan.westin@lir.gu.se (J. Westin)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

(DHNB) or with the interdisciplinary Association of Internet Researchers. DASH will play an important role as the first platform for PhD candidates to meet, exchange, and develop their research.

This paper presents an overview of the DASH doctoral school and provides experiences from its activities during the first year of the doctoral school in 2023-2024. Moreover, the paper highlights how such doctoral education emerges in interaction with the national research infrastructures and the local expertise. The challenges in providing education and training in DH relate especially to the interdisciplinary nature of digital humanities research, where various humanities disciplines engage with each other, but also with computational research approaches that have often been developed outside the humanities disciplines (see e.g. [3, 4, p. 4-8]). In DASH, these questions are addressed through a collaboration between the four partner universities, their research centres in DH, and their experts with different disciplinary backgrounds: the Centre for Digital Humanities and Social Sciences at Uppsala at Uppsala University, which is also the administering organisation; Humlab, Infrastructure and Centre for Digital Humanities at Umeå University; the GRIDH (Gothenburg Research Infrastructure in Digital Humanities) at Gothenburg University; and the DH Hub at Linnaeus University. DASH provides students access to knowledge and resources in DH spread throughout Sweden, strengthened even more by the collaborating universities' participation in the national humanities research infrastructures.

2. DASH – a National Doctoral School in Digital Humanities

The main aim of DASH is to provide PhD candidates in humanities and social sciences knowledge, skills, and critical perspectives in relation to digital research, its methods and tools. Moreover, DASH aims to build and strengthen the networks among early career scholars, and enable them to work in close collaboration and develop their thesis projects within the research environments that are part of DASH. The doctoral school admits doctoral candidates from all Swedish universities during the application period every spring. DASH functions as a network, a scholarly meeting place, and an interdisciplinary resource for PhD candidates, and does not fund individual doctoral studies. In order to complete the DASH doctoral school, participating PhD candidates must complete two courses, and take part in the advanced graduate workshops and the DASH National Summer School in Digital Humanities.

The courses included in DASH are coordinated at each partner organisation, but are also collaborations. The courses are given in different periods during the academic year, to ensure flexibility in choice for the doctoral candidates both regarding timing and the variety of disciplinary approaches. The courses include:

- *Introduction to Cultural Analytics* (Uppsala) introduces methods for computational text analysis from a humanities and social sciences perspective, and combines practical hands-on tasks with critical discussion and reflection on methodological concerns.
- *Reconstructing/Deconstructing Cultural Heritage in the Digital Age* (Uppsala and Umeå) covers a variety of aspects related to tangible and intangible cultural heritage in a digital age, including the digital transformation of cultural heritage institutions focusing on possibilities, skillsets and organisational perspectives.
- *Data Management and Open Science* (Umeå) which addresses the development toward open science and the increased need for data management skills in academia. The course is designed in order to address specifically the needs and challenges of research in the humanities and social sciences.
- *Digital Research. Methods and Perspectives in the Humanities and Social Science* (Umeå) provides an introduction to key qualitative and quantitative methods including digital media ethics. Digital research is the oldest PhD course in Digital Humanities in Sweden. It has been offered by Humlab at Umeå University since 2015.
- *Digitisation, XR-technologies, and digital diagnosis within the heritage sector* (Gothenburg), where PhD candidates carry out photogrammetric documentation of heritage objects, buildings and environments, and model, simulate and visualise 3D interpretations and associated research data.

The courses have been partly developed before DASH, and exemplifies how DASH expands the local expertise and scope into national level doctoral education. Moreover, DASH includes master's level courses that have been reshaped into doctoral level education by deepening their scope and adjusting them to knowledge and skills needed in doctoral training. Among these courses are:

- *Critical Theory and Digital Transformation and Ethics, Politics and Policies in Digital Humanities* (Linnaeus): the first course applies critical theory on the implementation of digital technologies and associated digital practices. The second course on the ethical and political aspects of DH research.
- *Tools and Methods: Critical Encounters* (Uppsala) introduces a selection of tools and methods used in DH, that will be discussed critically and in relation to theoretical literature and ongoing debates in DH.

The DASH Advanced Graduate Workshop is a yearly meeting in Spring each year where PhD candidates participate in a workshop at a partner institution to discuss their research and to receive feedback from the local research milieu, external experts, and their peers. The key aim of the advanced graduate workshops is to provide a critical understanding in the use of digital methods in humanities and social sciences research. The local and external experts help to identify methods and tools that are relevant for the candidates, and share best practices about their use. The work with the experts and the teachers at the Advanced Graduate Workshops, and also the National Summer School, enable to render visible and to investigate the challenges of interdisciplinary research inherent to DH. Moreover, the candidates are helped to develop an individual action plan. Activities will be coordinated by each partner institution and their DASH coordinator annually.

The DASH National Summer School in Digital Humanities builds upon previous efforts by LNU and Uppsala University in collaboration with University of Zadar in Croatia. The summer school comprises two online weeks and one intensive week of training focusing on digital tools and programming. This includes topics like text analytical tools, topic modeling and sentiment analysis, GIS and other visualisation tools as well as machine learning, procedural programming and the use of Python libraries, web scraping and data visualisation. The Summer School is enriched by lectures, workshops and exercises on tools and themes not previously covered, by the extended expertise of DASH's partners.

3. DASH – a Collaboration Between Local and National Research Infrastructures

Digital research infrastructures weave the fabric of DH. These ensure the flexible co-operation and knowledge exchange among researchers and other stakeholders (see [5, 6]). DASH is run by the four local digital research infrastructures in Uppsala, Umeå, Gothenburg and Linnaeus universities. These research and teaching units collaborate across their universities transcending IT, humanities and social sciences disciplines. All DASH research environments have expertise in digital technology and methods, data management and critical digital humanities albeit with varied foci in each research environment including: data science, such as structured data, data aggregation and modeling; artificial intelligence and machine learning methods such as image recognition; scientific visualisation such as networks and maps, 3D modeling and digitisation; critical perspectives, including reflexive use of digital methods and critical metadata and interface design.

More importantly, DASH works in close collaboration with the national research infrastructures Huminfra and InfraVis. Huminfra is a national research infrastructure for DH. It is led by the Humanities Laboratory in Lund and partners with 12 universities and organisations with expertise in e-scientific/digital materials, research tools, and experimental methods for the humanities. DASH both collaborates and complements Huminfra by offering PhD training and engaging with Huminfra's network of partners and their resources. InfraVis is a Swedish national infrastructure for the visualisation of research data. It pools state-of-the-art visualisation competence from nine partner universities to provide advanced visualisation services on research data from. In addition, all DASH

partner organisations are variably connected to other national infrastructures such as SweDigArch and SweClarin. SweDigArch is a national infrastructure for digital archaeology aimed at providing guidelines and new approaches for digital methods for archaeology. SweClarin is a node in CLARIN-ERIC (Common Language Resources and Technology Infrastructure) which develops national and European infrastructure for advanced language technology research. Also, DASH builds on the pedagogical research collaboration between CDHU and WASP-ED, the life-long machine learning programme with a national perspective.

4. DASH – Learning Through Courses and Workshops, Experiences from Fall 2023

Two introductory ‘kick-off’ days were held in Uppsala in September. This offered students, organizers, and instructors the opportunity to meet and explore ideas for upcoming research. The programme included an introductory session which presented the doctoral school’s overarching goals, as well as an overview of resources available to PhD students. This included presentations by each of the participating institutions and DH environments, information on the DASH summer school, DARIAH (the European Digital Research Infrastructure for the Arts and Humanities) and the Huminfra infrastructure and information portal.

The kick-off days included a “Design your own DH project!” workshop during which groups were formed comprising doctoral students, CDHU research engineers, DASH organizers, DH instructors from Uppsala’s Department of ALM. Students were asked to fill out a questionnaire about their research including research questions, primary sources/datasets, methods and tools, data work, and ethical approval. While the first student intake in 2023 prioritized those who were at early stages of their doctoral studies, this session provided a valuable opportunity to brainstorm research design and work through potential challenges with other PhD students as well as experienced research engineers, researchers, and teachers. The introductory days concluded with a guest lecture and workshop on “Comics as computation” by Ilan Manouach (FNRS, ULiège, Metalab at Harvard). The workshop invited students and participants to engage with notions of automation and production within artistic practices as well as digital ethics and ownership.

Students reported that the sessions were valuable for framing their ideas, making contact with other PhD students, obtaining feedback at an early point of research design, and finding approaches they had not yet considered. It created an important contact point linking organizers with PhD students and making explicit the wide breadth of resources within the Swedish DH infrastructure, especially in terms of courses, workshops, seminars, and tools available for use.

The first courses organised for DASH students in Fall 2023 were *Introduction to Cultural Analytics* and *Tools and Methods: Critical Encounters* taught at Uppsala University. *Introduction to Cultural Analytics* is a doctoral course that introduces methods for computational text analysis from a humanities and social sciences perspective. The course proceeds thematically through a research workflow, covering data collection, curation, and various text analysis methods such as NLP methods and basic machine learning. Hands-on exercises are conducted in Python using Jupyter notebooks, and students work with data and research questions stemming from their thesis projects. The course teachers have interdisciplinary expertise in the thematic areas. The second course, *Tools and Methods*, is a course taught in the master’s programme in DH. The DASH students acted as experienced peers for the master’s students. For instance, the DASH students planned one of the seminars, where they presented and reflected on their thesis work, and provided the other students a critical introduction to a tool used in DH.

The first seminar series, organised by Humlab, focused on the timely topic of AI and its implications for the humanities. The series took place in Fall 2023, with invited speakers Assistant Professor in Philosophy Dimitri Coehlo Mollo (UmU) and Associate Professor in Informatics Karin Danielsson (UmU). Under the title “AI Today: Between Reality and Hype”, Dimitri Coehlo Mollo invited the participants to reflect upon and question central issues about AI systems, such as who benefits from the hype around AI and how these narratives change our attitudes towards technology. In a second combined lecture and workshop, Karin Danielsson introduced perspectives, understandings, preferences, and relationships that humans have to AI and/or robotic system(s) in a talk entitled “You

and me(chanical) robot. Post-human futures – what comes after tomorrow and how we can understand it?”. The talk was followed by a discussion in which the participants were invited to bring in knowledge from their disciplines and discuss what kind of futures can be pictured. The seminars are open to other PhD candidates and researchers outside DASH, as a means for the DASH students to get the opportunity to meet other peers. Similar short seminar series are offered within DASH every semester, rotating between the four partner universities, and addressing various topics within the specific expertise and research interest of each responsible organization.

5. Conclusion

The aim of DASH is to provide the PhD students with knowledge, skills, and critical perspectives in regards to digital research, to offer the students support for their PhD work, and to establish networks among early career scholars and the partner DH environments. Thereby, DASH fills a gap by providing an arena for PhD candidates to access resources, develop skills, and create networks when specific knowledge and research environments in digital humanities are lacking at their home departments.

DASH activities build on existing courses, established resources, as well as new courses and national collaborations. They also encompass the *Advanced Graduate Workshop*, and the *DASH National Summer School in Digital Humanities*. To achieve its aims, DASH builds not only on the expertise at the four partner organisations but also on the resources available in the collaborative research infrastructures. This diverse range of disciplinary approaches and backgrounds facilitates open and engaging discussions regarding the interdisciplinary nature of digital humanities scholarship. Consequently, DASH itself emerges as an element of the national humanities research infrastructure, that offers education and competences for future researchers in digital humanities.

References

- [1] Berry, D. M. (2022). Critical Digital Humanities. In J. O’Sullivan (Ed.). *The Bloomsbury Handbook to the Digital Humanities* (pp. 125–136). London: Bloomsbury Academic. <http://dx.doi.org/10.5040/9781350232143.ch-12>
- [2] Nygren, T., Foka, A., and Buckland, P. (2014). The status quo of digital humanities in Sweden: past, present and future of digital history. In *H/Soz/Kult: Kommunikation und Fachinformation für die Geschichtswissenschaften*. <https://www.hsozkult.de/debate/id/fddebate-132273>
- [3] Van Es, K., Schäfer, M. T., Wieringa, M. (2021). Tool Criticism and the Computational Turn. A 'Methodological Moment' in Media and Communication Studies. *M&K Medien & Kommunikationswissenschaft*, 69(1), pp. 46–64. <https://www.nomos-elibrary.de/10.5771/1615-634X-2021-1-46/>
- [4] Dobson, J. E. (2019). *Critical Digital Humanities: The Search for a Methodology*. Urbana, Chicago, and Springfield: University of Illinois Press.
- [5] Foka, A., Misharina, A., Arvidsson, V., and Gelfgren, S. (2018). Beyond Humanities qua Digital: Spatial and Material Development for Digital Research Infrastructures. *Digital Scholarship in the Humanities*, 33(2), pp. 264–78. <https://doi.org/10.1093/lc/fqz042>
- [6] Golub, K., Göransson, E., Foka, A., and Huvila, I. (2020). Digital humanities in Sweden and its infrastructure: Status quo and the sine qua non. *Digital Scholarship in the Humanities*, 35(3), pp. 547–556. <https://doi.org/10.1093/lc/fqz042>

Research stories on Twitter

David G. Lorentzen¹ and Gustaf Nelhans¹

¹ University of Borås, Allégatan 1, Borås, Sweden

Abstract

This paper aims to study what type of research seems to interest the users of a social network platform and then complement the data with data from an open catalogue for research, exemplifying with Twitter and Open Alex. The basic idea is to get an overview of the stories the platform content tells during three months regarding topics, disciplines, and open access status. The findings suggest that the picture look very different between the approaches to map the topics, especially when looking at the articles most mentioned compared to the ones that are most retweeted. The study mainly highlights the methodological opportunities of combining text analysis and link relationships to explore the content and public interest in academic research.

Keywords

Twitter, Open Alex, topics, open access

1. Introduction

The study's relevance is the question of what scientific stories are told on a social media platform. The paper deals with the combination of sources, where one can be categorized as a streaming source, where posts are added continuously, and the other a more static source, where resources can be looked up for complimentary data. The study takes a digital methods perspective with a focus on social science research, which then implies that what we study is the stories that are told by the content [1]. The platform we take off with is formerly known as Twitter, now renamed X. Since data were collected before the rebranding, we use Twitter in this text.


2. Method and data

Data were collected using Focalevents [2]. At the time of data collection, we could use an academic developer account that allowed for searching the archive and streaming in real-time, with a download limit of 10 million tweets a month [3]. Table 1 lists the top base URLs with the number of tweets matching each URL. The data collection period was set to the first three months of 2023, searching for tweets with the base URL `<https://doi.org>`. This also matches URLs such as `<http://www.doi.org/10.51372/bioagro351.1>`

457,775 tweets were collected in this way. We selected all tweets written in English with DOI references (non-retweets) in the next step. Following pre-processing steps in which we unshortened shortened DOIs with Python requests and validated DOIs with python-doi, we ended up with 86,829 unique DOI references. Of these, 623 were invalid, for example `<https://doi.org/10.nuts>`. Using the Open Alex API, we then looked up more data, such as title, publication year, language, text type, open access status, source, keywords, abstracts, connection to sustainable development goals, citation data and retraction status. Data for 84,608 records (97 %) were returned from Open Alex.

We used Word2Vec from the Python gensim library on the abstracts to map topics. Stop words were removed, and words were stemmed using the Porter stemmer. The Word2Vec model was trained on the data for ten epochs. For each of the top 1,000 word stems, we looked up the 100 most similar terms and kept all relationships that were stronger than 0.5. These relationships were used to create term networks

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

 david.gunnarsson_lorentzen@hb.se (D. G. Lorentzen); gustaf.nelhans@hb.se (G. Nelhans)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

for further analysis. We did this for the entire dataset, the 1,000 most retweeted DOIs, and the 1,000 most mentioned DOIs.

Apart from the topical maps, we also performed descriptive statistical analyses.

3. Findings

3.1. Descriptive statistics

Of the 84,608 references, the vast majority were articles (82,414). Seven references were retracted. A large share of the references were open access. 25,993 had gold status, 16,291 were hybrid, 10,783 were green, and 5,242 were bronze. This entails that 50% of the references were open access and 69% if we include hybrid. According to the National Library of Sweden, this is quite in line with the share of published research from Swedish academics, which was 70% of the published scholarly articles in 2022. The most cited work was a book titled “Diagnostic and Statistical Manual of Mental Disorders” with 69,177 citations, and another 26 works had citation counts of at least 10,000. 342 were in the range between 1,000 and 10,000, and 3,162 were cited between 100 and 1,000 times. The data covered works from the most recent years, with 67,099 from 2022 and later, but also some historical works, with the oldest being “IV. An account of the tubera terræ, or truffles found at Rushton in Northamptonshire; with some remarks thereon” from 1693. 1,272 works were from 2000 and earlier, of which 26 were from before 1900.

Table 1
Publication types

Publication Type	Count
Article	82,414
Book chapter	981
Book	580
Report	197
Paratext	101
Reference entry	80
Dissertation	74
Dataset	65
Editorial	58
Other	58

As discovered by [4], many sources were from the natural sciences (Table 2). We see a variety of works when looking at the most overall mentioned DOIs, including retweets (Table 3). These are the most visible articles in the dataset across the three months. Most of these are from natural sciences and medicine, but there are also some examples from social sciences and psychology, such as the article about sharing misinformation.

Table 2
Source outlets

Source	Count
bioRxiv (Cold Spring Harbor Laboratory)	1,074
Nature Communications	829
Proceedings of the National Academy of Sciences of the United States of America	693
Scientific Reports	600
eLife	584
PLOS ONE	580
Nature	524
Science	333
Science Advances	309
Cell Reports	292

Table 3
Works grouped by mentions in tweets (including retweets)

Title	DOI	Mentions count
Serious adverse events of special interest following mRNA COVID-19 vaccination in randomized trials in adults	10.1016/j.vaccine.2022.08.036	859
Sharing of misinformation is habitual, not just lazy or biased	10.1073/PNAS.2216614120	629
The management of diabetic ketoacidosis in adults—An updated guideline from the Joint British Diabetes Society for Inpatient Care	10.1111/dme.14788	540
Integrating Molecular Biology and Bioinformatics Education	10.1515/jib-2019-0005	474
The Efficacy and Use of a Pocket Card Algorithm in Status Epilepticus Treatment	10.1212/CPJ.0000000000000922	424
The use of diuretics in heart failure with congestion	10.1002/ejhf.1369	396
2021 World Health Organization guideline on pharmacological treatment of hypertension: Policy implications for the region of the Americas	10.1016/j.lana.2022.100219	388
Metabolic syndrome – a new definition and management guidelines.	10.5114/aoms/152921	369
Management of Hyperglycemia in Type 2 Diabetes, 2022	10.2337/dci22-0034	291
Plant genome sequence assembly in the era of long reads	10.1017/qpb.2021.18	270

3.2. Topics

The first map (**Figure 1**) is based on all works cited in the dataset, where each abstract is treated the same. This map shows the diversity in topics, with several distinct clusters at the bottom showing terms related to molecular medicine, pathogens, climate research, agriculture and ecology. There are methodological and theoretical terms in the centre, while words related to academia and professions, the family, and psychological terms are found in the top right corner.

In the top left corner, different clusters distinguish words with linguistic functions, e.g. the purple cluster contains various types of conjunctions. At the same time, numbers, adverbs or adjectives relating to time and temporal sequencing, comparisons, measurements, and spatial or numerical relationships are found in different clusters. The other two maps zoom in on what the Twitter users find most interesting to redistribute (retweets) (Figure 2) and talk about (mentions) (Figure 3). Similarities include the focus on medical and clinical terms (bottom left and right, while the natural sciences, especially physics, are pretty well represented in the centre of Figure 2. In Figure 3, it is harder to distinguish topics, but the bottom cluster seems to relate to clinical medicine. In contrast, the academic and social cluster, including the mention of Chat GPT, is found in the orange cluster to the left.

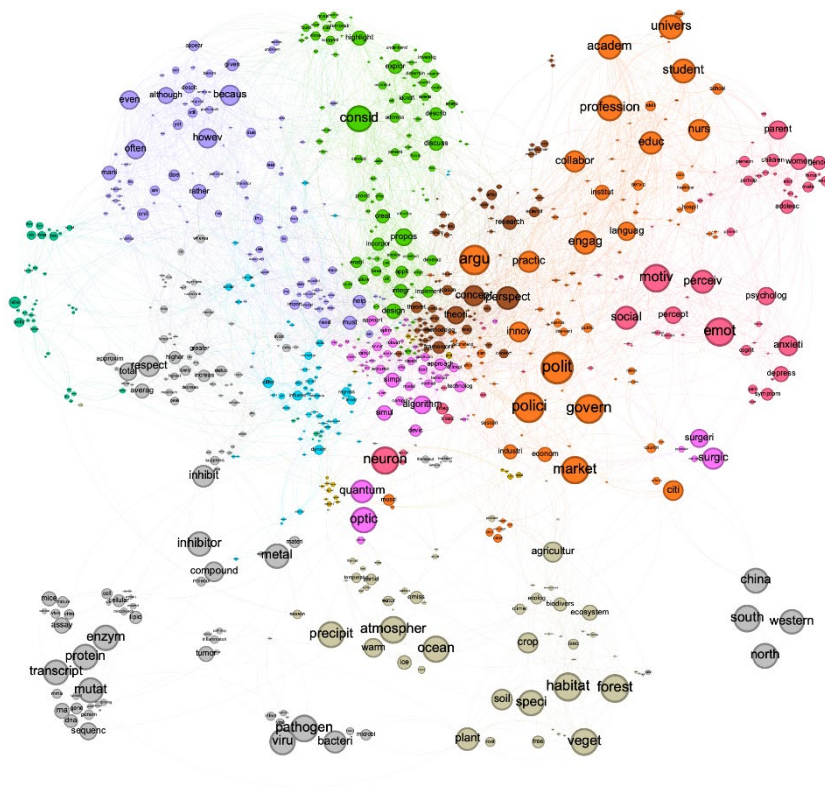


Figure 1: Word2Vec network from all cited abstracts.

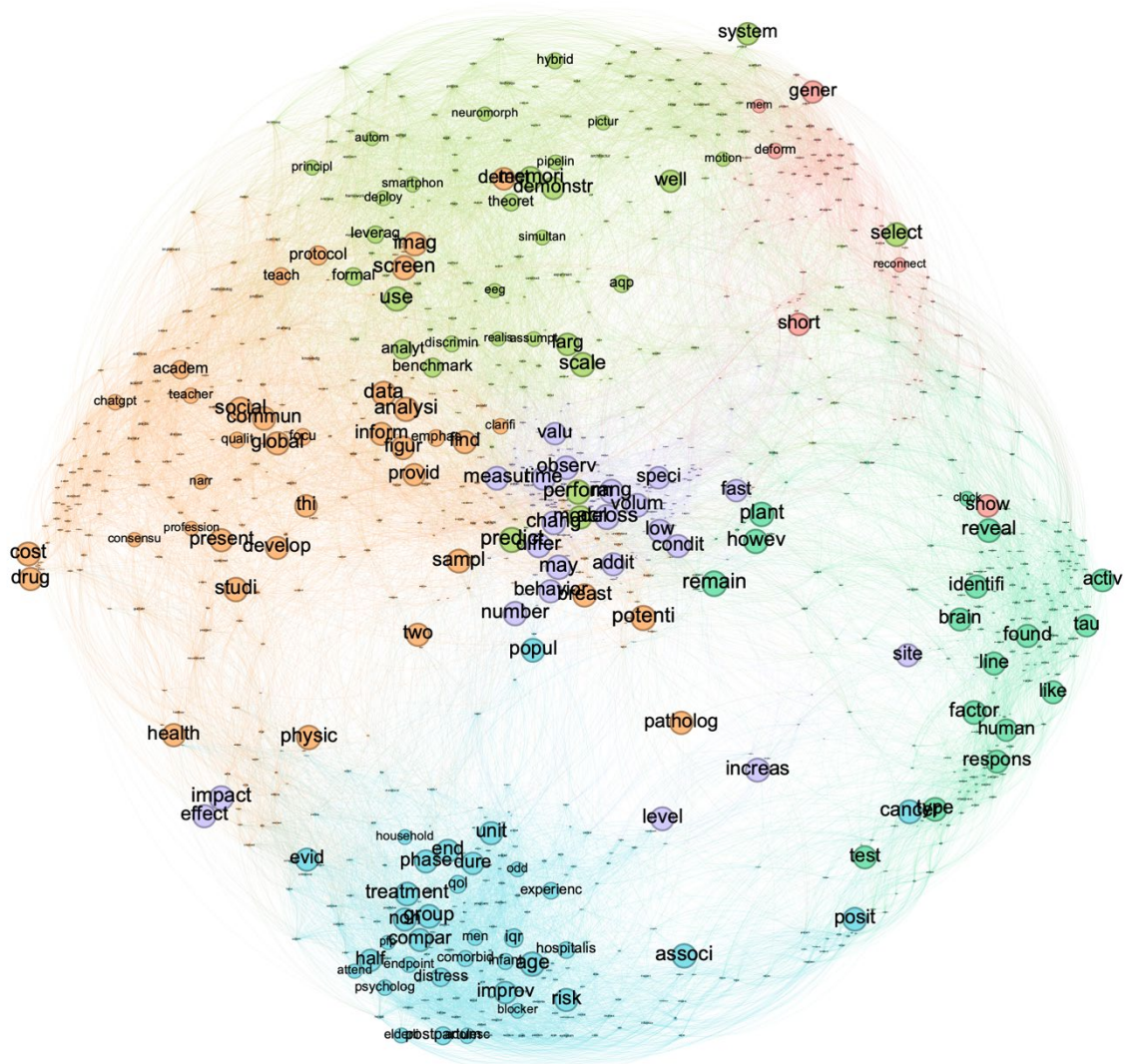


Figure 3: Word2Vec network from abstracts based on mentions.

4. Concluding remarks

What story has been told by the content? Firstly, we see a lot of open-access sources. While the most mentioned works are from the natural sciences, the topical maps show diverse coverage. However, a slightly different picture appears when the maps are filtered based on mentions and retweets. Medical themes and themes that relate to common discussions about academic issues in broader society seem prominent, i.e., the mention of Chat GPT. We have barely scratched the surface here, and some things can be learned from combining these two sources. With data from a social media platform, we gain insights into what the public finds interesting at a given moment. We can also investigate conversations and try to identify controversies surrounding academic subjects [5]. With the Open Alex API, we can get further information about the research cited by social media users. We can delve further into the practices of the users by looking at how the research is cited, for example, by looking at parts of speech to derive topics (nouns) and how something is cited (verbs) [6].

There are some drawbacks with the chosen methods that need to be highlighted. Most critical is that access to Twitter API is currently a paid service, and collecting a dataset of this size is quite costly.

Regarding data selection, we relied on DOI URLs but by doing so missing out on tweets referring to a direct URL to articles. Perhaps this decision limits the dataset to tweets created by people who are more accustomed to the academia. It is also important to keep in mind the technicalities of the platform at the time of the study. From a researcher point of view it was possible to search in the archive and collect up to ten million tweets a month, and also look up the conversations the tweets are part of. While a study of this type can reveal insights into what research the public is interested in, Twitter is not representative of the general public, and the sharing practices indicate of usage by researchers for self-promotion among other potential purposes [4]. However, when moving beyond the mere mentioning of research, argumentative patterns and practices can be revealed [7]. This paper has shown how one can use digital methods to study sharing practices on a platform in relation to a specific type of artifact, in this case research articles using their DOI URLs, and collecting additional information about what they share using an open data source. The use of digital methods to collect and analyse data makes it possible to uncover patterns that are not apparent when utilising manual analyses. Similar approaches can be used for other contexts in order to enhance the understanding of aspects of human culture.

Acknowledgements

This study was funded by Huminfra.

References

- [1] S. Niederer, *Networked Content Analysis: The Case of Climate Change*. Institute of Network Cultures, 2019
- [2] R. Gallagher, Social Media Focal Events Listener. URL: <https://focalevents.readthedocs.io/en/latest/index.html>, 2023
- [3] Twitter, Academic research access. URL: <https://web.archive.org/web/20230520042703/https://developer.twitter.com/en/products/twitter-api/academic-research>
- [4] G. Nelhans, & D. G. Lorentzen, Twitter conversation patterns related to research papers. *Information Research*, 21(2), paper SM2, 2016. URL: <http://InformationR.net/ir/21-2/SM2.html>
- [5] D. G. Lorentzen, J. Eklund, B. Ekström, & G. Nelhans, On the potential for detecting scientific issues and controversies on Twitter: A method for investigation conversations mentioning research. In *Proceedings of ISSI 2019*, 2189-2198, article-id 375, 2019
- [6] J. Eklund, & G. Nelhans, Probabilistic explorations of citation contexts: Citation roles and subject content of scientific references. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators, STI 2022* (sti22224), 2022.
- [7] A. Foderaro, & D. G. Lorentzen, Argumentative patterns and practices in debating climate change on Twitter. *Aslib Journal of Information Management*, 75(1), 131-148, 2023.

Humanistic AI: Towards a new field of interdisciplinary expertise and research

Mats Fridlund^{1,2}, David Alfter^{1,3}, Daniel Brodén^{1,2}, Ashely Green^{1,4}, Aram Karimi^{1,3}, and Cecilia Lindhé^{1,2}

¹ Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg, Renströmsgatan 6, Gothenburg, 412 55, Sweden

² Department of Literature, History of Ideas and Religion, University of Gothenburg, Renströmsgatan 6, Gothenburg, 412 55, Sweden

³ Department of Philosophy, Linguistics, and Theory of Science, University of Gothenburg, Renströmsgatan 6, Gothenburg, 412 55, Sweden

⁴ Department of Historical Studies, University of Gothenburg, Renströmsgatan 6, Gothenburg, 412 55, Sweden

Abstract

The Gothenburg Research Infrastructure in Digital Humanities (GRIDH) have participated in projects within various humanities fields that utilise as well as develop research tools and infrastructural resources that incorporate applications of ‘artificial intelligence’ (AI). These applications can include natural language processing, machine learning, computer vision, large language models, image recognition algorithms, classification, clustering, and deep learning. This paper advances the term ‘humanistic AI’ to describe an emergent form of interdisciplinary practice that uses and develops AI-based research applications to answer humanities research questions together with its entangled humanistic reflection. We coin this term to make implicit and visible the epistemological and material particularities of its practice and the new forms of knowledge its affordances make possible. The paper presents GRIDH projects within ‘humanistic AI’ together with its developed AI resources and applications.

Keywords

Research infrastructure, interdisciplinarity, critical digital humanities, artificial intelligence

1. Introduction

The recent surge in interest in the academic impact of ChatGPT and other applications of ‘artificial intelligence’ or ‘AI’, mainly overlook how humanities researchers have long been using and developing AI. Most prominently within humanities disciplines such as corpus linguistics and language technology, but also in digital humanities and traditional disciplines such as archaeology, comparative literature, and history. In the latter cases, this use is often less prominent in that it tends to be embedded in language technology tools and algorithms, such as topic modelling and word embeddings. With the expanding use of digital methods together with a rising critique within the humanities against the unreflective (dare we say, naive) use of biased and potentially dangerous AI applications, we propose a conceptualisation of ‘Humanistic AI’, to allow such uses to be discussed in a more structured and nuanced manner.

This paper is a first tentative attempt to develop Humanistic AI as a concept describing an emerging field of interdisciplinary research and expertise. We explain what we mean by Humanistic AI and lay out its main practical and conceptual dimensions, followed by describing the involvement of the Gothenburg Research Infrastructure in Digital Humanities (GRIDH, formerly Centre for Digital Humanities) in projects utilising, developing or interrogating AI. We conclude by discussing the concept’s usefulness in wider communications.

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

✉ mats.fridlund@gu.se (M. Fridlund); david.alfter@gu.se (D. Alfter); daniel.broden@gu.se (D. Brodén); ashely.green@gu.se (A. Green); aram.karimi@gu.se (A. Karimi); cecila.lindhe@gu.se (C. Lindhé)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Conceptualising Humanistic AI

What do we mean by ‘Humanistic AI’ and what are its different elements and dimensions? In short, the term refers to activities within humanities research and cultural heritage, that use, develop or study AI tools and applications. Below we briefly describe what we mean by ‘humanistic’ and ‘AI’, followed by a discussion of the emergence of the term and its different meanings within various AI fields and how we position ourselves in using the concept. Finally, we describe the three main elements of Humanistic AI as we understand the concept and its practice within the humanities.

2.1. Meanings of ‘Humanistic’ and ‘AI’

Within AI, ‘humanistic’ can be used to designate ‘humane’ or ‘human-like’ functionalities and behaviours as well as to describe aspects related to humanities disciplines or knowledge domains. For us, the term ‘humanistic’ in ‘Humanistic AI’, refers to the latter sense, to designate an activity or research-directed project within cultural heritage and the humanities. Notably, among historians, there is a consensus that ‘the humanities’ consists of a complex of socially and historically constructed academic disciplines and practices perceived as distinct and yet under continuous renegotiation [1]. For instance, in Sweden many humanistic disciplines move across different university faculties, and before the 1960s the faculty of humanities represented both the humanities and the social sciences [2].

Secondly, the meaning of ‘AI’ stands for artificial intelligence whose meaning is somewhat more problematic due to its increasingly widening meanings within its different academic and public contexts. To clarify and critique the various uses and abuses of the AI term a number of alternative and differentiating terms have been introduced such as ‘augmented intelligence’, ‘intelligence augmentation’, ‘automated approaches’, ‘autonomous systems’ and ‘intelligent systems. The “classic” AI textbook describes it as a field “concerned with not just understanding but also building intelligent entities—machines that can compute how to act effectively and safely in a wide variety of novel situations”, encompassing “logic, probability, and continuous mathematics; perception, reasoning, learning, and action; fairness, trust, social good, and safety; and applications that range from microelectronic devices to robotic planetary explorers to online services with billions of users.” [3]. When thinking about AI, many people without advanced technical expertise would imagine autonomous robots such as seen in TV shows, fearsome “creatures” that surpass human intelligence. While this certainly can be referred to as AI, our approach is more grounded in using it to refer to a *field that develops and studies intelligent machines, as well as to those algorithms and machines themselves*. In particular this refer to machine and software applications from the AI subfields of Expert Systems, Machine Learning (ML), Natural Language Processing (NLP), Speech Recognition, Computer Vision, Robotics, and Genetic Algorithms, which in themselves includes specific applications such as clustering, deep learning, image segmentation, text classification and topic modelling.

Despite these many alternatives to AI, we have, aware of the diversity and range of meanings discussed above, chosen to retain the use of the AI term as it encompasses many significant applications used within the humanistic domain. Furthermore, this is also due to the emerging and increasing contemporary use of AI connected to various purportedly ‘humanistic’ purposes which is something we see ourselves as well positioned to engage with.

2.2. Emergence of Humanistic AI

The last two decades have seen the term ‘Humanistic AI’ being used in different ways within the AI domain that partly overlaps our use. For instance, in 2003 the term ‘humanistic AI’ was used to describe one main trajectory within the design of intelligent machines, trying to emulate human cognitive capabilities, rather than mimicking the anatomical functioning of the human brain [4]. More recently, the term ‘Human-Centered AI’ (HAI) has also been used for similar AI activities and processes. Not rarely, such efforts are shaped by a rationale implying that HAI is not just efficient but also more fair, compatible, and ‘humane’, in augmenting rather than ‘replacing’ human decision-making. Furthermore, there are a range of AI development activities similarly drawing on

HSS perspectives, described in terms such as ‘Responsible AI’, ‘Ethical AI’, ‘Fair AI’, ‘Trustworthy AI’ and, at times, ‘Humanistic AI’ [5].

One example is the Media Lab at Swedish KTH Royal Institute of Technology that see the humanities provide “a critical perspective” as well as “a source of innovation in AI” and engages in interdisciplinary research combining “advanced engineering with philosophy, art, aesthetics and other disciplines from the humanities” to “develop a strong humanistic stance with respect to AI to avoid a situation of technological intelligence overrunning humanism” [6]. Another example is University of Bologna’s Humanistic AI unit that similarly describes Humanistic AI as a “novel branch” reframing the study of “the embodied human mind and social and cultural contexts, as well as their reciprocal relations”, applying AI techniques to humanities that “range from the classification, exploration, management, and preservation of cultural heritage, archives, or demo-ethno-anthropological materials” [7]. Our conceptualisation aligns with these efforts insofar as it concerns the application of AI to humanities rather than the design of HAI as well as developing digital resources in an interdisciplinary context augmented by the reflective and critical faculties of humanities scholars. It should also be noted that a similar but opposite trend is recently surging in popularity within AI, namely ‘human-in-the-loop architectures’, i.e. machine learning architectures where human knowledge is provided before or during the training phase in order to overcome the limits of modern AI [8].

2.3. Elements of Humanistic AI

AI is involved in humanistic research through three main areas of practice: humanistic researchers using various existing tools incorporating AI applications; the use of AI tools and knowledge to develop custom-made resources for humanities researchers; and humanists’ interrogating AI through analysis and reflection on the impact of the AI tools’ embedded positions (‘biases’) and affordances. All of this, either by individual humanistic researchers or as interdisciplinary collaborations. On one level, Humanistic AI can be understood in relation to other concepts, such as, Critical Digital Humanities [9] and Critical Code Studies [10], in that it concerns critical, interdisciplinary oriented reflection on sophisticated digital methods, software, tools, etc., as well as the socio-cultural production of knowledge in “digitalised” society. On another level, our notion of Humanistic AI is more of a tentative framework constructed around the somewhat contested concepts of ‘DH’ and ‘AI’ rather than an agreement on approaches and objects. Thus, we delineate a wide range of activities united by a heterogeneous aim to explore AI within the domains of the humanities.

2.3.1. Using AI

Applications of AI used in DH projects involve a range of diverse techniques and methods which includes vector representation for text, contextual search, data annotation, clustering, image classification, and recognition. Specific examples of applications implemented by GRIDH include advanced word embeddings (such as Word2Vec or FastText) to create vector representations of textual content allowing for semantic similarity analysis, topic modelling, and contextual understanding; word embeddings in combination with domain-specific ontologies to enhance the semantic understanding; topic modelling techniques, such as Dynamic Topic Modeling (DTM) or Term Frequency - Inverse Document Frequency (TF-IDF), to capture evolving themes and topics in historical text; semantic search to clarify meaning of queries and documents and to improve search recall precision; and image colour clustering based on similarity of embeddings and calculation of ‘nearest neighbours’.

2.3.2. Developing AI

DH research often involves complex issues not easily solvable by simply applying existing AI applications designed for general purposes and tasks. Thus, in contrast to simply *using* AI, DH projects may also set out to *develop* AI applications for their specific purposes. This can be done in different ways, such as training classifiers, fine-tuning existing or training new transformer models from scratch based on specific text or image corpora. Such GRIDH applications include computer vision and deep learning techniques for automatic image annotation, object detection and segmentation for image

labelling. However, developing more general AI applications requires a deeper understanding of the underlying principles (and implications), and large amounts of training data, a constraint often hard to satisfy in the humanities (e.g. the documents to be analysed are in extinct languages, or the artefacts under scrutiny no longer exists).

2.3.3. Interrogating AI

The last element entails applying humanistic research-based reflection and critique to interrogate the implications of the AI tools and methodologies used. The core of humanities scholarship concerns applying hermeneutics, (source) criticism and reflection concerning methods, tools and data used in producing humanistic knowledge, including potential societal impacts. Also, some humanities disciplines specialise in reflecting on the development, use and impact of digital technologies, such as digital humanities, information science, media studies, practical philosophy, and science and technology studies (STS). Such AI reflexivity at times comes as explicit interdisciplinary studies including humanities scholars, exemplified by an archeological study stating that “the outcome of any AI approach and its interpretation will differ,” enabling “us to reflect on the potential limitations of our digital technology to avoid taking its results as answers to our research questions about human beliefs, ideologies, and creativity.” [11] However, such interdisciplinary interrogation also comes tacitly in project conversations with humanist scholars probing the interpretative limits and affordances of the data generated by AI tools. This often entails making obtuse AI algorithms fathomable, as humanist researchers, to quote a digital humanist, “can never afford to treat algorithms as black boxes that generate mysterious authority” and to use them, “we have to crack them open and find out how they work.” [12]. Or at least try, as working out their inner workings can require elaborate analysis of models and outputs, at times involving separate projects analysing model performance and training that incorporates human-in-the-loop components or active learning.

3. Humanistic AI projects at GRIDH

The Humanistic AI projects at GRIDH focus on one of the three elements above or combine different forms of them and separates into two categories: text-based and multimodal AI projects. They somewhat overlap as text-based projects often include analyses or manipulation of images in the form of digital image files of text documents that are OCRed or not.

3.1. Text-based AI projects

3.1.1. The *Nordisk Familjebok* tool

A research infrastructure project initiated by GRIDH and developed together with Data as Impact Lab at the University of Borås, that created a open digital resource (*nordiskfamiljebok.dh.gu.se*) of the two first editions (published 1876–99 and 1904–26 respectively) of the encyclopedia *Nordisk Familjebok* (NF), a standard reference work for studying 19th and 20th century Swedish society. The scholarly use of NF was augmented by implementation of advanced ‘likeness’ search functionalities using a Word2vec-model based on the KB-BERT large language model of the KBLab of the National Library of Sweden, making it an AI use as well as AI development project.

3.1.2. The New Order of Criticism project

The project’s comprehensive approach aims to provide rich insights into how readers perceive and engage with books in the Swedish language, facilitating the development of user-centric applications like personalised book recommendations. To accomplish this large, pre-trained Swedish language models like BERT and ELMo are leveraged to conduct sentiment analysis and classification of newspaper book reviews. The project employs fine-tuning techniques to tailor these models to specific tasks, ensuring high performance. It goes beyond simple sentiment polarity analysis by implementing

aspect-based sentiment analysis, entity recognition, and emotion detection. The research in the project also includes advanced visualisation methods for presenting findings and addressing ethical considerations in AI.

3.2. Multimodal AI projects

In several GRIDH projects AI is used to study, analyse and manipulate non-text data, most often digital images and at times digital audio and video, and sometimes also associated with geospatial data. Thus, when we talk about multimodal projects, we mean it as a description of our non-text focused AI projects using one or several other data modes than text. Often these projects also involve analysis of text data. In some cases, these projects are ‘genuinely’ multimodal, where data of multiple types are combined for analysis to add additional context and at times also include analysis of multiple data types in parallel.

3.2.1. Projects using image clustering

GRIDH are using ML algorithms and developing interactive visualisations for image clustering of several types of image content. In the literary ‘lab’ developed for Litteraturbanken (LB), GRIDH use machine learning algorithms to cluster images of illustrations, initials, graphics ornaments, and sheet music extracted from the LB’s repository of 19th century works using object detection. The aim is to enhance the visualisation of literary reuse and similarity, as well as provide future researchers with easy data access. The *Ivar Aroseniusarkivet* (Ivar Arosenius Archive) project uses methods and concepts developed by Douglas Duhaime and the Yale DHLab [13], as well as the Nasjonalmuseet in Norway [14] to visualise the archive’s images of the artwork. The TSNE projections of RGB images and a gallery of each image’s nearest neighbours are displayed in an interactive frontend. GRIDH aims to use additional clustering algorithms and improved interactive visualisation to increase the archive’s accessibility to the public and researchers.

3.2.2. Projects using Augmented Reality

In the project ‘Rock Art in Three Dimensions’, an application was developed using two different Augmented Reality (AR) technologies; markerless image detection trained on natural features where a device tracks its position through image recognition of natural features, and plane tracking which recognises horizontal and vertical surfaces using the technique Visual Inertial Odometry (VIO). Markerless image detection made possible to detect and add contextual information to rock carvings without any physical additions to the rock art site. Thus, the interpretations could be served in a digital form while the physical environment was kept untouched, thereby aligning itself with the values of conservation [15].

3.2.3. Projects using automatic speech recognition

One frequently mentioned AI application is automatic speech recognition (ASR) which serves as the foundation of the project ‘Terrorism in Swedish politics (SweTerror)’ that studies parliamentary speech on terrorism 1968–2018. The speech technology trains and adapts deep neural networks to better cope with speaker biases, especially gender, and train, compare and inspect models trained on speech to detect historical changes in parliamentary speech. Also, the project’s text analysis relies on state-of-the-art LT methods using AI, such as word pictures, topic modelling and word vectors. [16].

4. Conclusions

In this paper, we have tentatively suggested Humanistic AI as an apt concept for discussing an emergent field in the intersection of the application of AI tools and the interests that fall within the domain of critical digital humanities and adjacent fields of the humanities. By addressing the core elements of what could be considered Humanistic AI and concretising it by presenting some projects involving

GRIDH, we have sought to demystify the notion of applying AI within the humanities. In a way, we have tried to show that humanists are already active within AI, sometimes without realising the depth, degree, or character of their involvement. In writing this paper, we have debated the ambiguous and partly contested concept of AI, and in this come to terms with the extent to which the notion of Humanistic AI can be useful in describing the work at GRIDH concerning the development of resources and infrastructure with elements of AI. This usefulness becomes apparent, not least when communicating what we “do” to external partners and fellow humanist researchers as well as other academic and non-academic stakeholders, interested in the applications of AI to the humanities.

Acknowledgements

We are very grateful for the support from and collaboration with our research and development partners and staff at CDH and GRIDH who made the projects discussed above possible, in particular Johan Eklund, Christian Horn, Johan Ling, Arild Matsson, Gustaf Nelhans, Rich Potter, and Victor Wählstrand Skärström.

References

- [1] R. Bon. *A new history of the Humanities*, Oxford University Press, Oxford, 2013.
- [2] A. Ekström & H. Östh Gustafsson (Eds.). *The humanities and the modern politics of knowledge*, Amsterdam University Press, Amsterdam, 2022.
- [3] Russell, Stuart J., and Peter Norvig. *Artificial intelligence: a modern approach*. Harlow, 2021. 4th ed, 7, 19.
- [4] Krishnakumar, Kalmanje. "Intelligent Systems for Aerospace Engineering: An Overview." *Von Karman Institute Lecture Series on Intelligent Systems for Aeronautics* (2002).
- [5] Saheb, Tahereh, Sudha Jamthe, and Tayebbeh Saheb. "Developing a conceptual framework for identifying the ethical repercussions of artificial intelligence: A mixed method analysis." *Journal of AI, Robotics & Workplace Automation* 1.4 (2022): 371-398.
- [6] <https://www.kth.se/hct/mid/research/media-lab/about-1.929121>
- [7] <https://centri.unibo.it/alma-ai/en/scientific-units/humanistic-ai>
- [8] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). "A survey of human-in-the-loop for machine learning". *Future Generation Computer Systems*, 135, 364-381.
- [9] D. Berry and A. Fagerjord. *Digital humanities: Knowledge and critique in a digital age*, Polity, London, 2017.
- [10] Mark C. Marino, *Critical Code studies*, MIT press, Cambridge, Massachusetts, 2020.
- [11] Horn, C., Ivarsson, O., Lindhé, C., Potter, R., Green, A., & Ling, J. (2022). "Artificial intelligence, 3D documentation, and rock art: Approaching and reflecting on the automation of identification and classification of rock art images." *Journal of Archaeological Method and Theory*, 29, 188–213.
- [12] T. Underwood. "Theorising research practices that we forgot to theorize twenty years ago", *Representations*, 127:1 (2014): 64–72.
- [13] <https://douglasduhaime.com/posts/identifying-similar-images-with-tensorflow.html>.
- [14] <https://www.nasjonalmuseet.no/en/about-the-national-museum/collection-management---behind-the-scenes/digital-collection-management/project-principal-components/>
- [15] Westin, J., Råmark, A. & Horn, C. (2023). "Augmenting the Stone: Rock Art and Augmented Reality in a Nordic Climate". *Conservation and Management of Archaeological Sites*.
- [16] J. Edlund, D. Brodén, M. Fridlund, C. Lindhé, L-J. Olsson, M. Ängsal and P. Öhberg, "A multimodal digital humanities study of terrorism in Swedish politics: An interdisciplinary mixed methods project on the configuration of terrorism in parliamentary debates, legislation, and policy networks 1968–2018", in: K Arai (ed.): *Intelligent Systems and Applications: Proceedings of the Intelligent Systems Conference (IntelliSys) 2021*, 2, Springer, Cham, 2022, pp. 435–449.

Designing digitally-driven integrative interdisciplinarity: Professionalism between protocol and judgement

Daniel Brodén^{1,2}, Mats Fridlund^{1,2} and Cecilia Lindhé^{1,2}

¹ Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg, Renströmsgatan 6, Gothenburg, 40530, Sweden

² Department of Literature, History of Ideas and Religion, University of Gothenburg, Renströmsgatan 6, Gothenburg, 40530, Sweden

Abstract

While there is a growing discussion of the importance of developing collaborative workflows for interdisciplinary research within DH, there is a lack of blueprints and consideration of specific expertise. This paper conceptualizes the practice of what we tentatively call *digitally-driven integrative interdisciplinary project design* in order to highlight a certain professional practice for integrating collaboration between technical expertise and traditional HSS researchers when developing research project applications, digital resources, etc. We begin by highlighting the need for *protocol* for workflow-oriented approaches to integrative interdisciplinary collaboration, but also an embodied expertise in need of being put into focus in discussions of integrative workflows within digital humanities. Then, we argue that *judgement* is also a crucial but often overlooked part of the professionalism involved. We conclude by discussing how to further develop the conceptualization of interdisciplinary digital project design and the expertise involved.

Keywords

Interdisciplinary research, research infrastructure, critical digital humanities

1. Introduction

It is well-known that although Digital Humanities (DH) projects combine humanistic and technical expertise in different ways, interdisciplinary collaboration can be challenging due to how different disciplinary rationales dovetail. It has been a worn truism that while computer scientists tend to be interested in pushing methodological development, researchers in the Humanities and Social Sciences (HSS) primarily seek to apply disciplinary methods to digital scholarship. However, it has also been argued that DH has given rise to a ‘third culture’, as imagined by C.P. Snow [1], where people from two essentially different disciplinary cultures take strides towards collaborating and aligning perspectives [2]. There is also an emergent DH discussion around developing and conceptualizing so-called integrative interdisciplinary workflows. For instance, the concept of “agile hermeneutics” [3] has been used to designate a collaborative process in which HSS scholars and data analysts are engaged in a constant dialogical and reflexive relationship [4]. Nevertheless, there still remains a palpable lack of blueprints and considerations of ‘embodied’ expertise for developing interdisciplinary DH projects. As Ahnert et al. [5] note in a recent study, “new projects and initiatives expend a lot of energy in their start-up period trying to establish collaborative values and project management strategies, often reinventing the wheel in the process”.

Partly in response to Underwood’s [6] call for the need within DH to make explicit, name and reflect upon our partly tacit working methods, the present paper highlights our conceptual work on the practice of *designing digitally-driven integrative interdisciplinary projects*, designating a certain approach to outlining and integrating collaboration between technical expertise and ‘traditional’ HSS researchers when developing, among other things, project applications and digital resources (tools, databases, etc.), involving multidisciplinary teamwork. Etymologically, the word *design* comes from the Latin word *designere*, meaning, among other things, to ‘mark out’, ‘point out’ or ‘devise’. The Italian verb

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

 daniel.brodén@gu.se (D. Brodén); mats.fridlund@gu.se (M. Fridlund); cecila.lindhe@gu.se (C. Lindhé)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

to do with both the spheres of ‘planning’ and ‘doing’ [7]. Nowadays, the term ‘design thinking’ is used to refer to a set of procedures in the process of designing and project organization in a general sense as well as to the body of knowledge developed about design ability [8] [9].

Against this background, we propose the notion of digitally-driven integrative interdisciplinary project design to put focus on an emergent form of expertise related to the processes of what has been called the fractured trading zones of DH [10]. On some level, this expertise has something in common with what Hunter labels ‘bridge people’, who are either fully competent in the ‘two cultures’ or at least ‘bilingual’ in that they can speak the languages of the two cultures [2]. Our paper concerns the latter of these two categories, directing attention to a ‘professionalism’ aimed at outlining and integrating collaboration between technical expertise and HSS researchers. On another level, this expertise can be described in terms of critical digital humanities ‘in practice’ in the sense that it entails a continuous reflection and dialogue on the computational methods, tools, databases, etc, as well as a concern for the socio-cultural organization, production and dissemination of knowledge [11]. To be clear, rather than discussing skillsets or best practices for project design, *we seek to conceptualize core aspects of a certain interdisciplinary-oriented professionalism that we argue is in need of being put into focus* in discussions of workflows for integrative interdisciplinary teamwork within DH. In this, we draw upon the experiences since 2015 as senior staff at the Gothenburg Research Infrastructure in Digital Humanities (GRIDH, formerly the Centre for Digital Humanities) at the Faculty of Humanities at the University of Gothenburg of initiating and designing research project applications and digital resources in collaboration with HSS researchers as well as working with the e-infrastructures Huminfra and Swe-Clarín.

1.1. Disposition

We begin our conceptualization of digitally-driven integrative interdisciplinary project design by discussing what might be called *protocol*. Engaging with Oberbichler et als. [12] discussion of an integrative approach to multidisciplinary teamwork centered around historical data, we highlight the need for structuring collaborative workflows, but also for grounding the application of protocol in embodied expertise. Then, drawing upon philosopher of education Gert Biesta’s writing about *judgement* [13] as a crucial but often overlooked part of professionalism, we argue that the exercise of judgement is also a critical part of digitally-driven integrative interdisciplinary project design. To concretize our discussion, we will briefly comment on some phases in the design of an ongoing mixed-methods project that brings together DH and HSS researchers, discussing how the integrative professionalism involved requires a dialectical relationship between protocol and judgement. We conclude by making some comments about how to further develop the conceptualization of digitally-driven integrative interdisciplinary project design and the professionalism involved.

2. Integrated interdisciplinary research and protocols

Many commentators on DH have argued for the need of robust results based on not only thorough engagement with data from a computational view, but also collaborative research grounded in interdisciplinary and mixed-methods approaches [11] [14]. Against this background, Oberbichler et al. [12] make a distinction between ‘multidisciplinary collaboration’ and ‘integrated interdisciplinary research’ within DH in the sense that while the former tries “to build something in between the disciplines so they share more than just the problem”, the latter mean that “people from different scientific fields come together, collaborate, and study a common question or problem with the goal of reaching common conclusions”. Here, one should, of course, keep in mind the long-running and often contradictory debates within academia about the specific meaning of terms such as multi- and interdisciplinarity. However, rather than the specific choice of terminology, most important in our context is the character and the quality of these collaborative processes. Oberbichler et al. argue that integrated interdisciplinary research “requires going deeper than just saying something about a phenomenon from different perspectives” and this also includes “the understanding of how each field works and which approaches are used for problem-solving” [12]. “In order to provide a process for successful collaboration and communication, the differences and commonalities between disciplines

need to be considered. Merging of applications, tasks, and traditions, involving mixed method approaches as well as increased interaction between the disciplines, has been identified as a possible common objective” [12], they write.

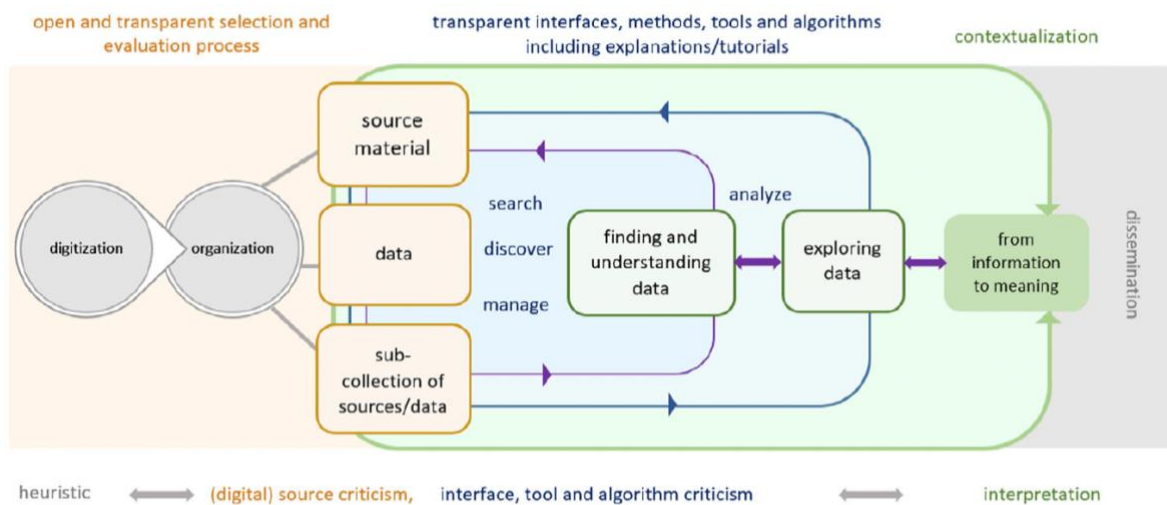


Figure 1: Interdisciplinary digital hermeneutics workflow by Oberbichler et al. [12].

To achieve an integrated interdisciplinary research process, Oberbichler et al. propose a workflow (Figure 1) that emphasizes the importance of iterative collaborative steps between data analysts and HSS researchers to gradually better understand, explore and analyze the data, methods and tools in focus. In our work developing a range of project applications and digital resources at GRIDH and the e-infrastructures Huminfra and Swe-Clarín, we have found that an integrative workflow following iterative steps centred on the data in focus roughly in line with the one proposed by Oberbichler et al. works well for facilitating a dialogue between the parties involved to gain deeper insights into the data and each other’s concerns. However, our point is not to argue for their model as an ‘ideal’ or ‘all-purpose’ tool. For instance, one could discuss if not the category digitization should also be included in the iterative segment of the workflow (Figure 1), since this is hardly a static and fixed process [15]. Rather, the point we want to make is that following some kind of workflow and protocol is key in the practice of digitally-driven and integrative interdisciplinary project design, as it provides a collaborative framework for finding common ground and shared understanding of the different research perspectives and tasks involved.

However, before proceeding, we need to think about the somewhat overlooked issue of integrative expertise. Oberbichler et al. [12] propose that their model could serve as a starting point for planning future projects, but they describe it “as a concept rather than a concrete architecture”. Thus, while highlighting a set of protocols as well as appropriate interdisciplinary skills – what they call an “art of in-betweenness” – notably, these elements are ‘free-floating’ rather than ‘embodied’ in the sense that they are not associated with any distinct functions or roles in the model. Here, we could return to Hunter’s notion of bridge people [2] who in the context discussed by Oberbichler et al. could be understood as people in some way proficient in in-betweenness. Thus, from the perspective of digitally-driven integrated interdisciplinary project design, it seems reasonable to argue that much time and energy could be saved by making sure that every project in some way involves bridge people. However, there is something ‘static’ with the notion of bridge people in the sense that it seems implied that they somehow naturally manifest their competence in the two cultures, and we would like to rather stress the point of understanding both the application of protocol and in-betweenness dynamically, as a ‘doing’ rather than a ‘being’.

3. Professionalism and judgement

To get a better understanding of the application of protocol and the art of in-betweenness in the context of mediating different disciplinary rationales and facilitating collaboration in DH projects, we can turn

to a notion of educational philosopher Biesta [12] to understand the process as enacting a certain acquired professionalism. Arguing against a general tendency to instrumentalize teaching as the “application of protocols”, Biesta [13] points to the general misconception that protocols and guidelines can tell teachers “what they should do on the assumption that particular forms of research can provide clear and unambiguous knowledge about ‘what works’”. Rather, Biesta stresses the crucial importance of judgement, in that teachers as professionals bring something critical to the educational situation as all desirable education requires instant and spontaneous judgement to be made about what to do, both on a general level and in concrete situations.

Transferred to our context, we argue that digitally-driven integrated interdisciplinary project design can only be organized so far according to protocol or, for that matter, according to the idea that bridge people somehow naturally mediate convergence between the two cultures and implement an integrative workflow. Rather, the process will, to some extent, depend more on a specific professionalism being enacted. Similar to how Biesta conceives teaching as an act of judgement, we suggest that it is productive to understand the application of protocol and in-betweenness as based on acts of judgement. Here, it might be useful to remember a ‘classic’ definition of professionalism as an approach tied to sets of specialized and exclusive knowledge and skills [13] [16]. Thus, similar to what Biesta writes about education, bridge people will constantly be “confronted with situations that, in some respects, are always new and hence call for judgement” [13]. Furthermore, we should keep in mind that professionalism is also tied to relationships of knowledge-based authority and trust. Authority can, of course, be questioned with regards to the relationships between professionals and their ‘clients’, but it nevertheless provides a justification for judgements made. In the context of an integrated interdisciplinary research process, there is arguably judgement to be made about, among other things, the balance between the different steps in a workflow, the identification of aligned goals and what needs to be done (and not done) at a certain point. As Biesta [13] notes, the whole point of professional practices “is that they do not just service the needs of their clients, but also play a crucial role in the definition of those needs”.

4. Enacting protocol and judgement: The SweTerror project

We will now turn to a brief case study to concretize our discussion and discuss our enactment of protocol and judgement in two specific phases in the digitally-driven integrated interdisciplinary research project ‘Terrorism in Swedish politics: A multimodal study of the configuration of terrorism in parliamentary debates, legislation, and policy networks in Sweden 1968–2018’ (SweTerror, 2021–2026) [16]. First, we will comment on some elements of the design of the project application and, then, the re-design of research questions during the ongoing research process.

4.1. Designing an interdisciplinary project team

The SweTerror project originated from a call in 2020 from the DIGARV (Digitisation and Accessibility to Cultural Heritage Collections) research program for data-driven HSS research based on digitized material available in Swedish GLAM institutions. Drawing together an interdisciplinary team from, among other things, digital humanities, terrorism studies, history, linguistics, political science, language and speech technology, the SweTerror project was designed as a comprehensive mixed methods study of the political discourse on terrorism, as represented by both the text transcripts and audio recordings of Swedish parliamentary debates. In the assessment, the Swedish Research Council (Vetenskapsrådet, VR) lauded the composition of the research team, stating that the SweTerror project “creates a real opportunity to change the nature of scholarship on the politics of terror” by bringing together a “remarkably interdisciplinary team of researchers” [18].

The process behind the project application began with the core project partners (DH scholars Brodén and Fridlund, and language technology scholar Edlund) jointly drawing up the overarching scope of the project in relation to the digitized material in focus, that is the digitized records from the Swedish Parliament (both speech and text), partly in line with the proposed workflow in Figure 1 that takes the data as its point of departure for dialogue and organizing. However, in the context of the practice of digitally-driven integrated interdisciplinary project design, it is also clear that the enactment of protocol

and judgement tended to overlap even at the early stage of creating the application. For instance, in the enrollment of researchers, we (Brodén and Fridlund) had to more or less rely on judgement alone when tasked with putting together an interdisciplinary project team with broad and complementary domain and methodological expertise. Essentially, our bringing together a team of HSS researchers and data analysts had more to do with a series of judgements about what kind of research the data enabled and the specific expertise required, rather than application of any distinct interdisciplinary protocol.

4.2. Designing interdisciplinary research questions

A key part of the writing of the project application was the collaborative, iterative negotiation of the research questions, with the questions gradually emerging from a critical dialogue within the team. Among other things, the research questions were as much the result of a negotiation between the analytical interests of the HSS researchers as of the data analysts' view on what analysis the data would enable. In this process, our experience as bridge people played an important part, insofar as we were able to acknowledge the different disciplinary perspectives and also, to some extent, mediate between the 'two cultures'. Without going into details about our individual backgrounds, we should note that we as senior researchers in the humanities share not only substantial experiences of designing externally-funded interdisciplinary projects and conducting research in different academic contexts, but also a certain approach toward collaboration in the sense that we emphasize the importance of problem solving over disciplinary interests. Since most team members had not collaborated before, as bridge people with certain experiences and sensibilities we also sought to build in some kind of a relationship of trust in our judgement that, for instance, played a part in the writing of research questions when deciding trade-offs, the use (or not) of domain specific concepts, exact formulations, etc.

However, after the project was approved and started, the team also had to, to some extent, re-design and re-negotiate the collaborative tasks necessary for answering the research questions, not least in the process of co-writing. For instance, in writing conference papers, the HSS researchers' work very much rested on what specific data the data analysts could extract from the parliamentary corpora at that particular time, since these are in a continuous flux subject to ongoing curation [19]. This called for facilitating a interdisciplinary dialogue within the team, but also judgements about how much the data needed to be curated, refined and improved in order to allow valid conclusions to be drawn about the parliamentary discourse on terrorism based on different disciplinary standards and concerns. Furthermore, this included potential 're-calibration' of the research questions in relation to the specific data available.

5. Conclusions

In this paper we have sought to conceptualize core aspects of an interdisciplinary-oriented expertise that is in need of being put into focus in discussions of collaborative workflows within DH. We have proposed the practice of digitally-driven integrative interdisciplinary project design to highlight a practice aimed at organically integrating collaboration between technical expertise and traditional HSS researchers when developing project applications, digital resources, etc. By discussing the need for enactment of both protocol and judgement we have delineated crucial and complementary distinctions concerning the professionalism involved. To concretize the argument, we have discussed two phases in the design of the SweTerror project that illustrates how this practice is grounded in an overlapping, dialectical relationship between protocol and judgement.

However, while we have addressed the core elements of protocol and judgement, the practice of digitally-driven integrative interdisciplinary project design needs to be developed further and in more detail. Among other things, since the notion of bridge people concerns both those who are either fully competent in the two cultures or at least bilingual, further elaboration is required on how these different categories relate to the workflows and professionalism discussed here. Likely, the notion of seniority and experience also feeds into this context and needs to be further taken into account on some level. Furthermore, in the paper we have used the terms 'technical expertise' and 'data analysts' in a somewhat monolithic way, whereas in practice these categories are often inhabited by people with different

degrees of interdisciplinary experience and expertise. Just as we as humanities researchers through our interdisciplinary experience and expertise are partly bilingual in the two cultures, the same applies to data analysts, who can be partly or fully ‘fluent’ in HSS research concerns. Thus, there is a need to deepen the discussion both about the conceptualization of digitally-driven integrative interdisciplinary project design and about the distribution of the professionalism described in this paper among different individuals and institutions in the field of DH.

Acknowledgements

This paper draws upon the collaborative work carried out with our colleagues at GRIDH and the researcher team in the SweTerror project (<https://sweterror.se>) funded by the research program DIGARV (VR, RJ and the Royal Swedish Academy of Letters, History and Antiquities). The presented results has also emerged from work within the national research infrastructures Huminfra (funded by VR, contract no. 2021-00176) and Swe-Clarin (funded by VR, contract no. 2017-00626).

References

- [1] J. Ingvarsson. *Towards a digital epistemology: Aesthetics and modes of thought in early modernity and the present age*, Palgrave Macmillan, Cham, 2021.
- [2] A. Hunter. "Digital Humanities as third culture, *MedieKultur*", 57 (2014): 18–33.
- [3] G. Rockwell and S. Sinclair. *Hermeneutica: Computer-assisted interpretation in the Humanities*, MIT Press, Boston, Mass., 2016.
- [4] A. Fickers. *Digital hermeneutics in history: Theory and practice*, De Gruyter, Oldenburg, 2019.
- [5] R. Ahnert, E. Griffin, M. Ridge and G. Tolfo. *Collaborative historical research in the age of big data*, Cambridge University Press, Cambridge, 2023.
- [6] T. Underwood. "Theorising research practices that we forgot to theorize twenty years ago, *Representations*", 127:1 (2014): 64–72.
- [7] Svenska Akademiens Ordbok, <https://www.saob.se/>
- [8] P. Rowe. *Design thinking*. MIT Press. Cambridge, Mass., 1987.
- [9] N. Cross. *Design thinking: Understanding how designers think and work*, Berg, London and New York, 2011.
- [10] P. Svensson. "The Digital Humanities as a Humanities Project", in: M Gorman (Eds.) *Trading Zones and Interactional Expertise: Creating New Kinds of Collaboration*, MIT Press, Cambridge, Mass., 2011, pp. 42–60.
- [11] D. Berry and A. Fagerjord. *Digital humanities: Knowledge and critique in a digital age*, Polity, London, 2017.
- [12] S. Oberbichler, E. Boroş, A. Doucet, J. Marjanen, E. Pfanzelter, J. Rautiainen, H. Toivonen and M. Tolonen. "Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians". *Journal of the Association for Information Science and Technology*, 73:2 (2021): 225–239.
- [13] G. Biesta. "What is education for? On good education, teacher judgement and educational professionalism", *European Journal of Education*, 50:1 (2015): 75–87.
- [14] B. Johnson, A. Onwuegbuzie and L. Turner. "Towards a definition of mixed methods research, *Journal of Mixed Methods Research*", 1:2 (2007): 112–133.
- [15] A. Smith and B. Whearty, "All the work you do not see: Labor, digitizers and the foundations of digital humanities", In: M Gold and L Klein (Eds). *Debates in the digital humanities*, University of Minnesota Press, Minneapolis, 2023.
- [16] E. Freidson. *Professionalism reborn: Theory, prophecy, and policy*, University of Chicago Press, Chicago, 1994.
- [17] J. Edlund, D. Brodén, M. Fridlund, C. Lindhé, L-J. Olsson, M. Ängsal and P. Öhberg, "A multimodal digital humanities study of terrorism in Swedish politics: An interdisciplinary mixed methods project on the configuration of terrorism in parliamentary debates, legislation, and policy networks 1968–2018", in: K Arai (ed.): *Intelligent Systems and Applications: Proceedings of the Intelligent Systems Conference (IntelliSys) 2021, 2*, Springer, Cham, 2022, pp. 435–449.

[18] VR assessment 2020–05052.

[19] D. Brodén, M. Fridlund, L-J. Olsson, M. Ängsal and P. Öhberg. "The Diachrony of the New Political Terrorism: Neologisms as Discursive Framing in Swedish Parliamentary Data 1971–2018", *Digital Humanities in the Nordic and Baltic Countries Publications*, 5:1 (2023): 79–89.

From the Arctics to Antarctica - A multimodular visualisation of data

Jonathan Westin¹, Tristan Bridge¹, Matteo Tomasini¹

¹ Gothenburg Research Infrastructure in Digital Humanities, Göteborgs universitet, Renströmsgatan 6, Göteborg, 405 30, Sverige

Abstract

This paper outlines the structure of Multimodal Map, a tool developed at GRIDH to access and visualise place-based datasets. The Multimodal Map frontend, which is developed with a Vue3 framework that fetches data from a backend built in Django, is arranged as a series of distinct and interconnected views that lets the user interact with the material at different scale and level of abstraction. To support the wide variety of formats the different projects need to handle, Multimodal Map makes use of both custom solutions and several open frameworks and libraries. These include Open Layers for the geographical visualisations, OpenSeadragon for IIIF-images, potree.js for point clouds, 3D Heritage Online Presenter (3DHOP) for meshes, and relight-viewer.js for RTI Photography.

Keywords

Research Infrastructure, User Interface, Data model

1. Introduction

The Gothenburg Research Infrastructure in Digital Humanities (GRIDH) have developed a package of user interface modules organised around a data model specifically aimed at spatio-temporal visualisations. The core package, which we refer to as Multimodal Map (henceforth MuM) was first fully developed for the project *Extended Rephotography* where the researchers needed both a system to register data collected in the Arctics and a tool through which to visualise the spatio-temporal relations in the material. The dataset consisted of glacier observations, historical and present photography and rephotography, measurements, and 360-degree video recordings. Since April 2023, MuM has been adapted and developed to accommodate the needs of several subsequent projects, including *Reading the Signs*, *Göteborgs Jubileum 1923*, *Etruscan Chamber Tombs*, *Sonora*, *Stokkastovan*, *The Inscriptions of Saint Sophia*, and *Built Cultural Heritage in Antarctica*. Through these projects, MuM has been expanded with capabilities to view, perform measurements on, and evaluate 3D data, explore reflectance transformation imaging (RTI), browse and filter visual galleries of datasets, and group and sort documentation according to date or type. Hence, rather than offering semantic annotation and structuring or processing of text and images, functions handled much better by mature tools such as Recogito (<https://recogito.pelagios.org>), MuM is defined by a “linear modularity”, a semi-rigid structure that moves from the visual establishment of context to the exploration of digitisations through media-specific tools. The common denominator for the MuM projects is that the data, primarily visual in nature, is organised around an exact geographical position and a moment or event in time, which allows both for spatial exploration and chronological presentation and filtering. However, data-modelling informs the structure of a dataset by establishing hierarchies. In the MuM projects these hierarchies – and therefore the data model – are instigated by the concept of

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.



jonathan.westin@lir.gu.se (J. Westin); tristan.bridge@lir.gu.se (T. Bridge); matteo.tomasini@lir.gu.se (M. Tomasini)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

approaching the material at different scales, from abstract to concrete. Thus, we argue that MuM also offers a conceptual *method* for data registration that guides how a dataset is structured.

The present projects that make use of the MuM frontend and backend span a wide array of categories and geographies, from the Arctics to Antarctica, but all collect and make available datasets related to heritage that have been structured around the concept of place and scale through their adaption to MuM. These datasets include documentation of inscriptions from threatened environments and inscriptions in Ukraine, documentation of Swedish pipe organs, data collections about street signs in Rwanda, historical photographs from Gothenburg, immersive documentation of glaciers and the remains of the polar expeditions, and high-resolution point clouds of Etruscan chamber tombs in Italy and log houses on the Faroe Islands.

2. General description

The MuM frontend, which is developed with a Vue3 framework that fetches data from a backend built in Django, is arranged as a series of distinct and interconnected views that lets the user interact with the material at different scale and level of abstraction; *the map view* (A), *the gallery view* (B), *the place view* (C), and *the object view* (D).

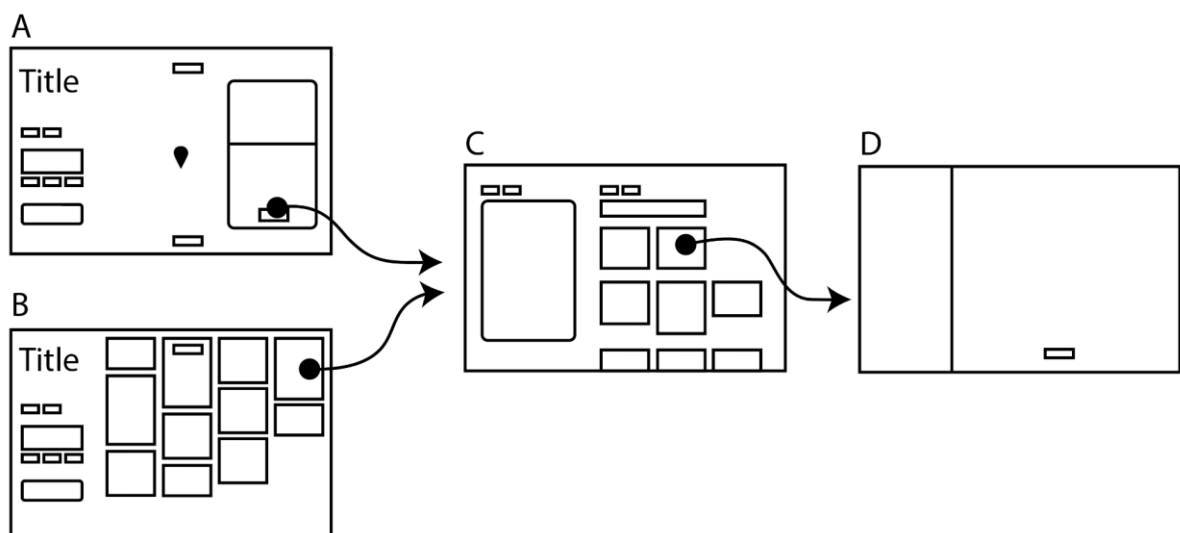


Figure 1: The conceptual connection between the different views. Illustration: J. Westin.

The main interface, *the map view*, is organised around the core component of a graphical representation of a Euclidean space where the data is spatially presented either as interactive markers or as overlays. Depending on the datasets that need to be visualised, this *map view* could be limited to the floor plans of one or several buildings (*Stokkastovan, The Inscriptions of Saint Sophia*) or be expanded to include an entire countryside or the better part of a continent (*Etruscan Chamber Tombs, Sonora, Extended Rephotography*). On the left hand of the screen, the user has access to a set of widgets through which to filter down the displayed data. These can be specified for each individual project and include controls that let the user limit the visible data to a particular dataset or period, view only data from a delimited geographical area or of a certain type, and focus on data that has been associated with a particular tag. As an example, in the *Etruscan Chamber Tombs* project, the user can filter by dataset, type of data (3d models, plans, or all data), time period (ranging from unknown to 400-200 BC), site, necropolis, and type of tomb. Such controls enable a level of data manipulation that is non-hierarchical through the ability to interact with specific sites as well as a more general filtering.

In parallel with the *map view* there is a *gallery view* which offers a graphical representation of the markers. Depending on the project, the gallery is populated with either a single visual representation of each of the places on the map (*Etruscan Chamber Tombs, Sonora, Stokkastovan, and The Inscriptions of Saint Sophia*), or all photographic data (*Extended Rephotography*,

Göteborgs Jubileum 1923). The map and the gallery mirror each other regarding what data sources they display, meaning that if the user filters down the dataset to a certain area on the map or a certain type of place, only images from that area or that type of place will be shown in the gallery.



Figure 2: The map view and gallery view of *Etruscan Chamber Tombs* with filter-widgets on the left.

When a marker is selected, data associated with that place is shown. Presently MuM has three possible interfaces for displaying this information; a place card that overlays the right hand side of the map and assembles available photos from the selected place into a carousel with a preview of the metadata presented below (*Etruscan Chamber Tombs*, *Sonora*, *Stokkastovan*, and *The Inscriptions of Saint Sophia*), a compact scrollable column that overlays the right hand side of the map with place data and previews of associated visual media (*Extended Rephotography* and *Göteborgs Jubileum 1923*), or as an expandable area to the right of the *map* view populated both with previews and metadata (*Reading the Signs*).

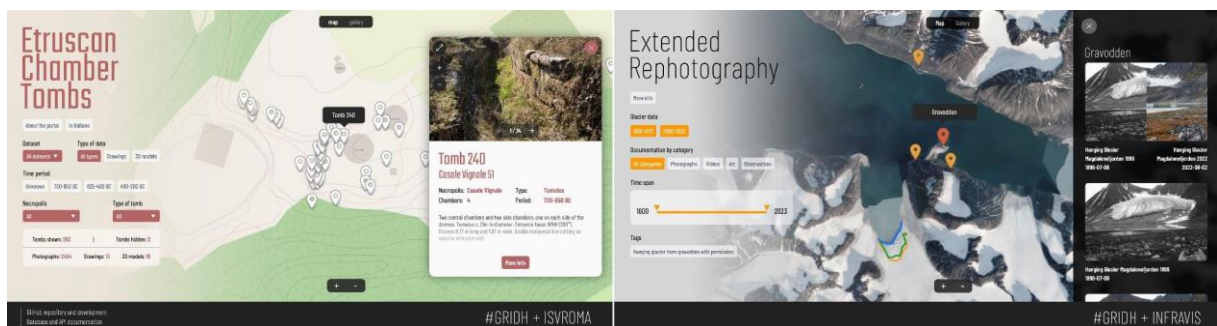


Figure 3: The place card in *Etruscan Chamber Tombs* and the column view from *Extended Rephotography*.

The projects that make use of the place card to preview the data all have an additional view, the *place view*, that the place card and images in the gallery link to. This view collects all the data from a place in an expanded interface that lets the user sort the associated data by type or date and presents a more generous space for descriptions and metadata connected with the place. In order to display and let the users interact with the wide variety of formats the different projects need to support, MuM makes use of both custom solutions and several open frameworks and libraries. These include Open Layers for the geographical visualisations, OpenSeadragon for IIIF-images, potree.js for point clouds, 3D Heritage Online Presenter (3DHOP) for meshes, and relight-viewer.js for RTI Photography. These libraries all come with their own user interfaces and have therefore been redesigned to present a coherent experience for the MuM user. When the user selects a preview image for visual data, the interaction is handed over to either the built in MuM *object view* for images, rephotography, and videos (both standard and 360) or for point clouds, meshes and RTI photographs to an auxiliary web-app built to handle that type of data.

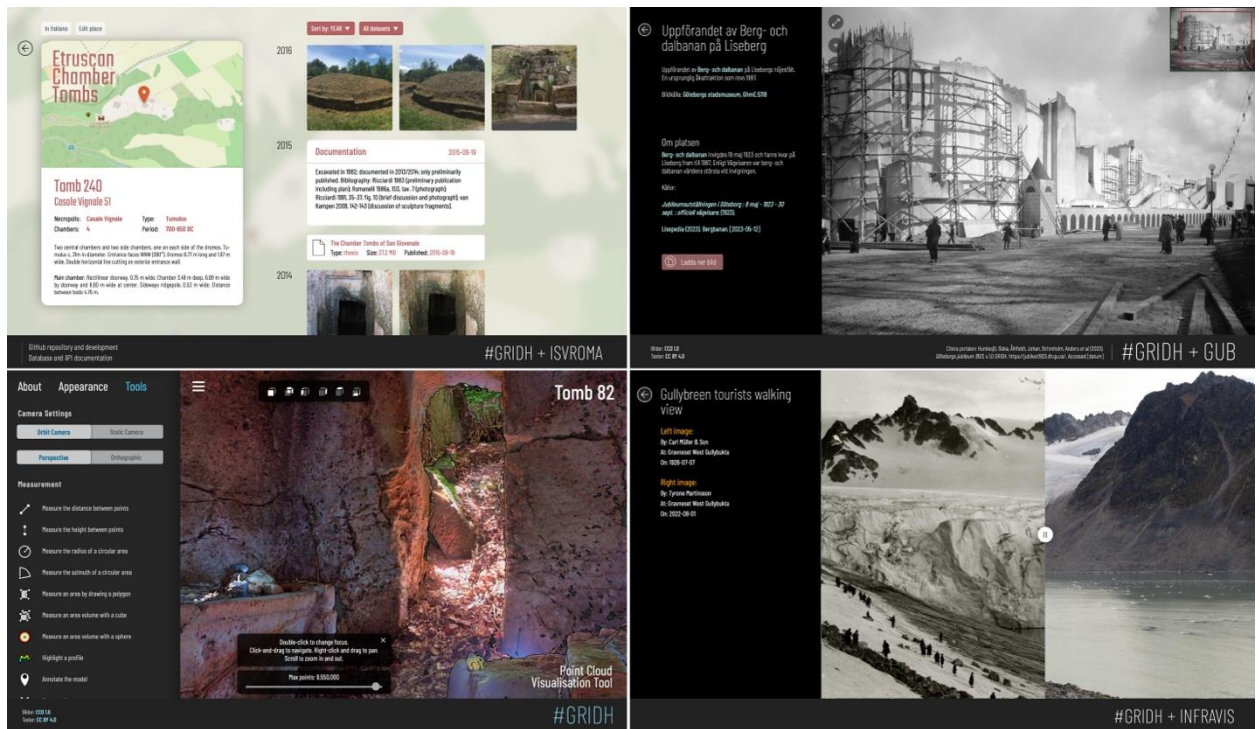


Figure 4: Place view and Point Cloud Viewer from Etruscan Chamber Tombs, Image viewer from Göteborgs Jubileum 1923, and Rephotography viewer from Extended Rephotography.

3. Frameworks and libraries used for MuM

In order to visualise meshes, point clouds, and RTI photography, the 3DHOP (<https://3dhop.net>), potree.js (<https://github.com/potree/potree>), and relight-viewer.js (<https://vcg.isti.cnr.it/relight/>) libraries were of particular interest. These come equipped with a wide range of features that facilitate user interaction and fast downloads of 3d assets through a pre-pyramidisation of the 3D data where only the necessary information is loaded at runtime. 3DHOP, for instance, provides built-in support for controlling scene lighting and carrying out measurements. Similarly, potree.js offers capabilities for camera manipulation and both distance, area, and volume measurement within a point cloud, and relight-viewer.js light manipulation and shader appearance. These features were considered invaluable for researchers interested in working with such material.

However, there are technical challenges involved in adding these libraries to MuM: 3DHOP and potree.js are predominantly jQuery-based libraries, while the main frontend framework for MuM is Vue3. The architecture and reactive databinding in Vue3 differ significantly from jQuery's more direct manipulation of the DOM (Document Object Model). Furthermore, Vue3 is designed to have primary oversight over its designated section of the webpage, and if jQuery makes any changes to that area, Vue3 may overwrite them during its next update. This leads to complexities in integration, as bringing jQuery into a Vue3 environment can result in conflicts and unexpected behaviour. Hence, a separate frontend/backend interface using Express and JavaScript were instead constructed. By building an auxiliary site with a near seamless connection with the main site, MuM-projects are able to utilise the 3DHOP, Potree and Relight libraries without the constraints of Vue3's architecture. This auxiliary site operates in parallel to MuM and serves as a unified platform for visualising and collecting all the point clouds and models across GRIDH's various platforms.

In addition to the auxiliary site built on the 3DHOP, potree.js, and relight-viewer.js libraries, MuM has currently support for streaming video and layered rephotography visualisation built on

open standards, and geographical representations and IIIF-image visualisation through OpenLayers (openlayers.org) and OpenSeadragon (openseadragon.github.io). Through its modular design, MuM can in the future easily be expanded with libraries and custom modules designed to access and visualise additional types of datasets as need arises, for instance through image clouds or WebXR, without breaking the overarching structure of the interface and the data model.

4. Diana and the data model

For its backend, MuM relies on a database coordination solution built by GRIDH, called Diana. Diana was written in Django with PostgreSQL, and it consists of an app providing base functionalities and abstract data models both for data input by the users and for making data accessible through generated REST APIs. With the tooling offered by Django, we can provide direct access to the database to our end-users, and limit their access to the projects they collaborate in. Through the Django admin site, the end-users can easily upload a variety of data without coding knowledge, and we can provide more tools for complex tasks such as batch uploads of data.

For each MuM project, a new application is written and installed in the Diana framework. Each application is generally centered around a data model indicating a Place, which includes some naming, categorization and, most important, a geography data field in the form either of geographical coordinates, or of a polygon indicating an area of interest. Places are then connected to other data models representing features such as images, 3D models, authors and/or reporters, observations, and various other forms of documentation, via Django's ForeignKey and ManyToMany fields (for more details, see Django Documentation). Tag models are used for categorization of other data models: in *Etruscan Chamber Tombs*, for example, we created tag-type models to describe different types of documentation, techniques used to develop 3D models, but also the epoch of datation of tombs. The data models for each project inherit some of their properties from the abstract data models provided through Diana. This ensures consistency in the database structure, while at the same time providing flexibility for the specificities of each project. For example, independently on what data models are specified within each specific application, each model comes with fields "created_at", "modified_at" and "published", that get automatically populated whenever a new data point is added to the database or modified. Some of the abstract models in Diana include *Tag models* and *Image Models*. *Tag models* consist of short case insensitive text, ideal for creating categories to which data points can be assigned. *Image models* include a field to upload IIIF-images (through GRIDH's IIIF server) as data points, as well as generating Universal Unique Identifiers automatically. This is but a short description of Diana's data models and how they interact with MuM, but a full treatment of Diana's capabilities is outside of the scope of this paper.

In addition to providing each MuM project with a database requiring minimal boilerplate code to be functional, Diana applications can potentially share data models and become an interactive powerhouse that gets more powerful the more projects make use of its functionalities. Diana shines also when it comes to the serialisation and generation of generic and consistent views in the form of REST APIs (including GeoJSON API), through the Django REST framework. This ensures the creation of compliant web APIs that the frontend relies upon. The flexibility of Django makes it easy to tailor these APIs to the needs of the frontend.

5. Conclusions

While it is close at hand to describe MuM as a *tool*, a software through which to register, access, and visualise a certain type of dataset, as has been shown it is to an equal amount a *method* for data organisation and curation; it informs an analytical approach to the material where a defined spatial position function as a fixture-point in Euclidean space for data of various types from various times. Hence, each point in space also becomes an archive of its own that organises data

pertaining to that place. The user approaches the dataset from an abstract representation of the data, as markers or analytical layers on a map or plan, but each step the user takes from there brings her closer to a more detailed representation of the data; first through a description of the spatially grounded place, and then through the individual representations of that spatial context served through the expanding set of visual data modules. The backend solution upon which MuM is developed allows for consistent data input and facilitates the interaction of end-users with the data shown in the frontend.

Acknowledgements

During the development of MuM, there have been several persons involved in providing code and solutions for both frontend and backend. Victor Wählstrand Skärström and Jonathan Westin instigated an early version of the frontend for *Reading the Signs*, then Arild Matsson, Tristan Bridge, and Jonathan Westin realised the first completed version for *Extended Rephotography* with backend support from Aram Karimi in Diana, which was initially developed by Victor Wählstrand Skärström. Tristan Bridge, Kristin Åkerlund and Jonathan Westin, with backend assistance from Johan Åhlfeldt, completed the *Göteborgs Jubileum 1923* and *Reading the Signs* iterations. Matteo Tomasini, together with Tristan Bridge and Jonathan Westin developed *Etruscan Chamber Tombs*, which served as a basis for *Sonora*, *The Inscriptions of Saint Sophia*, *Stokkastovan*, and *Built Cultural Heritage in Antarctica*.

The DIGARV Platform: A collaborative platform for working with cultural heritage data and research data

Johan Åhlfeldt¹ and Arild Matsson¹

¹ GRIDH, University of Gothenburg, Box 100, 405 30 Gothenburg, Sweden

Abstract

This article covers an easy-to-use research tool for collaborative work. The tool has been adapted for structured data and high-resolution images within four research projects at GRIDH. The platform is especially designed for working with temporal and spatial data. Furthermore, the platform gives researchers access to a relational database system through input forms and access to external cultural heritage data including high-resolution images. This way the platform also aims to utilize external data published as Linked Open Data (LOD) and, at the same time, prepare its own research data for publishing as LOD. Because of the spatial and temporal nature of the data, it is visualized in time and space through maps and timelines to give overview and context during the data management phase.

Keywords

Database management, Linked Open Data, IIIF, SQL, Online publishing, Online collaboration, GIS.

1. Introduction

In 2019 the Centre for Digital Humanities (CDH), now known as the Gothenburg Research Infrastructure in Digital Humanities (GRIDH), was part of two successfully funded projects in the *Digitisation and accessibility of cultural heritage* (DIGARV) call, <https://digarv.se/en/>. The two projects were *Expansion and Diversity*¹ and *Mapping Lived Religion*². Both projects are characterized by data driven approaches, curated datasets, large projects groups and collaborative work in collecting and managing data from various sources, both digital, printed sources and manuscripts.³ The responsibility of GRIDH was to facilitate the collection of existing datasets into an information system, one database for each project, and set up a system for collaborative data management among the researchers.


Despite the different project focuses, Medieval history & Art history and Cultural studies respectively, both are methodologically oriented towards visualizing and analyzing data in time and space on maps. The map and timeline visualize the pattern and distribution of the phenomenon under study on the aggregate level (distant reading) and at the same time, give users access to individual cases with all its evidence in full detail (close reading). Both projects maintain a need to import and manage data online.

We decided to build two spatial enabled and relational databases (SQL) from data models that emerged from the needs of each project and their respective research questions. The platform, which we at GRIDH called the *DIGARV Platform*, was built at the same time to accommodate the needs of both projects. We decided to build both projects on the same code base split between a backend and a frontend

¹ <https://www.gu.se/en/research/expansion-and-diversity-digitally-mapping-and-exploring-independent-performance-in-gothenburg-1965-2000>

² <https://lnu.se/en/research/research-projects/mapping-lived-religion-medieval-cults-of-saints-in-sweden-and-finland/>

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

 johan.ahlfeldt@lir.gu.se (J. Åhlfeldt); arild.matsson@gu.se (A. Matsson)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

³ Nilsson *et al* 2023. Liepe, L., & Ellis Nilsson, S. (2021)

part, as detailed in later sections, the only difference lies in their different configuration files. Each project was assigned a dedicated sub domain on top of the domain for GRIDH which is at <https://dh.gu.se>, <https://expansion.dh.gu.se> and <https://saints.dh.gu.se> respectively. The intention was to build a common platform that could also be used in future projects which was the case for the project *Pehr Strands flöjtur* (Pehr Strand's Pipe Organs) [3] and the internal project at GRIDH *Iconographia Mediaevalis Suetica*⁴.

The DIGARV Platform is primarily used as a space for collaborative work. However, both projects decided to keep the platform open to the public from the beginning of its existence, except for the part of editing data. This open approach was important in the *Expansion and Diversity* project to communicate with the informants of the project and, in the *Mapping Lived Religion* project the members of the reference group. In the *Mapping Lived Religion* project, the platform is also used by a similar project at Tampere university in Finland, *Lived Religion in Medieval Finland*, together more than 10 researchers. During the project period the *Mapping Lived Religion* project have had 16 students editing data as part of their academic training. Users can be added to the system as members to gain edit privileges by the system administrator at GRIDH. The platform has been continuously evaluated and adjusted to better meet the needs of the projects and its members and to create a smooth workflow.

This paper doesn't cover the resulting public interfaces for the two projects which were developed later.

2. User interface

The DIGARV Platform was created with a split frontend–backend architecture. The frontend is implemented in the Vue 2 JavaScript framework⁵ and notably uses the OpenLayers map library⁶ and the OpenSeaDragon image viewer (see **Section 7**), as well as multiple smaller libraries for interactive components, etc.

The user interface is designed to visualize spatial and temporal data on maps. The setting for the temporal data connected to places on the map is adjusted by a time slider, using Min, Max and step values, see the top left in **Figure 1**. The width of the map and the tab pane can be adjusted seamlessly keeping the center of the map. Multiple layers can be displayed at the same time and different background layers can be selected.

To the right of the map is a set of five tabs, see the top right in **Figure 1**. The *Home* tab is a short introduction to the project in question. The *Layers* tab is where the user selects background layers and available layers as overlays. The *Search* tab is where the user search for entities in the database and selects one of them for display or edit. From the *Search* tab users can also create new records for each entity. The *Show* tab displays all data connected to one record, including high resolution images and relations to other objects. Finally on the *Edit* tab records can be created or edited.

Navigation between tabs is automatic but users can also shift between the tabs at any time and content is preserved. This way, a multitude of information and settings are displayed in a very compact and easy to access way.

The *Edit* and *New* buttons and the *Edit* tab are only displayed when a user is logged in, see near the top right in **Figure 1**. When users are logged in, individual contributions are tracked and saved to the database. This is important for responsibility, traceability and later in the public interface, proper attribution.

The default language is English. However, we have prepared for several interface languages, in the case of *Mapping Lived Religion*, the public interface will also be available in Swedish and Finnish.

⁴ <https://iconographia.dh.gu.se>, no English version available.

⁵ <https://v2.vuejs.org/>

⁶ <https://openlayers.org/>

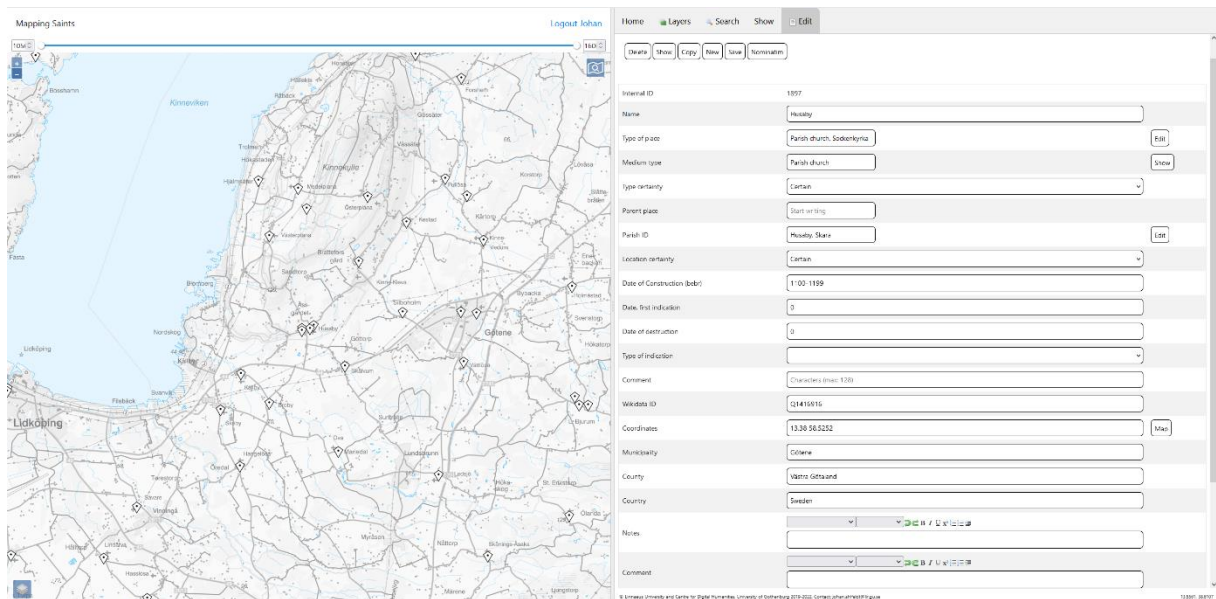


Figure 1: The interface of the *Mapping Saints* application. The map is centered around Husaby parish church, Götene municipality in Sweden. The Edit tab is available when logged in as a member. The terrain map (in gray scale) by Lantmäteriet (Land Survey) is selected as the background layer. The place overlay in this example represents all places available in the database and information about a particular place can be accessed by clicking the marker on the map, or by searching for the place name in the search interface. Map navigation takes place through pan and zoom. A full resolution screenshot is available at https://dh.gu.se/data/figure1_mapping_saints_husaby.png. The *Show* tab of the Husaby parish church is available here, <https://saints.dh.gu.se/place/1897> with further public access to all tabs except the Edit tab.

3. Database implementation and population

According to the aim of each project and their data, two different database models were created. The commonality of their structures made it easy to develop a similar code base. The databases are implemented in MySQL, with database tables of two sorts.

The first is for the main entities in the database, e.g., cult, place, parish, persons in the *Mapping Lived Religion* project, and organisation, person, place, production, event tables in the *Expansion and Diversity* project.

The second type of tables are the relation tables. These connect the entities to each other and sometimes with a qualifier property such as certainty or a specific role of an agent.

There are also tables connecting entities in our database with entities in external databases, especially the Swedish Open Cultural Heritage database (SOCH)⁷. This allows us to explore and reuse cultural heritage objects within our own interface in real time and can subsequently be used as evidence in our research (see below **Section 4** on Linked Open Data).

In all tables, an integer field id is used as primary key, i.e., a unique identifier for each row in the database table. If a field is a reference attribute or foreign key to another table, it is always the id field that is referenced. For geographical data we always use the GEOMETRY data type, capable to hold different kinds of geometries, i.e., POINT, LINESTRING or POLYGON. This column is always named *geom*. Many of the tables contains fields like *created*, *updated* and *modified*. They can be changed automatically with the information that a logged in user provides or the current date and time.

Aside from the object and relation tables containing the main project data, each project database also contains a special *Column Labels* table. Each row describes a field in a database table and contains a human-readable label, instructions for the editing form and whether the field should be hidden from

⁷ <https://www.raa.se/in-english/digital-services/about-soch/>

display, read-only and searchable. For fields that function as references to other tables, they also indicate the table and field name of the reference target. These values are included when the backend application serves detailed data for a given object, helping the frontend application to use and render the fields appropriately.

In the early phase of the *Expansion and Diversity* project, we received a copy from Göteborgs Stadsmuseum [the City Museum of Gothenburg] of their database Carlotta, regarding data about non-institutional performing art groups as well as related persons and events. There were differences between this data model and our own – for instance, we separate Production entries (e.g. a play or a dance piece) from Event entries (the performance of a Production at a certain time and place). Nevertheless, we were able to use this data firstly for populating parts of our own database and secondly for maintaining identifiers of corresponding data entries in SOCH. When our list of performing art groups was expanded as a result of our research, we made a simple API for the museum to use to collect our persistent identifiers for groups already in their collection and the new ones.

By assigning subdomains to projects and minting persistent identifiers for the main entities in the database, we can reference everything in the database directly by HTTP URIs. These URIs can be used by other datasets to reference and download content from our information system in real time. They can also be used in scholarly literature. For instance, <https://saints.dh.gu.se/place/1897> denotes Husaby parish church, Götene municipality, Sweden, and <https://saints.dh.gu.se/cult/88> denotes the Altarpiece with five sculpted saints and paintings depicting scenes from St. Lawrence's legend in Adelöv parish church.

4. Linked Open Data (LOD)

The concept of Linked Open data is a compilation of principles and methods for data publishing reusing/sharing of data. When everyone publishes their data online with an open license, makes them machine readable, describes them in a standardized way and links them to other data, the Internet itself can be used as a common database. This enables anyone to build digital services based on data from several providers, for instance, Cultural heritage and Research institutions. In our case, we can integrate data from remote providers in our own platform instead of making copies and storing them on our servers.

Since many of the digital sources we utilize are published in SOCH, it is a good idea to also publish our research data in SOCH. Our contribution, especially in the Mapping Lived Religion project, is to bring scholarly comments and concepts on cultural heritage objects and put them in a historical and analytical context. This is accomplished by referring to pieces of source material by their persistent identifiers. Linked open data is also helping us define our concepts and make them compatible with other online datasets. Besides Wikidata, the Art & Architecture Thesaurus⁸ and Iconclass⁹ are two examples of such classification systems. Another contribution of the Saints project is the compilation of a list of medieval cult places, with coordinates, variant and historically attested names and with references to online sources, such as Bebyggelseregistret [protected buildings] and Fornsök [archaeological sites] at Riksantikvarieämbetet, the Swedish National Heritage Board.¹⁰

5. Backend application

The backend of the DIGARV Platform is a lightweight PHP 7-driven server application. Except for some specific functionality, the code is free from dependencies to any specific research project. It features a database query layer and an API (see next section) for listing, describing and editing data.

⁸ <https://www.getty.edu/research/tools/vocabularies/aat/> by the Getty Research Institute.

⁹ <https://iconclass.org> is a classification system for the content of images and maintained by the Henri van de Waal Foundation.

¹⁰ Nilsson et al (2023).

There is a strict separation of concerns between data content and appearance, so the backend accepts data in GET and POST HTTP requests and responds with data in the JSON format.

The generalized code is specified to a certain research project by means of a *Tables File*. This is a PHP file that declares the data model: Object types, relationships between them and what fields to use for labelling objects, searching them, positioning them on a map, et cetera. The file contains a single multi-dimensional array following a documented structure.

To deploy the backend for a certain project, the code is copied to a web-accessible root directory on the server. A config file is added, containing database credentials and a pointer to the Tables File associated with the project.

6. API

The web application programming interface (API) of the backend offers ten endpoints for listing, finding, describing and editing data. Describing three of these may be sufficient to outline the capacity of the service.

- The *Search* endpoint lists items of a given type according to given criteria. As an example, a usage could be to list *Organisations* founded after 1980 and having a page on Wikipedia.
- The *Map* endpoint does something similar but returns the list in the *GeoJSON* format [5], compatible with GIS software such as the map display in the frontend.
- The *Edit* endpoint, given an object type and an identifier, returns the full details of the indicated object. As the name suggests, the returned data is used not only for displaying information, but also to generate an edit form (by a user with the appropriate permissions).

7. Serving high-resolution images with IIIF

Newly added data points in the Mapping Lived Religion project are being connected with previously existing data in the *Medeltidens bildvärld* photographic collection of the Swedish History Museum¹¹ and the *Iconographic register*¹² as part of the digital archive at the National Heritage Agency. These collections are available by API through SOCH, but high-resolution images could only be obtained offline (on physical storage media, in batches).

The *IIIF* protocol was developed by a consortium of actors in the libraries and academia, founded in 2015, to solve various problems with delivering media on the web. Its Image API specifies a method of saving bandwidth when serving high-resolution images¹³. In essence, it relies on fixed-size slices of each image at different resolutions. Thus, a single low-resolution slice may be enough to show a thumbnail. Conversely, when zooming in on a section of the image, only the few high-resolution slices overlapping with the viewport are necessary.

Among the multiple open-source IIIF implementations, the *IIPImage server*¹⁴ was selected for its speed and installed in the local server infrastructure at GRIDH. As a prerequisite, images need to be in a multi-resolution format. To this end, the *VIPS* image processing utility¹⁵ was executed on the server after uploading the original *Medeltidens bildvärld* images.

The frontend application employs the *OpenSeaDragon* viewer component¹⁶ to use the IIIF-supported image service. This allows visitors to the Mapping Saints website to explore photographic representations of cult objects in great detail, without much lag or disruption. The images from the

¹¹ Medeltidens bildvärld has its website (with only low-resolution images) at <https://medeltidbild.historiska.se/medeltidbild/default.asp>

¹² Liepe, L., & Ellis Nilsson, S. (2021).

¹³ Read the IIIF Image API 3.0 at <https://iiif.io/api/image/3.0/>

¹⁴ <https://iipimage.sourceforge.io/documentation/server>

¹⁵ <https://www.libvips.org/API/current/using-cli.html>

¹⁶ <https://openseadragon.github.io/>

Iconographic register (register cards with photographic images) were much smaller in size and didn't need to be in a multi-resolution format to display quickly in the *OpenSeaDragon* viewer.

8. Legacy

The platform backend code is almost entirely developed internally by research engineers at GRIDH. Unlike most open digital research projects, it does not utilize many external code packages. Consequently, it suffers from some issues including incomplete security, susceptibility to bugs and a high time cost of adding common functionality.

This motivated GRIDH, in later years, to create a new system with similar goals as the DIGARV Platform but relying more on free and open-source software frameworks. Since then, the new system, named Diana, has been the preferred collaboration platform for new research projects.

Diana builds on top of Django¹⁷, a Python library for managing database systems, to create input forms and API endpoints. Django also have a robust authentication and authorisation system down to the database table level for individual users and groups. The admin interface of Django (the input forms) is generated by directives written in Python code. The output are HTML pages which can be customized further for the purpose of displaying source material like documents and images. We have already implemented a display of PDF-documents and a static IIF-client to display images next to the input forms. There is also a map editing interface which can modify existing geometries and create new ones.

As discussed above, the DIGARV Platform has many of these features, some of them are implemented in a rudimentary way, others are implemented in a more advanced way, highly adapted for the research task at hand. The advantage of the DIGARV Platform at this point is that it implements a visualization of all places and optionally more specific layers important for the editing process. The platform also fully implements a IIF-client where images can be panned, zoomed and rotated. Much work remains to build a custom platform on top of Django with the same functionality as the DIGARV Platform.

9. Conclusions

The DIGARV Platform is a highly adaptable platform for collaborative research working with structured data and relational databases. It is also a platform for the integration of data from external information system, including image evidence as source material for scholarly work. To researchers and the general public alike, it presents data for distant and close reading, and specializes in spatial-temporal data most notably through an interactive map with an adjustable time window.

Compared to the later Django-based system, the DIGARV Platform has both advantages and drawbacks. While most functionality comes included with Django, certain specific features are easier to develop in the bespoke in-house platform.

The programming code for the entire platform will be published on GRIDH's Github account, <https://github.com/gu-gridh>

Acknowledgements

The Expansion and Diversity and the Mapping Lived Religion projects are funded by the Digitisation and accessibility of cultural heritage (DIGARV) research programme through The Swedish Research Council, The Royal Swedish Academy of Letters, History and Antiquities, and Riksbankens jubileumsfond.

¹⁷ <https://www.djangoproject.com/>

References

- [1] Ellis Nilsson, S., Zachrisson, T., Fröjmark, A., Liepe, L., & Åhlfeldt, J. (2023). 'Mapping Saints: creating a digital spatial research infrastructure to study medieval lived religion'. In: *Digital Spatial Infrastructures and Worldviews in Pre-Modern Societies*, A. Petrulovich & S. Skovgaard Boeck (eds.), Arc Humanities Press, 2023, p. 33-58
- [2] Liepe, L., & Ellis Nilsson, S. (2021). 'Medieval Iconography in the Digital Age: Creating a Database of the Cult of Saints in Medieval Sweden and Finland'. Iconographisk post: *Nordisk tidskrift för ikonografi*. (2). 45-63.
- [3] Norrback, Johan (2021). "Pehr Strands flöjtur." Centrum för digitala humaniora, Göteborgs universitet. URL: <https://strand.dh.gu.se> (version 1).
- [4] von Rosen, A., Holgersson, H., Strömberg, M., Adler Sandblad, F. & Grehn, S. "Expansion och mångfald - en relationell forskningsdatabas". Expansion och mångfald: Digital kartläggning och analys av den utominstitutionella scenkonsten i Göteborg 1965-2000 – Institutionen för kulturvetenskaper, Göteborgs universitet, 2021.
- [5] Butler, H., M. Daly, A. Doyle, Sean Gillies, T. Schaub, and Stefan Hagen. "The GeoJSON Format." Request for Comments. Internet Engineering Task Force, August 2016. <https://doi.org/10.17487/RFC7946>.

Samförfattande som datadriven tvärvetenskap: Pragmatiska lärdomar från SweTerror-projektet

Daniel Brodén^{1,2}, Mats Fridlund^{1,2}, Leif-Jöran Olsson^{3,4}, Magnus P. Ängsal⁵, Patrik Öhberg⁶

¹ Göteborgs forskningsinfrastruktur för digitala humaniora (GRIDH), Göteborgs universitet, Renströmsgatan 6, Göteborg, 40530, Sverige

² Institutionen för litteratur, idéhistoria och religion, Göteborgs universitet, Renströmsgatan 6, Göteborg, 40530, Sverige

³ Nationella språkbanken, avdelning Språkbanken Text, Göteborgs universitet, Renströmsgatan 6, Göteborg, 40530, Sverige

⁴ Institutionen för svenska, flerspråkighet och språkteknologi, Göteborgs universitet, Renströmsgatan 6, Göteborg, 40530, Sverige

⁵ Institutionen för språk och litteraturer, Göteborgs universitet, Renströmsgatan 6, Göteborg, 40530, Sverige

⁶ SOM-institutet, Institutionen för journalistik, medier och kommunikation, Göteborgs universitet, Seminariegatan 1B, Göteborg, 413 13, Sverige

Abstract

Terrorism i svensk politik (SweTerror) är ett storskaligt tvärvetenskapligt forskningsprojekt med forskare från såväl human- och samhällsvetenskaperna som datavetenskaperna. Samtidigt använder och utvecklar SweTerror nationell forskningsinfrastruktur för riksdagsdata. Detta paper beskriver användningen av samförfattande som en datadriven tvärvetenskaplig praktik för att integrera olika vetenskapliga perspektiv och skapa samsyn i projektforskningen. Vi tar fasta på betydelsen av valet att koncentrera samarbetsformen kring konferenspapers inom specifikt digitala humaniora och diskuterar erfarenheten av att samskrivande försvagar vetenskapligt revirtänkande, liksom ett iterativt förhållningssätt till forskningsdata kopplade till forskningsinfrastrukturer under uppbyggnad. Avslutningsvis betonar vi datadrivet samförfattande som en pragmatisk praktik för att stärka kollaborativt samarbete och kunskapsbryggor inom en tvärvetenskaplig forskargrupp.

Keywords

Tvärvetenskap, samförfattande, digitala humaniora, forskningsinfrastruktur

1. Introduktion

‘Terrorism i svensk politik (SweTerror)’ [1] (2021–2026) är ett stort tvärvetenskapligt projekt som utforskar den parlamentariska diskursen om terrorism i Sverige genom både data-intensiva och ‘traditionella’ kvalitativa studier av texttranskriptioner och ljudinspelningar från riksdagsdebatterna i Sverige 1968–2018. SweTerrors forskningsdesign är uppbyggd kring mixade metoder [2] och omfattar forskare från såväl human- och samhällsvetenskaperna (HS) som teknikvetenskaperna. I projektet samarbetar experter inom bland annat digitala humaniora, terrorismstudier, idéhistoria, lingvistik, statsvetenskap, språk- och talteknologi. En grundtanke är att projektet genom att förena olika vetenskapliga tanke- och arbetssätt kan generera en kunskap som varit svår att åstadkomma tidigare inom ramen för de enskilda disciplinerna. I bedömningen av projektet lade Vetenskapsrådet (VR) vikt vid sammansättningen av forskargruppen och skrev att SweTerror “creates a real opportunity to change the nature of scholarship on the politics of terror” genom ett “remarkably interdisciplinary team of researchers” [3].

Betydelsen av ‘tvärvetenskap’ har varit föremål för en utdragen och inte sällan motstridig akademisk debatt med delvis varierande innebörder beroende på sammanhang. För enkelhets skull tar vår framställning en utgångspunkt i Raymond Millers breda förståelse av *interdisciplinarity*, eller tvärvetenskaplighet, som “the generic all-encompassing concept and includes all activities which juxtapose, apply, combine, synthesize, integrate or transcend parts of two or more disciplines” [4].

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

✉ daniel.brodén@lir.gu.se (D. Brodén); mats.fridlund@lir.gu.se (M. Fridlund); leif-joran.olsson@svenska.gu.se (L.-J. Olsson);

magnus.petterson.angsal@sprak.gu.se (M.P. Ängsal); patrik.ohberg@som.gu.se (P. Öhberg)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Likaså koncentrerar vi oss på tvärvetenskaplighet som något mer eller mindre entydigt fruktbart, medvetna om att det i det ämnesöverskridande också ligger en risk att tunna ut eller släta över vetenskapliga särarter.

Detta paper synliggör och diskuterar datadrivna, tvärvetenskapliga samarbetsformer utifrån den forskning som genomförts inom SweTerror-projektet av de projektdeltagare (Brodén, Fridlund, Olsson, Ängsal och Öhberg), som hittills samarbetat kring textmaterialet från riksdagen med fokus på dataanvändning och språkteknologisk analys, samt forskningsinfrastruktur. Det specifika syftet är att diskutera samförfattande av vetenskapliga publikationer som en konkret tvärvetenskaplig praktik för att integrera olika vetenskapliga perspektiv. Centralt är att samskrivning kan förena ett “larger network of individuals with a variety of skills and knowledge” och skapa en form av “synergies that is often not found in solitary research” [5].

1.1. Disposition

Nedan diskuterar vi samskrivning av konferenspapers som en integrativ tvärvetenskaplig praktik för gemensam projektforskning och utveckling av forskningsinfrastruktur. Vi börjar med att beröra idén om olika vetenskapliga kulturer och integrativ tvärvetenskaplighet för att etablera ett ramverk för diskussionen av de datadrivna samarbetsformerna i SweTerror. Därefter reflekterar vi över konkreta aspekter av vårt samförfattande med fokus på en publiceringsstrategi för att försvaga vetenskapligt ‘revirtänkande’ och understödja ett iterativt förhållningssätt till forskningsdata. Vidare diskuterar vi hur den kollaborativa arbetsprocessen är kopplad till forskningsinfrastrukturer under uppbyggnad. Vi avslutar med att sammanfatta våra centrala slutsatser om samskrivande som en praktik för att skapa kunskapsbryggor inom en tvärvetenskaplig forskargrupp.

2. Integrerad tvärvetenskap

Som bekant, skrev C.P. Snow [6] om ‘de två kulturerna’ med avseende på hur särskilt humaniora och natur- och teknikvetenskaperna skiljer sig åt beträffande inriktning, syfte, kunskapssyn, et cetera, vilket försvårar möjligheten till gemensam kommunikation och ömsesidig förståelse. Samtidigt har det talats om möjligheten till en ‘tredje kultur’ genom tvärvetenskapliga överbyggningar [7]. Digital humaniora, som nämns i detta sammanhang [8], har flera likheter med natur- och teknikvetenskaperna, inte bara med avseende på kombinationen av kvantitativa och kvalitativa metoder, utan också vetenskapliga praktiker som projektforskning och publiceringsformer.

Det finns inom digital humaniora en diskussion om utvecklingen av kollaborativa processer för projektforskning där flera kommentatorer har framhållit att robusta resultat inom fältet är beroende av inte bara grundlig datavetenskaplig analys, utan också tvärvetenskapligt samarbete. Oberbichler m.fl. [9] skriver att en *integrerad* tvärvetenskaplig projektforskning innebär att “people from different scientific fields come together, collaborate, and study a common question or problem with the goal of reaching common conclusions”. Oberbichler m.fl. betonar också vikten av att skapa en samsyn inom forskargruppen: “the differences and commonalities between disciplines need to be considered. Merging of applications, tasks, and traditions, involving mixed method approaches as well as increased interaction between the disciplines, has been identified as a possible common objective” [9].

Mot en sådan bakgrund har ett tvärvetenskapligt projekt som SweTerror behov av plattformar för att utveckla samsynen mellan forskarnas olika perspektiv. Till exempel höll vi i inledningsskedet av projektet en serie seminarier, där de olika deltagarna fick presentera sin forskning med fokus på teoretiska och metodologiska utgångspunkter. För att överbygga den ‘teknologiska’ klyftan mellan HS-forskare och språkteknologer i projektet hålls kontinuerliga arbetsmöten med utgångspunkt i den parlamentariska datan, dess egenart och analysmöjlighet – det är knappast givet hur HS-forskare utan kunskap om premisserna för datadriven textanalys ska kunna begreppsliggöra exempelvis ordvektorer.

Såttillvida kan föreliggande studie beskrivas i termer av ‘tillämpad kritisk digital humaniora’ i avseende på att utvecklingen av robust och tillförlitlig datadriven HS-forskning är beroende av kritisk reflektion om de egna metoderna, verktygen och praktikerna [10].

3. Praktiker för datadrivet samförfattande

För att på konkreta sätt etablera en kollaborativ arbetsform och tvärvetenskaplig dialog beslutade vår forskargrupp på ett tidigt stadium för att koncentrera projektforskningen kring samförfattande. Centralt är att digital teknik förändrat förutsättningarna för kollaborativ tvärvetenskaplighet med det webbaserade samskrivningsverktyget Google docs som en symbol för utvecklingen [11]. Genom att kombinera Google docs och den webbaserade videomötestjänsten Zoom har vi etablerat en distribuerad samarbetsplattform som inte är bunden av platsmöten och som gör att vi enkelt kan skriva tillsammans och dryfta olika aspekter av texten i realtid. En annan fördel med samskrivande i videomöten är att det samlar deltagarna för konkret och intensivt arbete, med flexibilitet för deltagare att vid behov kan lämna Zoom stundvis för skrivande på textavsnitt på egen hand. Samtidigt som sådana 'digitala skrivarstugor' knappast är något unikt, har skrivarbetet i SweTerror också fått en viss särprägel, som är kopplad till vårt val att koncentrera den kollaborativa processen kring konferenspapers inom företrädesvis digitala humaniora.

3.1. Publikationskulturer och försvagade revir

I en diskussion av så kallad integrativ humaniora skriver Ekström och Sörlin [12] om en 'andra gradens specialisering', såtillvida att sammanvävningen av olika vetenskapliga kulturer bidrar till att ny kunskap utvinns om ett studieobjekt, men också att forskningsarbetet bedrivs med 'svaga akademiska revir' (i positiv mening). En sådan beskrivning ter sig träffande med avseende på vårt samskrivande, som är avsett att stimulera rörelse över invanda ämnesgränser, men också tillit till kollegornas specifika ämnesexpertis. Publikationskulturen kring konferenser inom digital humaniora har präglats av fältets bakgrund i datavetenskapliga ämnen, däribland informatik, språkteknologi och datalingsvistik. Medan till exempel publikationer i statsvetenskapliga konferenser tenderar att utgå från publikationsfärdiga artiklar, kretsar konferenser inom digital humaniora på ett annat sätt kring prioriterade idé- och metod-papers för jämförelsevis snabb publikation i en konferensvolym i nära anslutning till konferensen. Formatet har för vårt samskrivande inneburit relativt korta ledtider och behov av att 'snabbt komma i gång'. Det innebär att vi snarare än att reda ut meningsskiljaktigheter har fokuserat på vad vi kan komma överens om. På så vis har skrivprocessen fått ett starkt drag av pragmatism. Men samtidigt som vi sökt en pragmatisk samsyn i olika frågor, har arbetsprocessen varit beroende av en tillit till andras expertis, såtillvida att de olika deltagarna har ansvarat för skrivandet av 'det som de är bra på'. Till exempel har statsvetaren (Öhberg) haft ett huvudansvar för textavsnitt om den parlamentariska kontexten, medan språkteknologen (Olsson) fokuserat på datadriven textanalys. Även om projektet hittills inte mött några större utmaningar i det tvärvetenskapliga samarbetet, har vi beslutat att i kommande artiklar, riktade till ämnestidskrifter, följa principen att vid behov låta respektive huvudförfattare fatta beslut om utformning av innehållet utifrån sina specifika disciplinära perspektiv.

3.2. Datadriven projektforskning

Samförfattandet i SweTerror utgår från ett iterativt utforskande av vår forskningsdata. Ett sådant datadrivet arbetssätt har setts som viktigt för att skapa integrerad tvärvetenskaplig projektforskning [9], men är i sig knappast unikt för digital humaniora. Till exempel utgör den nationella enkätdata som SOM-institutet (Samhälle, Opinion och Medier) [13] tillhandahåller ett underlag för projektforskning inom hela HS-området med mer eller mindre tvärvetenskaplig karaktär. Likväl har koncentrationen av vår forskningsprocess kring riksdagsdata visat sig vara ett effektivt sätt för hela gruppen att bygga upp kunskap om både själva datan och dess kontext, inte minst riksdagens arbetsformer. Det har direkt påverkat vårt sätt att ställa forskningsfrågor om den parlamentariska diskursen. Till exempel har vi genom att använda data från riksdagsledamöters motioner undersökt riksdagspartiernas responsivitet i relation till medborgerlig opinion (kring politiskt våld) [14]. Vår 'datafokuserade' utgångspunkt har vidare legat till grund för det viktiga beslutet att strukturera hela vårt arbete med både data-set och analyser kring *riksdagsår* (som löper från höst till höst) istället för kalenderår [15]. Såvitt vi kan se, har inga andra forskningsprojekt inom digital humaniora kring riksdagsdata tagit fasta på den inom statsvetenskaplig forskning grundläggande utgångspunkten att riksdagsår, vilket styr mandatperioden för regering och riksdag, liksom dynamiken kring propositioner och motioner, riksdagsdebatter, et

cetera. Till exempel har SweTerror-projektet [14] visat att riksdagspartiernas aktivitet när det gäller motionerande i hög grad påverkas av om de befinner sig i regeringsställning eller opposition.

3.3. Publikationer som flerstegsraket

En grundtanke bakom vårt samskrivande har varit att successivt arbeta fram en forskningsartikel genom en serie konferenspapers. Denna karaktär av ‘flerstegsraket’ kan illustreras med processen bakom en av projektets första antologipublikationer [16]. Eftersom det vid projektstart saknades tillräckligt kurerad riksdagsdata, började vi med att studera nationell opinionsdata från SOM-institutet om terrorism och politisk extremism som en slags referenspunkt för riksdagsdiskursen. Skrivarbetet utmynnade, bland annat, i ett bokkapitel i SOM-institutets årsbok [17], men även konferenspapers som jämförde opinionsdata med data från Valforskningsprogrammet om riksdagspolitikens attityder. Jämförelserna visade att det fanns en klar vänster-högerdimension i medborgarnas och politikernas förhållningssätt, där personer till vänster var mer oroliga för extremism, medan personer till höger var mer oroliga för terrorism. Resultaten riktade vårt analytiska intresse mot huruvida mönstret återspeglades i skrivandet av riksdagsmotioner. Därför skrev vi ett konferenspaper som kombinerade en kvantitativ analys av data om motionerandet kring dessa frågor och en kvalitativ fallstudie med närläsning av relevanta motioner från två riksdagsår [14]. Baserat på denna serie av publikationer, skrev vi slutligen ett syntetiserande kapitel till en antologi, som för närvarande är under utgivning på internationellt förlag [16]. Centralt är att detta kumulativa skrivarbete både genererat delresultat som tillsammans blivit ‘något mer’ och att denna flerstegsraket drivit den kollaborativa arbetsprocessen framåt som helhet.

3.4. Infrastruktur under utveckling

En utmaning i det kollaborativa arbetet har varit att projektet både utgår från och bidrar till forskningsinfrastrukturer som är under uppbyggnad. De ofta tidskrävande datainhämtnings- och kureringsprocesserna hanterades genom att HS-forskarna i ett inledande skede främst utforskade textmaterial från riksdagen som redan var tillgängligt hos Språkbanken Text och Riksdagens öppna data. Centralt är att HS-forskarna då i stor utsträckning närläste riksdagsmaterialet på ‘traditionellt’ vis, men ju längre utvecklingen av forskningsinfrastrukturen framskridit, ju större roll har den storskaliga språkteknologiska analysen spelat. Ett sätt att beskriva förskjutningen är att säga att samarbetet mellan HS-forskare och språkteknologer blivit mer utvecklat och framskjutet. Samtidigt ska det betonas att det inledande kvalitativa närläsningens arbetet varit värdefullt då det bidragit till den konkreta materialkännedom, som stärkt det reflekterande synsättet på och förutsättningarna för den storskaliga språkteknologiska textanalysen. Vidare har SweTerror inlett ett samarbete med projektet WeStAc (Welfare State Analytics, 2019–2024, DIGARV), [18] och forskningsinfrastrukturen Swerik (2023–2026, RJ), [19] som bygger upp och tillgängliggör ett fullständigt digitaliserat textmaterial av bland annat riksdagsdebatter och motioner från Riksdagsbiblioteket. Samtidigt som WeStAc-korpusen [20] ger SweTerror tillgång till debattprotokoll med förbättrad OCR-kvalitet och metadata om de svenska riksdagsledamöterna, bidrar SweTerrors språkteknolog (Olsson) till WeStAc och Swerik genom databerikning och -kurering, liksom tekniska standarder (FAIR) och kvalitetskontroll. Vi har också i vår kollektiva arbetsprocess identifierat luckor i de digitaliserade debattprotokollen och utvecklar för närvarande ett konceptuellt ramverk för skapande av kontext-orienterade data [21] [22] med utgångspunkt i våra specifika forskningsintressen [23].

4. Sammanfattning

I detta paper har vi diskuterat våra datadrivna samarbetsformer för integrativ tvärvetenskap i projektet SweTerror. Framställningen har betonat möjligheten att tidigt etablera en kontinuerlig dialog över ämnesgränser genom fortlöpande seminarier, iterativt arbete kring data och främst samskrivning av publikationer. Utöver att allmänt framhålla värdet av samförfattande som en gemensam arbetsplattform, har vi pekat på hur en publikationskultur inom digitala humaniora med relativt korta deadlines för konferenspapers kan ge pragmatiska och resultatorienterade förhållningssätt till skrivprocessen.

Formatet har för vårt samskrivande inneburit en förskjutning i gemensamt fokus från vetenskapliga meningsskiljaktigheter mot vad vi kan enas om. Vidare har vi lyft fram poängen med att arbeta med en serie konferenspapers som en 'flerstegsraket', där det kumulativa arbetet utmynnar i en syntetiserande forskningsartikel. Samtidigt har vi tagit fasta på möjligheten att bygga upp tvärvetenskaplig projektforskning kring ett iterativt utforskande av forskningsdata. Slutligen har vi beskrivit betydelsen av att vårt samskrivande skett i relation till en forskningsinfrastruktur under uppbyggnad, där den till en början begränsade tillgången till data indirekt medfört att vi byggt upp 'kvalitativ' materialkännedom, som bidragit till den senare storskaliga, språkteknologiska analysen. I förlängningen har det datadrivna samskrivandet stärkt inte bara det tvärvetenskapliga samarbetet utan också vår tillit till varandras specifika vetenskapliga expertis.

Acknowledgements

Projektet SweTerror (<https://sweterror.se>) finansieras av forskningsprogrammet DIGARV (VR, RJ och Kungliga Vitterhetsakademien). Arbetet som presenteras i föreliggande paper har också kopplingar till de nationella forskningsinfrastrukturerna Huminfra (VR, kontrakt nr. 2021-00176) och Swe-Clarin och Nationella Språkbanken (VR, kontrakt nr. 2017-00626).

References

- [1] J. Edlund, D. Brodén, M. Fridlund, C. Lindhé, L-J. Olsson, M. Ängsal and P. Öhberg, "A multimodal digital humanities study of terrorism in Swedish politics: An interdisciplinary mixed methods project on the configuration of terrorism in parliamentary debates, legislation, and policy networks 1968–2018", in: K Arai (ed.): *Intelligent Systems and Applications: Proceedings of the Intelligent Systems Conference (IntelliSys) 2021*, 2, Springer, Cham, 2022, pp. 435–449.
- [2] B. Johnson, A. Onwuegbuzie and L. Turner. "Towards a definition of mixed methods research", *Journal of Mixed Methods Research*, 1:2 (2007): 112–133.
- [3] VR assessment 2020–05052.
- [4] R. C. Miller, "Varieties of interdisciplinary approaches in the social sciences: A 1981 overview", in: *Issues in Integrative Studies*, 1, pp. 1–37.
- [5] L. Siemens. "It's a team if you use 'reply all': An exploration of research terms in digital humanities environments", *Literary and linguistic computing*, 24:2 (2009), 225–233.
- [6] C. P. Snow, *The two cultures and the scientific revolution*, CUP, Cambridge, 1959.
- [7] J. Brockman. *The third culture: Beyond the scientific revolution*, Simon & Shuster, New York, 1995.
- [8] J. Ingvarsson. *Towards a digital epistemology: Aesthetics and modes of thought in early modernity and the present age*, Palgrave Macmillan, Cham, 2021.
- [9] S. Oberbichler, E. Boroş, A. Doucet, J. Marjanen, E. Pfanzelter, J. Rautiainen, H. Toivonen and M. Tolonen. "Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians". *Journal of the Association for Information Science and Technology*, 73:2 (2021): 225–239.
- [10] P. Snickars, "Google docs, digital humaniora & akademiskt samarbete", in: P-O. Erixon and J. Pennlert (Eds.), *Digital humaniora: Humaniora i en digital tid*, Daidalos, Göteborg, 2017.
- [11] D. Berry and A. Fagerjord. *Digital humanities: Knowledge and critique in a digital age*, Polity Press, Cambridge, 2017.
- [12] A. Ekström and S. Sörlin. *Alltings mått. Humanistisk kunskap i framtidens samhälle*. Norstedts, Stockholm, 2012.
- [13] www.gu.se/som-institutet
- [14] P. Öhberg, D. Brodén, M. Fridlund, V. Wählstrand Skärström and M. P. Ängsal. "Unifying or divisive threats? Anxiety about political terrorism and extremism among the Swedish public and parliamentarians", 1986-2020, in: K. Berglund, M. La Mela and I. Zwart (Eds.), *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, pp. 145–158.

- [15] https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/riksdagsordning-2014801_sfs-2014-801/#K3
- [16] D. Brodén, M. Fridlund, P. Öhberg, V. Wählstrand Skärström and Magnus P. Ängsal. "Who's afraid of terrorism? Divise worries and securitarian concerns in Swedish public opinion and parliamentary action 1986–2020", in: A. Rostami and C. Edling and (eds.), *Violent extremism: a Nordic outlook*. Lexington Books, Lanham (fc.), 19 pp.
- [17] D. Brodén, M. Fridlund and P. Öhberg. "Vem räds terrorismen? Kluven oro och säkerhetsivrare i Sverige", in: U. Andersson, A. Carlander, M. Grusell and P. Öhberg (eds.), *Ingen anledning till oro?: SOM-undersökningen 2020*. SOM-Institutet, Göteborg, pp. 375–385.
- [18] www.westac.se
- [19] <https://swerik-project.github.io/>
- [20] <https://github.com/welfare-state-analytics/riksdagen-corpus>
- [21] K. Bode, *A world of fiction: Digital collections and the future of literary history*, University of Michigan Press, Ann Arbor, MI, 2018.
- [22] J. Guldi. *The dangerous art of text mining: A methodology for digital history*, Cambridge University Press, Cambridge, 2023.
- [23] M. Fridlund, D. Brodén, L-J Olsson and M. P. Ängsal, "Codifying the Debates of the Riksdag: Towards a Framework for Semi-automatic Annotation of Swedish Parliamentary Discourse", in M. La Mela, F. Norén & E. Hyvönen, eds., *DiPaDA 2022: Proceedings of Digital Parliamentary Data in Action (DiPaDA 2022) Workshop Co-located with the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, Uppsala, Sweden, March 15, 2022, CEUR-WS vol. 3133 (Aachen: CEUR-WS.org, 2022), 167–175.

Accessing centuries of documentation - Resources to improve access to Swedish rock art documentation and metadata

Ashely Green^{1,2,3}, Tristan Bridge³, Christian Horn^{1,2}, Siska Humlesjö³, Aram Karimi³, Johan Ling^{1,2}, and Jonathan Westin³

¹ Institutionen för historiska studier, Göteborgs universitet, Renströmsgatan 6, Göteborg, 41255, Sverige

² SvensktHällristningsForskningsArkiv, Göteborgs universitet, Renströmsgatan 6, Göteborg, 41255, Sverige

³ Gothenburg Research Infrastructure in Digital Humanities, Göteborgs universitet, Renströmsgatan 6, Göteborg, 41255, Sverige

Abstract

The archive of rock art documentation maintained by SHFA provides a valuable resource to archaeologists and others who study rock art. The archive includes images of rock art documentation, sites, and the documentation process, from the 17th century to the more recent high resolution 3D recording and visualizations. In the last few years, GRIDH, in collaboration with SHFA, have begun to improve access to the archive through a Django-based solution and new digital resources. In this paper, we introduce the images in the archive, provide details on the new digital resources, and reflect on how the new website will impact data availability and rock art research.

Keywords

Research infrastructure, digital resource, archaeology

1. Introduction

As part of developing core methods for visualising large digital archives, the Gothenburg Research Infrastructure in Digital Humanities (GRIDH) have developed new web resources to display and share SvensktHällristningsForskningsArkiv's (SHFA) digital archive of rock art documentation. SHFA is a research infrastructure at the University of Gothenburg which curates and manages the digital archive of rock art documentation from Sweden and several other countries.

The idea to establish a database to provide digitalised documentation of the rock art in Sweden to the public and the academic community has been around since the early 1990's and the idea gained momentum after the rock carvings in Tanum were inscribed on the UNESCO world heritage list in 1994. Third-party funded projects, including EU funding, studied rock art documentation, developed methods, and began to digitalise older recordings. In 2007, Riksbankens Jubileumsfond granted money which matched funding by the Swedish National Heritage Board (Riksantikvarieämbetet). This funding allowed for the kick-off of SHFA the following year with Ulf and Catarina Bertilsson, Kristian Kristiansen, and Gerhard Milstreu. An inventory study was conducted by Åsa Fredell in 2008 to get an overview of the amount of data that needed to be digitized, where this material was stored, and which rock art sites these covered.

Since then, SHFA has digitised ca. 80000 documents of different kinds of rock art documentation including Indian ink drawings, rubbing, tracings, photographs and more. Where these documents represent portions of a single panel they are aligned and merged into a single image for upload to the SHFA database. In addition, SHFA staff's own documentation projects

Huminfra Conference 2024, Gothenburg, 10-11 January 2024.

✉ ashely.green@gu.se (A. Green); tristan.bridge@gu.se (T. Bridge); christian.horn@gu.se (C. Horn); siska.humlesjo@gu.se (S. Humlesjö); aram.karimi@gu.se (A. Karimi); johan.ling@gu.se (J. Ling); jonathan.westin@gu.se (J. Westin)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

were conducted with a special focus on 3D methods including laser scanning and photogrammetry.

Since the documentation of rock art is an ongoing process, the original projected data that needed to be digitised has grown considerably. In addition, third-party funded projects that SHFA collaborated with generated new documentation in Spain, Portugal, Italy, Denmark, and more which is also aimed to be provided to the public.

To provide this material, a website was established in 2009 based on an older system which categorised entries by originator of the documentation. The data was published under a CC-BY-NC-ND license, which, while being an open access license, is the most restrictive Creative Commons license. To account for the increased amount of material and demand, it was decided to transform the old database into a more modern, topographically sorted system that provided metadata through its folder structure which was completed in 2021. This was used as the start for the current backend, which was developed by GRIDH in 2022, for a new website that was launched in Summer 2023. The digitized documentation is now provided under a less restrictive CC-BY license.

2. General description

The SHFA resource is comprised of a Django backend and a web-based user interface. The image metadata is stored in a database but uploaded and edited in GRIDH's implementation of a Django-based platform – *Diana*. Recorded metadata includes unique identifiers to maintain links to each image, as well as fields that provide researchers with detailed information about the site location, image credits, image attributes, and interpretations of the motifs visible in each image.

The frontend offers content in both Swedish and English, as well as displays in light and dark mode. The main page provides users with a header containing links to documentation followed by a 2-column layout containing a search panel with three search options and a gallery of sample images. In the desktop version, the search panel is visible on the left and the gallery on the right; however, in the mobile version, the search frame is at the top and the gallery below it. The size of each panel is adjustable.

Users are provided with a search guide, description of the archive, and external links to the SHFA homepage and SHFA news page in the header. The search guide details instructions for the three search options as well as the available keywords and datings, and descriptions of the available image types. The about page provides information on the images in the database, the progress of migrating from the previous system, and guidance on how to cite the images.

The search frame has a form input, the *simple search*, where users can input free-text and all content in the API is searched and items where an explicit match is found, or the search term is a substring of the record entry are returned. This search input is also accompanied by buttons which allow users to trigger example searches. The example searches familiarise users with the new interface layout as well as a range of rock art and image styles. Below the simple search are *map display* (visible by default) and *advanced search* components with buttons that toggle between them. The map displays markers of all sites, with the default view randomised on loading. The advanced search allows users to search specific fields – site, author, institution, image type, keywords, and dating. Each input accepts free-text or the user's selection from the autocomplete suggestions. Clicking on a map marker will also trigger a search for the site.

Once a search is executed, the map zooms to the extent of all sites included in the search results. For further information on the sites, hovering over markers on the map will display pop-ups with site identifiers from the relevant heritage registers or the placename stored in the database. Search results are returned and paginated in the gallery with 25 images per page. Hovering over an image thumbnail in the gallery also displays the site identifiers or placename.

The metadata panel is only visible once an image thumbnail is selected and displays the IIF image at the top; however, the image can also be viewed in full screen mode. This image can also be downloaded using the button overlaid on, or to the left of, the image. Below the image, all the available metadata is displayed. For sites in Sweden, the site description is fetched from Fornsök,

the Swedish National Heritage Board's search service for ancient and cultural remains, with a link directly to the site entry in Fornsök.

3. Technical description

3.1. Backend

The image metadata is stored in a PostgreSQL database and the site metadata, including geometry, is linked based on the site identifier and stored in a PostGIS database. GRIDH use *Diana*, their Django-based database coordination system, to make the data accessible to the frontend using by generating REST APIs. The primary returned fields can be described as follows:

- "id": incremental numeric identifier
- "uuid": unique 32-digit alphanumeric identifier
- "iiif_file": storage location of the higher resolution display image which is generated using the International Image Interoperability Framework (IIIF)
- "site": site identifiers and the site location information
- "collection": collection of images based on a common institution, region, or creator
- "author": name(s) of image creators
- "institution": affiliation(s) of image creators
- "year": year image was taken or created
- "rock_carving_object": group of panels or region an image belongs to
- "type": descriptive image type (e.g., photo, 3D visualisation, night photo, etc.)
- "keywords" & "dating_tags": archaeologists' interpretation of the images and motifs

Diana also offers an accessible frontend solution to allow users to upload new data to the database and take advantage of auto-filled data based on existing entries in the database to an extent.

3.2. Frontend

The SHFA frontend uses the Vue3 framework [1] and features a 3-column layout using the Split.js library. The leftmost column showcases the map and search functionalities, the middle column is an image gallery using the Vue Masonry library, and the right-most panel shows a high quality (IIIF) version of the selected image and relevant metadata. The panels from left to right become more specific and enable a workflow where the data can be filtered down from a regional scale to a specific site.

The frontend fetches the images and metadata using the API generated from the Django backend, but the site description is fetched from Fornsök using HTML parsing. The website's language can be toggled between English and Swedish through a button in the top right-hand corner. The Swedish-English text is manually generated using i18n [2] for the general website texts and keywords in the metadata panel. However, the remaining text is stored in the database. The map uses OpenLayers [3] with a colour-corrected OpenStreetMap layer as a basemap and site markers in a WebGL point layer with the interactive pop-ups.

The simple search function searches the site identifiers, carving, keyword, dating, location, image type, author, and institution fields and returns images where the entered text matches a substring of the data in any of the searched fields. Clicking on one of the suggested search buttons adds the text to the search bar and automatically triggers the search. The advanced search function has separate inputs for each field and supports free-text as well as clicking on one of the autocomplete suggestions which are generated using the entered text as a search query to the relevant API. The advanced search function allows users to narrow their search results by searching for a value in multiple fields, e.g., laser scans from the site Tanum 1:1. Clicking on a marker on the map or the 'Search in map view' button also triggers a search for images by site.

Search results are then grouped by image type and the image types are returned in the order defined by the archaeologists at SHFA which highlights the documentation over images of the site or documentation process. Results are displayed as image thumbnails in the gallery panel with a section header indicating the image type group. Clicking on an image in the gallery opens the image and metadata panel.

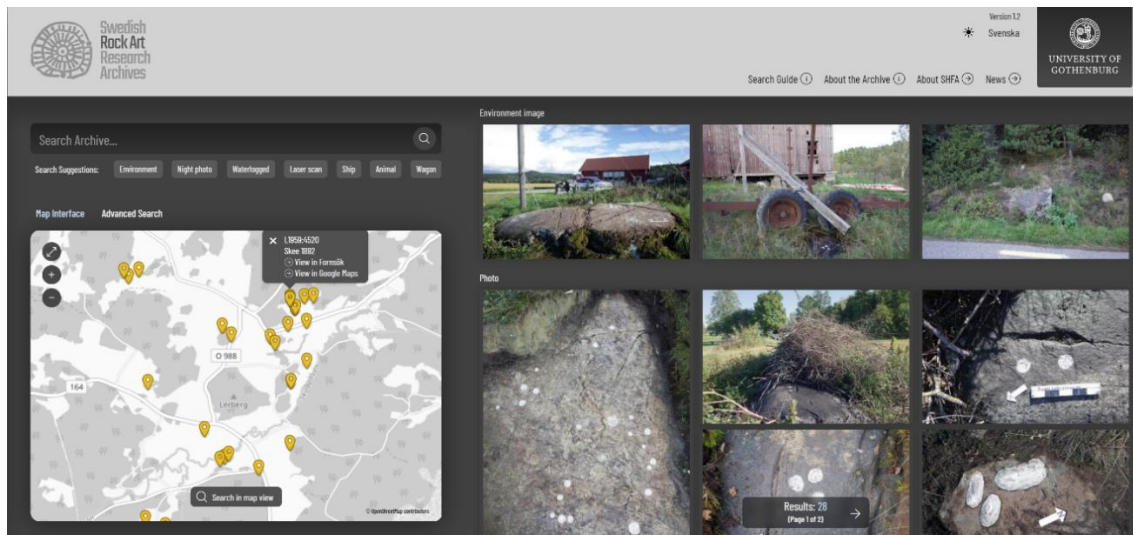


Figure 1: Example of the main page for shfa.dh.gu.se on a widescreen display

The website is responsive, meaning that the layout is preserved across various screen widths and heights. Individuals can access the platform across all browsers and devices while maintaining the same user interface. As demonstrated in Figure 1 and 2, all elements are available on various screen sizes, but the elements are rearranged to improve the ease of use for users. The website's analytics are handled through the open-source library Matomo [4] with custom composites that track searches performed.

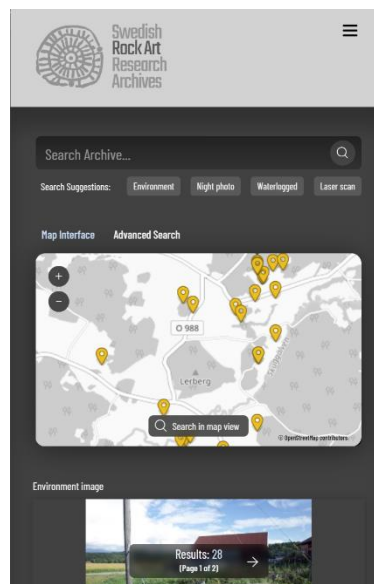


Figure 2: Example of the main page for shfa.dh.gu.se on a narrow display

The IIIF image and controls are rendered with OpenSeadragon [5]. The available metadata is returned below the image in a table. Headers are only returned if data is present in the field. A reference for each image is constructed from the author, year, image type, site identifier, and image id fields with the date accessed calculated from a JavaScript Date instance [6] and

returned language determined by an if/else statement. The keyword and dating lists are concatenated as dating information is not available for all images or motifs. Links to the IIF images and sites can be shared by copying the URL in the address bar.

4. SHFA data and research

While there have been over 11,000 visits to the site in the four months since it became public, it is still difficult to estimate the short-term impact on research in such a timeframe. However, there is data gathered for the old website of the SHFA which can be discussed, with some limitations that must be considered. The data of the use of the website over the last six years only includes visits to the website which did not include any user information. That means it is impossible to say whether the contacts were established by academic users, members of the general public, or other originators. There is also no information about unique users. Gathering user statistics including academic affiliation and more is inherently difficult, because it would require, for example, to establish a login which in turn might diminish the user base because people do not want such information collected or simply because it would make the use of the website much more cumbersome.

A better metric for the impact on research is the use of rock art documentation provided by the SHFA in academic publications. However, even this is not free from problems. The SHFA provides documentation under a creative commons license that requires the user only to acknowledge the originator of the documentation, but not the SHFA as provider. That means SHFA relies on the circumspection and kindness of authors to mention in text, figure descriptions, or acknowledgments where the images they used came from. However, a representative sample of publications were gathered that allow some insights. Between 2017-2022 the SHFA supported 98 academic publications with peaks in 2018 and 2022. Peer reviewed publications outweigh considerably non peer reviewed publications (85%). To assess the quality and rank of publications, the University of Gothenburg uses the ranking provided by the The Norwegian Directorate for Higher Education and Skills. This system ranks publications into three levels (0-2) with the higher number indicating a higher rank. Most publications that were supported were level 1 followed by level 2 publications. The number of supported publications increased clearly in 2022, including the number of level 2 publications.

The SHFA also supports many BA, MA, and PhD theses. Although, it is again difficult to ascertain any concrete numbers. Safer ground provide research projects undertaken by the SHFA driven by and in support of its infrastructure. Currently, the SHFA is involved in six different projects including the Riksbankens Jubileumsfond program “Maritime Encounters”, the Swedish Research Council project “Modelling Bronze Age societies”, and the Swedish Research Council national research infrastructure “InfraVis”. SHFA also supports external projects such as the Swedish National Heritage Board supported project “Digitala bilder för forskning och publik” by Prof. Fredrik Fahlander (Stockholm University).

In 2022, the Riksbankens Jubileumsfond infrastructure for research project “Rock art in three dimensions” ended successfully. The project created the first working image recognition model to successfully semi-automatically identify Scandinavian rock art motifs. In addition, over the course of the project two different ways to better visualize 3D rock art documentation were developed and eventually bundled up into an individual app called Topographic Visualization Toolbox which is open source (<https://tvt.dh.gu.se/>) [7], [8]. Furthermore, two prototypes for mobile apps including augmented reality technologies were programmed to provide information about rock art at home and on-site both [9]. The infrastructure not only supported 20 scientific publications directly and indirectly linked to researchers in the project, but the work also laid important groundwork for the newly launched website and database of the SHFA.

In addition, the SHFA contributed to the national discussion around important cultural heritage including the UNESCO world heritage area “Rock art in Tanum”. This particularly concerns preservation issues through the tradition to paint the rock art in on touristic sites. There

is a long debate about this practice and SHFA co-organised, for example, the “Ren sten” (Clean stone) conferences together with Vitlycke Museum, Riksantikvarieämbete and others to drive the debate forward [10] Preliminary results of a new pilot study indicates that the paint actively contributes to the erosion of the images on the rocks [11].

Overall, the increased user-friendliness of the relaunched SHFA database and especially the website is expected to further be a strong driver for high impact research. A less restrictive CC licence and the suggested citation for each image will allow easier use of the material and the newly implemented map feature makes it easier to find local and transregional comparative rock art. The future implementation of features like viewers for 3D meshes and point clouds will allow users to explore and download material directly, which increases accessibility and researchability of these documentations. This will not only support new publications, but also new research projects advancing our knowledge about those making rock art, and how to protect and communicate rock art. The upload of international material from Norway, Denmark, Italy, Spain, and Portugal from the database to the website will make the new web resources a research resource and a hub for a wider international audience of researchers supporting their work.

5. Discussion and Conclusion

Providing rock art documentation and associated metadata digitally and under a creative commons license using web-based distribution has had many positive aftereffects. Sites are often widely distributed and comparable sites might be separated by several hundred kilometres. Furthermore, the sites are immovable making them disparate even if they are located within a couple of metres from each other. Analogue documentation is only a partial remedy to this situation, because it is generally 1:1 in size and large spaces are necessary to compare the data which may in addition be stored in different locations. The wide availability of digitised rock art documentation has empowered research by increasing the opportunities to compare data not only of different sites, but also subsequent documentations of the same site. In addition, providing data outside of Sweden has enhanced to possibilities for international, diachronic comparative studies which have the opportunity to provide new insights into Bronze Age mobility, travel, exchange networks, and ideologies across Europe. The larger amount of available data also means that more evidence is available for researchers to test their hypotheses. This has rejuvenated rock art research which has experienced an increased research output in monographs, articles, theses, and new projects making important contributions to method development, computational studies including Big Data, and it also allows deeper analysis of social roles that were understudied so far, for example the carvers in Bronze Age societies [10], [11]. By digitising and disseminating rock art documentation through a web-based resource, the data is accessible to the public as well as researchers, allowing for engagement with rock art from multiple perspectives.

The resources developed by GRIDH offer a solution to create future resources where the primary components are images and geospatial data. Using a backend solution that provides the research partners with an adaptable user interface makes uploading data easier for those with a technical background. As viewing the rock art in-person requires a visit to the site, offering a website optimised for both desktop and mobile use makes it easier to view the existing documentation while at a site. User feedback is also considered as we continue to make improvements to the frontend. Data is continuously added to the SHFA website and as migration continues to the new resources, additional data types, such as 3D meshes, will be made available.

Acknowledgements

We are thankful to the many individuals and organisations who have contributed data and feedback to SHFA without whom making such a repository of rock art images would not have been possible. We would also like to thank the other staff of SHFA and GRIDH who provided valuable feedback in the process of developing the new resources. This work received financial

support from Riksbankens Jubileumsfond (grant nos. IN18-0557:1; M21-0018) and the Swedish Research Council (grant nos. 2020-01097; 2020-03817).

References

- [1] Vue.js, 'Vue.js - The Progressive JavaScript Framework | Vue.js'. Accessed: Nov. 09, 2023. [Online]. Available: <https://vuejs.org/>
- [2] i18next, 'i18next: learn once - translate everywhere'. i18next, Nov. 09, 2023. Accessed: Nov. 09, 2023. [Online]. Available: <https://github.com/i18next/i18next>
- [3] openlayers, 'OpenLayers'. OpenLayers, Nov. 09, 2023. Accessed: Nov. 09, 2023. [Online]. Available: <https://github.com/openlayers/openlayers>
- [4] Matomo, 'Matomo - The Google Analytics alternative that protects your data', Analytics Platform - Matomo. Accessed: Nov. 09, 2023. [Online]. Available: <https://matomo.org/>
- [5] openseadragon, 'OpenSeadragon'. openseadragon, Nov. 08, 2023. Accessed: Nov. 09, 2023. [Online]. Available: <https://github.com/openseadragon/openseadragon>
- [6] Mozilla, 'Date.prototype.toLocaleString() - JavaScript | MDN'. Accessed: Nov. 09, 2023. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference/Global_Objects/Date/toLocaleString
- [7] C. Horn, D. Pitman, and R. Potter, 'An evaluation of the visualisation and interpretive potential of applying GIS data processing techniques to 3D rock art data', *J. Archaeol. Sci. Rep.*, vol. 27, p. 101971, Oct. 2019, doi: 10.1016/j.jasrep.2019.101971.
- [8] C. Horn, O. Ivarsson, C. Lindhé, R. Potter, A. Green, and J. Ling, 'Artificial Intelligence, 3D Documentation, and Rock Art—Approaching and Reflecting on the Automation of Identification and Classification of Rock Art Images', *J. Archaeol. Method Theory*, vol. 29, no. 1, pp. 188–213, Mar. 2022, doi: 10.1007/s10816-021-09518-6.
- [9] J. Westin, A. Råmark, and C. Horn, 'Augmenting the Stone: Rock Art and Augmented Reality in a Nordic Climate', *Conserv. Manag. Archaeol. Sites*, vol. 23, no. 5–6, pp. 258–271, Sep. 2023, doi: 10.1080/13505033.2023.2232416.
- [10] U. Bertilsson, C. Horn, and J. Ling, 'Scandinavia and Northern Europe (2015–2019)', in *Rock Art Studies: News of the World VI*, P. Bahn, N. Franklin, and M. Strecker, Eds., Archaeopress Publishing Ltd, 2021. doi: 10.2307/j.ctv1zm2tkx.
- [11] C. Horn, J. Ling, and M. Peternell, 'Bohuslän Rock Art', in *Encyclopedia of Global Archaeology*, Cham: Springer International Publishing, 2021, pp. 1–16. doi: 10.1007/978-3-319-51726-1_3050-1.

Konsten att bedriva svensk ordforskning utan att kränka upphovsrätten

Gerlof Bouma¹, Markus Forsberg¹, Justyna Sikora² and Emma Sköldberg¹

¹Institutionen för svenska, flerspråkighet och språkteknologi, Språkbanken Text, Göteborgs universitet

²KB-labb, Kungliga biblioteket

Abstract

Vi beskriver KB-labb och Språkbanken Texts samarbete för att underlätta ordforskning på de upphovsrätts-skyddade korpusar som finns i Kungliga bibliotekets samlingar. Satsningen har hittills lett till två öppna datasamlingar, Kubord 1 och 2, som ger tillgång till ordstatistik och ordsamförekomststatistik. Vi beskriver även Kubord-fastText, en samling vektormodeller som är baserade på samma korpusar, som är under utveckling.

Keywords

ordforskning, ordvektorer, ordstatistik, lexikografi, tidningstext

1. Inledning

Vid Göteborgs universitet bedrivs det sedan flera decennier tillbaka forskning kring ämnena lexikografi, lexikologi och fraseologi. Digitaliserade textsamlingar – korpusar – har länge spelat en avgörande roll inom denna forskning. Korpusarna är även, och har länge varit, centrala för den ordboksverksamhet som pågår inom Språkbanken Text och som bland annat mynnar ut i Svenska Akademiens samtidsordböcker (SAOL [1] och SO [2]; se vidare [3]). Underlaget för senare upplagor av de aktuella ordböckerna har i huvudsak bestått av redigerade texter i form av tidningstext, men även en del romaner. Av upphovsrättsliga skäl är dock tillgången till digitaliserade tidningar och romaner begränsad och det är givetvis ett bekymmer för de forskare som studerar svenskans ordförråd och för de språkvårdare som till exempel ska ge rekommendationer kring ordval. Ordforskningen behöver tillgång till stora korpusar, med olika slags texter, särskilt moderna texter men också texter från olika tidperioder, för att kunna undersöka ovanliga ord, ords spridning, nya ord och deras utveckling både i betydelse och i form, för att nämna några aspekter. Men bristen på korpusar drabbar inte bara språkforskningen utan också annan humanistisk och samhällsvetenskaplig forskning, även om det inom dessa inriktningar sällan är själva språket som är i fokus, utan det som språket förmedlar.

För att motverka denna situation pågår det nu ett samarbete mellan KB-Labb och Språkbanken Text, som syftar till att tillgängliggöra upphovsrättskyddade texter – särskilt moderna pressmaterial – utan att hamna i konflikt med upphovsrätten, på sätt som främjar forskningen. Arbetet går också ut på att vidareutveckla språkteknologiska metoder som kan stärka den vetenskapliga

Huminfra Conference 2024, Gothenburg, 10–11 January 2024.

✉ gerlof.bouma@gu.se (G. Bouma); markus.forsberg@svenska.gu.se (M. Forsberg); justyna.sikora@kb.se (J. Sikora); emma.skoldberg@svenska.gu.se (E. Sköldberg)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

kvaliteten på ordforskning, och inte minst ordboksarbete. Samarbetet har bland annat resulterat i datasamlingarna Kubord 1 och Kubord 2 som, exempelvis via forskningsverktyget Korp, är fritt tillgängliga hos Språkbanken Text. Därtill är ännu en datasamling, Kubord-fastText, på väg att göras publik. I denna artikel kommer resultaten av det pågående samarbetet att presenteras och diskuteras. Vi kommer även att blicka framåt och resonera kring hur det pågående samarbetet kan fördjupas och stärkas ytterligare i framtiden.

2. Kubord 1, Kubord 2 och Kubord-fastText

2.1. Kubord 1

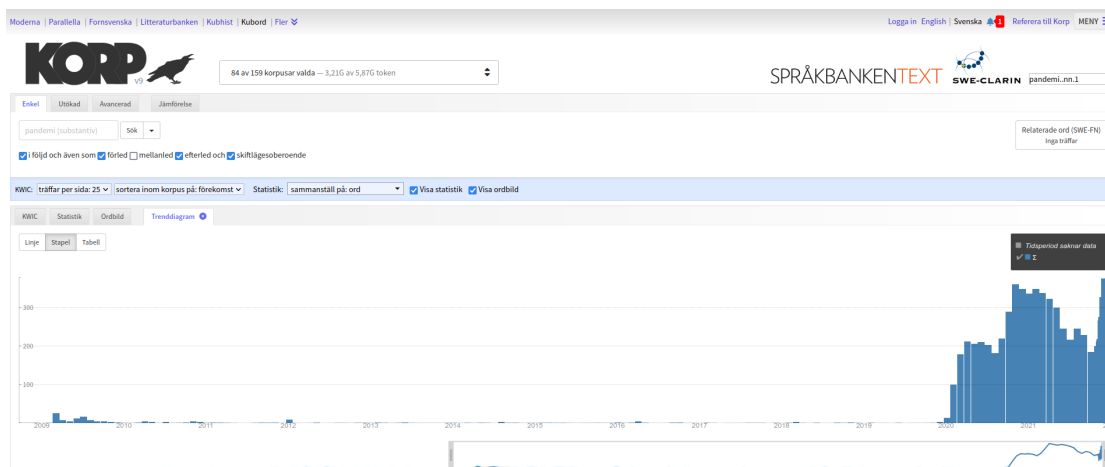
Det första konkreta resultatet av det pågående samarbetet går under beteckningen Kubord 1, en datasamling som finns fritt tillgänglig via Språkbanken Text.¹ Kubord 1 bygger på drygt 80 delsamlingar med olika pressmaterial, sammanlagt lite mer än 3 miljarder tokens. Tidningsmaterialen är publicerade mellan åren 2000–2021 i bland annat morgontidningar (*Dagens Nyheter* och *Svenska Dagbladet*), kvällstidningar (*Aftonbladet* och *Expressen*) och landsortspress (*Östgöta Correspondenten*). Att Kubord 1 enbart innehåller tidningsmaterial kan tyckas begränsande, men tidningar innehåller trots allt många olika slags texttyper. I pressmaterial behandlas också en rad olika ämnen vilket resulterar i en stor spridning (såväl ämnesmässig som stilistisk) bland de ord som påträffas i korpusarna.

På grund av upphovsrätten kan Kubord 1 inte göras sökbar på samma sätt som andra korpusar inom Språkbanken Texts forskningsinfrastruktur, via exempelvis Korp: det går att slå upp ord, men förekomsterna visas inte i vanligt konkordansformat. Det man kan få, däremot, är detaljerad frekvensinformation. Som allt annat textmaterial inom Språkbanken Text är Kubord-materialet försett med metadata som källhänvisningar och är automatiskt berikat med hjälp av Språkbankens analysplattform Sparv [4]. All denna information kan användas för att förfina sökningarna och statistiken. För att kunna analysera ett ord, till exempel bestämma ordets grundform, ordklass eller betydelse, krävs det för det mesta att man har tillgång till den språkliga kontexten. Med andra ord är vi beroende av kontexten i det automatiska analyssteget. Men i och med att KB inte har möjlighet att dela materialet med Språkbanken Text, på grund av upphovsrätten, körs Sparv därför på KBs servrar. När analysen är gjord, tas kontexten bort. Det går därmed inte att ens återskapa textfragment, och på så sätt kan vi tillgängliggöra en förädlad datasamling med ord utan att strida mot upphovsrätten.

För att illustrera vilken sorts information man kan få ut ur Kubord 1, återges resultaten av en sökning i figur 1, där vi tittar på ett ord vars användning har ökat lavinartat de senaste åren, nämligen substantivet *pandemi*. Ordet har använts drygt 86 000 gånger i de aktuella texterna och det används framför allt i bestämd form singularis. En sökning i Kubord 1 visar bland annat också att den vanligaste sammansättningen med *pandemi* som för- och efterled är *pandemiåret*, *pandemilagen* och *pandemitider* respektive *viruspandemi* och *influensapandemi*. I figur 1 visas ett trenddiagram för ordet, inklusive användning i sammansättningar. Ökningen i förekomster syns mycket tydligt med start i 2020, det vill säga, när Covid-19-pandemin bröt ut.

I samband med exempelvis lexikografiskt arbete är det ofta viktigt med den språkliga kontexten.

¹Kubord 1 finns tillgänglig via <https://spraakbanken.gu.se/resurser/kubord>.



Figur 1: Trenddiagram för ordet *pandemi* i Kubord 1.

Det sammanhang som orden används i behövs för att man ska kunna avgöra ords betydelse(r) och fastslå i vilka konstruktioner orden typiskt uppträder i. Källhänvisning finns förvisso i Kubord 1, men det kräver en hel del extra arbete att finna rätt på den faktiska källan för en viss förekomst. Kubord 1:s utformning begränsar därmed givetvis användningsområdena inom ordforskningen. Samtidigt kan forskaren komma en bra bit på väg med frekvensangivelser gällande berikade ordformer. Exempelvis räcker Kubord 1 till att undersöka likheter och skillnader i ordförrådet mellan två årgångar av en och samma tidning. Sådana jämförelser är användbara inom bland annat det nyordsarbete som bedrivs inom ordboksprojektet (se vidare [5, s. 74]). Inom fältet mat och dryck är ord som *streetfood*, *charkbricka*, *gastropub*, *miso*, *pommes*, *prosecco*, *salsiccia*, *syrah* och *teriyaki(sås)* exempel på ord som kommit starkt under senare år. Många av dessa ord, som huvudsakligen är substantiv, säger något om hur svenska ord bildas. De säger även något om vår samtid, till exempel om våra matvanor och hur det omkringliggande samhället förändras över tid. Nya ord som dessa kan antingen bli uppslagsord i kommande upplagor av ordböcker eller fungera som språkexempel, då i form av sammansättningar eller avledningar i aktuella lexikografiska verk (se vidare [6]). Jämförelser i ordförråd går inte bara att göra mellan årgångar av tidningar i Kubord 1:s material, men också mellan material från Kubord och andra material i Korp, exempelvis tagna från sociala medier eller webbsidor.

Innehållet i Kubord 1 fungerar även som stöd vid utmönstring av uppslagsord. Just nu pågår ett arbete med att jämföra ordförrådet i de aktuella dagstidningarna med förteckningen med uppslagsord i SAOL för att de hur väl ordförrådet i ordlistan speglar de ord som faktiskt används i dagspressen av idag.

2.2. Kubord 2

Kubord 1 är, som sagt, begränsat till statistik över enstaka ord. I den andra samlingen som KB-labb och Språkbanken Text har tagit fram, Kubord 2, får man också tillgång till statistik över ordpar som står i syntaktisk relation till varandra, såsom verb-subjekt eller substantiv-attribut.

Moderna | Parallella | Fornsvenska | Litteraturbanken | Kubhist | Kubord | Fler

KORP v9 75 av 159 korpusar valda — 2,65G av 5,87G token

Enkel | Utökad | Avancerad | Jämförelse

pandemi (substantiv) Sök

i följd och även som förted mellanled efterled och skiftlägesberoende

KWIC: träffar per sida: 25 | sortera inom korpus på: förekomst | Statistik: sammanställ på: ord Visa statistik Visa ordbild

KWIC | Statistik | **Ordbild**

pandemi..nn.1 (substantiv)

Preposition	Attribut	pandemi	Efterställt Attribut	Pandemi	verb	Verb	pandemi
1. under	16266	1. global	438	1. av influensa	91	1. hantera	569
2. på grund av	3437	2. framtida	136	2. ha	140	2. bekämpa	197
3. före	1770	3. dödlig	97	3. göra	80	3. klara	201
4. av	6756	4. ny	253	4. till trots	42	4. klara ²	201
5. efter	2157	5. corona	42	5. drabba	57	5. stoppa	148
6. på	3948	6. ond	75	6. med diabetes	30	6. ha	536
7. trots	816	7. allvarlig	45	7. slå	37	7. pågå	106
8. innan	628	8. svår	43	8. pågå	35	8. komma	202
9. mitt i	536	9. aktuell	24	9. vara	78	9. ta	165
10. till följd	407	10. långvarig	17	10. skörda	24	10. överleva	74
11. i och med	123	11. jävla	17	11. på sätt	39	11. möta	98
12. med tanke på	132	12. fullskalig	12	12. härja	22	12. utnyttja	60
13. ur	194	13. historisk	21	13. på allvar	26	13. orsaka	67
14. mitt uppe i	65	14. fler	20	14. lamslå	18	14. tackla	42
15. kring	155	15. eventuell	22	15. i tid	35	15. hejda	39

Figur 2: Ordet *pandemi* i Kubord 2.

Med andra berikas Kubord-orden därmed med viss kontextuell information. Även Kubord 2 är fritt tillgänglig via Språkbanken Text.² Samlingen är baserad på i princip samma tidningsmaterial som Kubord 1, och har försetts med samma metadata och språkteknologiska analys.

Informationen om grammatiskt relaterade ordpar möjliggör så kallade *ordbilder* för de ord som förekommer i korpusarna. Ordbildsvisningen är ett mycket användbart verktyg inte minst när ordforskaren vill undersöka en lexikal enhets typiska medspelare i en sats. Låt oss åter ta substantivet *pandemi* som exempel. I de pressmaterial som ingår i Kubord 2 används detta substantiv som sagt inte mindre än drygt 86000 gånger. Med hjälp av ordbildsvisningen kan forskaren skapa sig en överblick över alla dessa fall och lättare se mönster i hur ordet brukar användas. Visningen effektiviserar och förbättrar därmed analysarbetet avsevärt (se också [5,

²Kubord 2 finns tillgänglig via <https://spraakbanken.gu.se/resurser/kubord2>

s. 76]).

Ordbilden för *pandemi* i Kubord 2 visas i figur 2. Av ordbilden i figuren framgår bland annat att ordet *pandemi* preciseras av adjektiv som *global*, *dödlig* och *ny*. Vidare står ordet som objekt till verb som *hantera*, *bekämpa*, *klara*, *stoppa* och *pågå*. Ordbilden visar också att ordet förekommer som subjekt till verb som *slå till* och *bryta ut*.

Ett av de aktuella verben som brukar uppträda tillsammans med *pandemi* är alltså partikel verbet *bryta ut*. En ordbildssökning i Kubord 2 på just det ordet tydliggör i sin tur att detta verb har flera betydelser (se vidare bland annat [5, s. 77] om värdet av ordbilder vid identifiering av ords olika betydelser). För det första kan *bryta ut* utgöra en synonym till ‘lösgöra ur en större helhet’. Ett återkommande objekt till ordet är då substantivet *del*. För det andra kan verbet betyda ‘inleda’, ‘starta’. Verbet uppträder då tillsammans med subjekt eller objekt som *brand*, *krig*, *slagsmål*, *protest*, *strejk* och *orolighet*. Som synes rör det sig ofta om negativt laddade ord. Ett objekt som avviker i ordbilden är *jubel*. Verbet *bryta ut* kan då sägas betyda ‘brista ut’.

En uttömmande beskrivning av såväl *pandemi* som *bryta ut*, dess betydelser och kombinatoriska drag, ska sålunda ge information om bland annat detta. Uppgifter om hur de aktuella orden kombineras med andra ord är mycket viktiga uppgifter i ordböcker, i synnerhet sådana som vänder sig till inlärare av ett språk. I definitionsordboken SO har beskrivningen av olika slags ordkombinationer nått långt men den kan onekligen förbättras, inte minst genom att fler återkommande ordkombinationer läggs till (se vidare bland annat [7]). I samband med det arbetet kommer användningen av ordbildsvisningen i Kubord 2 att spela en viktig roll.

2.3. Kubord-fastText

Hittills har vi uppehållit oss kring Kubord 1 och 2 som redan är öppet tillgängliga via Språkbanken Text. Detta avsnitt handlar om Kubord-fastText, som är en resurs under utveckling.

Inom projektet *Svenska Akademiens samtidsordböcker* har vi tidigare genomfört pilotstudier med ordvektorer, som bland annat kan förse lexikografer med ord vars språkliga kontexter liknar varandra, och utforskat hur användningen av sådana kan berika det lexikografiska arbetet. Av [8] framgår metodens användbarhet i lexikografiska sammanhang på flera sätt. Ordvektorerna kompletterar de uppgifter lexikograferna får fram med hjälp av konkordanser och ordbilder. Bland grannarna i de vektorrymder som granskas återfinns till exempel ofta semantiskt besläktade ord till det ord som studeras. Dessa grannord kan ligga till grund för tillägg av fler hänvisningar till synonyma, antonyma och kohyponyma ord inom ordboksartiklarna. Metoden kan därmed, på ett förhållandevis objektivt och datadrivet sätt, förtydliga kopplingar mellan befintliga uppslagsord i ordboken och sådana som läggs till i samband med en revidering. Bland grannarna finns det också många sammansättningar som kan tjäna som morfologiska språkprov i ordboksartiklar. Därtill kan ordvektorer ringa in olika slags semantiska fält och ge information om ords värdeladdning.

För att konkretisera återvänder vi en sista gång till substantivet *pandemi*. Bland grannarna i vektorrymden till detta ord finner man ord som *coronapandemi*, *covidpandemi*, *pandemiår*, *pandemivåg* och *pandemiläge* (som alla innehåller det aktuella ordet), men också andra ord som är något mer avlägsna (till form och/eller innehåll) men ändå klart relaterade till det aktuella substantivet och den samhälleliga krissituation som pandemin orsakade. Exempel är *epidemi*, *hälsokris*, *folksjukdom*, *covidrestriktioner*, *lockdown*, *smittspridning*, *smittovåg*, *platsbrist* och *vaccinbrist*.

Ett problem i samband med dessa pilotstudier har dock varit en bristande kontroll över vad som utgör ett ord i vektorrymden, vilket begränsar användningen och därtill ger en hel del brus i vektorrymder. Vidare är det önskvärt att ha ökad kontroll över vilka korpusmaterial som ordvektorerna baseras på. Inte bara för att det är centralt att ha tillgång till källhänvisning, utan även för att kunna jämföra ordvektorer över exempelvis tid och material. Att försöka ta sig an dessa problem har varit centralt i arbetet med Kubord-fastText. Ett viktigt mål för vår pågående utredning är att få fram en väl underbyggd rekommendation för hur vektorrymderna ska vara beskaffade för att vara så användbara som möjligt.

Som namnet redan anger använder vi oss av fastText [9] för att skapa våra vektorrymder, vilket är en metod som skapar vektorrymder där varje ord representeras utifrån dess delar. Detta står i kontrast till metoder som alltid behandlar ord som en helhet, se till exempel [10] som använder sig av en sådan metod i ett lexikografiskt sammanhang. Användningen av fastText möjliggör hantering av ord som inte har observerats i träningsdatan, så länge delar av ordet har blivit det, vilket är en viktig egenskap för svenska språket med sin rika produktion av sammansättningar.

3. Sammanfattning och framåtblick

I den här artikeln har vi i korthet redogjort för resultaten av ett pågående samarbete mellan KB-labb och Språkbanken Text. Arbetet går ut på att, inom ramarna för upphovsrättsskyddet för olika källmaterial, utveckla nya datasamlingar som är av största möjliga nytta för svensk ordforskning.

De samlingar som har tagits fram har begränsningar, bland annat i och med att ordens kontexter inte visas upp, men samlingarna utgör trots det ett viktigt bidrag till ordforskare. Inte minst har arbetet kring dessa samlingar höjt den vetenskapliga kvaliteten på det nyordsarbete som bedrivs inom projektet Svenska Akademiens samtidsordböcker vid Göteborgs universitet. Samarbetet mellan KB-labb och Språkbanken Text banar också väg för metodutveckling på mer generell nivå. Exempelvis har det redan resulterat i studier av ordvektorers roll inom lexikografin.

I nuläget sträcker sig Kubord-materialen fram till och med år 2021, men materialen kommer att kompletteras med nyare pressmaterial allt eftersom, vilket är viktigt för exempelvis ordforskare med fokus på de ord som tillkommit eller vunnit terräng på senare år och hur dessa används.

4. Efterord

Detta arbete har möjliggjorts av *Nationella språkbanken* och *HUMINFRA*, båda finansierade av Vetenskapsrådet (20182024, kontrakt 2017-00626; 20222024, kontrakt 2021-00176) och deras samarbetsorganisationer samt av projektet *Svenska Akademiens samtidsordböcker*, finansierat av Svenska Akademien.

Referenser

- [1] Svenska Akademiens ordlista, 14 ed., 2015. Tillgänglig via <https://svenska.se>.
- [2] Svensk ordbok utgiven av Svenska Akademien, 2 ed., 2021. Tillgänglig via <https://svenska.se>.
- [3] S.-G. Malmgren, E. Sköldberg, The lexicography of Swedish and other Scandinavian languages, *International Journal of Lexicography* 26 (2013) 117–134.
- [4] M. Hammarstedt, A. Schumacher, L. Borin, M. Forsberg, Sparv 5 User Manual, Technical Report, Göteborg, 2022.
- [5] A. Kilgarriff, Using corpora as data sources for dictionaries, in: H. Jackson (Ed.), *The Bloomsbury Handbook of Lexicography*, Bloomsbury Academic, London, 2013, pp. 71–88.
- [6] E. Sköldberg, Hur fångar vi upp svenskans nya ord med hjälp av kubord, *Språkbanksbloggen*, 2022. Tillgänglig via <https://spraakbanken.gu.se/blogg/20221128-hur-fangar-vi-upp-svenskans-nya-ord-med-hjalp-av-kubord>.
- [7] E. Sköldberg, Phraseological theory, evidence in corpora and lexicographical practice: on collocations in a monolingual dictionary of Swedish, in: K. Blenselius (Ed.), *Valency and constructions. Perspectives on combining words*, number 46 in Meijerbergs arkiv för svensk ordforskning, Meijerbergs institut, 2022, pp. 155–182.
- [8] M. Forsberg, E. Sköldberg, Ordvektorer i lexikografiskt arbete, in: E. Volodina, D. Dannélls, A. Berdicevskis, M. Forsberg, S. Virk (Eds.), *Live and Learn. Festschrift in honor of Lars Borin*, Institutionen för svenska, flerspråkighet och språkteknologi, Göteborg, 2022, pp. 37–41.
- [9] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics* 5 (2017) 135–146. URL: <https://aclanthology.org/Q17-1010>. doi:10.1162/tacl_a_00051.
- [10] N. H. Sørensen, S. Nimb, Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings, in: J. Čibej, V. Gorjanc, I. Kosem, S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, Ljubljana University Press, Faculty of Arts, Ljubljana, Slovenia, 2018, pp. 819–826.

Linköping Electronic Conference Proceedings 205
ISBN 978-91-8075-512-2
ISSN 1650-3686
eISSN 1650-3740
<https://doi.org/10.3384/ecp205>