DHNB 2022 CONFERENCE

Digital Humanities in the Nordic and Baltic Countries 6th Conference

UPPSALA 15–18 MARCH 2022

BOOK OF ABSTRACTS

Karl Berglund, Matti La Mela, Inge Zwart & Clelia La Monica (eds.)
CDHU/Department of ALM, Uppsala University, March 2022

DHNB

UPPSALA UNIVERSITET

Contents

(Alphabetically by first author)

A-H

Anne Agersnap[1], Kirstine Helboe Johansen[2], Katrine Baunvig Frøkjær[1]

The legacy of the Danish 'church fathers' N.F.S. Grundtvig and S.A. Kierkegaard in 11,955 contemporary sermons

Keywords *Contemporary sermons, Grundtvig, Kierkegaard, word embedding, associative patterns*

Affiliation 1: The Grundtvig Study Center, Aarhus University, Denmark
2: Department of Theology, Aarhus University, Denmark

Contribution short paper

Abstract N.F.S. Grundtvig (1783-1872) and S.A. Kierkegaard (1813-1855) are important theologians in Denmark. They lived through the turbulent process of Danish mid-nineteenth-century nation building holding the establishment of the Evangelical-Lutheran Church in Denmark (ELCD) in 1849 as a pivotal mark. Though contemporaries, Grundtvig and Kierkegaard have come to represent two differing theological stances and clerical mentalities that have influenced modern Danish church history each in their own way. Grundtvig was one of the key figures behind the establishment of the ELCD in 1849, and his theological legacy has been important in Danish church life since then – in particular through his huge contribution of hymns to the official Danish hymnal. Kierkegaard, on the other hand, is known for his philosophical and theological pondering over the human condition of the individual in Modernity; his position is often categorized as proto-existentialistic. To him, faith was a matter of individual choice. Despite their prominent status in the ELCD, Kierkegaard's and Grundtvig's influence on the theology practiced in the church today has never been subject to a thorough and systematic investigation. It is the aim of this paper to begin to remedy this fact. Accordingly, we present an analysis of a text corpus of 11,955 Danish sermons and investigate how Grundtvig and Kierkegaard are interpreted and associated with the Christian conceptual triad of faith, hope and love. The study thus scrutinize the interpretative links between Grundtvig and Kierkegaard, particularistic authorities of Danish church life, and universalistic features of Christianity in the form of central concepts.

The sermon corpus is composed of sermons from 2011-2016 written by 95 pastors in the ELCD, and metadata includes time labels and basic sociodemographic labels. The sermons are sampled opportunistically, but are dispersed in terms of geography, age and gender of the pastors.

We deploy word embedding analysis to map the semantic fields of seed terms at different complexity levels. In the first analysis, we generate and interpret the semantic networks of Grundtvig and Kierkegaard respectively. In the following analysis, we observe these networks of Grundtvig and Kierkegaard in relation to the semantic networks of the seed terms faith, hope, and love. Measuring the core networks of the theologians against the triad networks unveils subtle, but significant variations between Grundtvig and Kierkegaard in the network structures. These variations entail that the Grundtvig cluster and the Kierkegaard cluster are oriented differently towards the triad network; that significant terms in the triad networks vary in their proximity to the clusters of either Kierkegaard or Grundtvig; and that certain terms originating from the networks of triad seeds are absorbed to the clusters of either Grundtvig or Kierkegaard.

Niko Marius Kustaa Aho[1], Anna Ristilä[1], Reima Välimäki[1], Mila Oiva[2], Aleksi Vesanto[1]

Text re-use in large internet data: workflow and BLAST optimization

Keywords *Text re-use, BLAST, Workflow, Pseudohistory*

Contribution short paper

Affiliation 1: University of Turku, Finland
2: CUDAN, Tallinn University

Abstract The paper presents the workflow of analyzing text re-use in a large and heterogeneous dataset of online discussions and webpages. The study was conducted in the research project “The Ancient Finnish Kings: a computational study of pseudohistory, medievalism and history politics in contemporary Finland and Russia”. The analysis was done with a modified version of BLAST (Basic Local Alignment Search Tool), a program traditionally used in bioinformatics for calculating and finding similarities between biological sequences. The version used in this study has been redesigned to efficiently detect text reuse in input data and to form clusters of these cases (Vesanto et al. 2017).

The project utilized BLAST with two separate sets of online data, one in Russian (300,000 texts) and the other in Finnish (24,500,000 texts). The texts were gathered from web pages and online discussion forums, as well as from existing databases provided by academic institutions. The data included material from Finnish Suomi24, Homma and Tiede forums and news articles from Russian Integrum portal. The resources for running the program were provided by CSC (Finnish IT center for science). Due to the program’s high need for processing power, it was run in batches on multiple cores on CSC’s Puhti supercomputer.

The program found a total of 955,000 clusters of reuse within the materials. The clusters were used to create a database with a search engine, which was then utilized by the project members to study the circulation of pseudohistorical texts online.

Working with Text reuse BLAST on large quantities of web data posed some challenges. Therefore measures needed to be taken to filter the data and generally optimize the process. The goal of the paper is to document the workflow of preprocessing and analysing online discussions, as well as discuss ways to optimize BLAST processing with such materials. We also present some challenges specific to text data from online sources, such as instances of text reuse resulting purely from the metadata of the scraped websites, e.g. menu panel texts.

Bibliography

Vesanto, A., Nivala, A., Rantala, H., Salakoski, T., Salmi, H., & Ginter, F. (2017). Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910. Proceedings of the 21st Nordic Conference of Computational Linguistics., 54–58. http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf

Karolina Andersdotter[1], Malin Nauwerck[2]

Secretaries at work: accessing Astrid Lindgren's stenographed manuscripts through expert crowdsourcing

Keywords *Citizen science, expert crowdsourcing, hackathons, Astrid Lindgren, shorthand*

Contribution long paper

Affiliation 1: Åbo Akademi University, Uppsala University Library;
2: Swedish Institute for Children's Books, Uppsala University

Abstract The digitisation of GLAM collections has made large parts of our digital heritage available online. However, the collections have often been difficult to access in a meaningful way, still requiring manual handling of digital images of text to decipher manuscripts and printed materials due to e.g. limited OCR/HTR capabilities or insufficient metadata. While these technologies are under rapid development, in some cases they require training data to learn both machine and handwritten texts, and in some cases it just makes more sense to transcribe texts manually.

Enter crowdsourcing: a method where a crowd of people is involved to transcribe, describe or otherwise enrich the digital heritage collections with data [1]. However, the labour and cost efficiency of crowdsourcing in a cultural heritage context has been questioned [2] – is the quality of the crowdsourced results worth the investment in launching and running a crowdsourcing project?

Ongoing project the Astrid Lindgren Code [3] explores Swedish author Astrid Lindgren's original manuscripts in shorthand/stenography. Lindgren's stenography has for long been considered "undecipherable" [4, 5] and never been subject to research, making manual interpretation the only existing possibility of accessing the material as well as providing training data for future research [6].

Nevertheless, crowdsourcing has proven to be unexpectedly successful in producing transliterations of Lindgren's stenographed notepads. With 170 volunteers signing up for decoding, prolific attempts during spring 2021 have resulted in a full transliteration of the drafts to novel The Brothers Lionheart (1973) in only a couple of weeks.

Transliterating stenography is a particular skill, situated in time and associated with the profession of the former secretary. While substantially limiting a recruitable crowd of volunteers, the expert skill required has also been central in the success and the methodological development of the project.
This paper presents the method development securing this successful crowdsourcing process, focussing on the importance of joint ownership, planned communication efforts, and community building through online hackathons, as well as how the particular circumstances of a pandemic year might have had an impact on the avid response from a crowd which during a normal scenario might have lacked the confidence and time to participate.

Bibliography

[1] M. Terras, Crowdsourcing in the Digital Humanities, in: S. Schreibman, R. Siemens, J. Unsworth (Eds.), A New Companion to Digital Humanities, Wiley Blackwell, Chichester, 2015, pp. 420–438. doi: 10.1002/9781118680605.ch29.

[2] T. Causer, K. Grint, A.-M. Sichani, M. Terras, 'Making such bargain': Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription, Digit. Scholarsh. Humanit., Jan. 2018, doi: 10.1093/llc/fqx064.

[3] M. Nauwerck, Riksbankens jubileumsfond, Astrid Lindgren-koden: Astrid Lindgrens stenograferade originalmanuskript genom digital bildanalys, genetisk kritik, bok- och mediehistoriska perspektiv (dnr: P19-0103:1), 2020. URL: https://www.rj.se/anslag/2019/astridlindgren-koden-astrid-lindgrens-stenograferade-originalmanuskript-genom-digital-bildanalysgenetisk-kritik-bok—och-mediehistoriska-perspektiv/.

[4] V. Edström, Astrid Lindgren och sagans makt, Rabén & Sjögren, Stockholm, 1997.

[5] L. Törnqvist, Rapport : Astrid Lindgrens arkiv – nya forskningsmöjligheter, Barnboken Tidskr. för barnlitteraturforskning 34 (2011) 59–67.

[6] R. Heil, M. Nauwerck, A. Hast, Shorthand Secrets: Deciphering Astrid

Lindgren's Stenographed Drafts with HTR Methods, in: D. Dosso, S. Ferilli, P. Manghi, A. Poggi, G. Serra, G. Silvello (Eds.), Proceedings of the 17th Italian Research Conference on Digital Libraries, Padua, Italy (virtual event due to the Covid-19 pandemic), February 18-19, 2021, pp. 169-177. URL: http://ceur-ws.org/Vol-2816/short5.pdf.

Everita Andronova[1], Anna Frīdenberga[2], Lauma Pretkalniņa[1], Renāte Siliņa-Piņķe[2], Elga Skrūzmane[2], Anta Trumpa[2], Pēteris Vanags[2]

User-friendly search possibilities for the early Latvian texts: challenges posed by automatic conversion

Keywords *historical corpus, conversion of old spelling into modern, replacement algorithms*

Contribution short paper

Affiliation 1: Institute of Mathematics and Computer Science, University of Latvia;
2: The Latvian Language Institute, University of Latvia

Abstract Diachronic corpora are of high importance not only for linguistic research but also for those interested in other fields of humanities (literature, history, sociology, etc.). Historical spelling is a considerable obstacle for broader use of the Corpus of early written Latvian texts (senie.korpuss.lv). The task of providing user-friendly search possibilities in the Corpus has been put forward recently.

The scope of the Corpus is Latvian texts from the beginning of the written tradition in the early 16th century till 1800. Launched in 2003, it is still being supplemented with new sources. At the end of 2021 its volume is 1.1 mil. running words. The original texts were mostly printed using Fraktur and other blackletter typefaces, but in the corpus, they are presented in Latin transliteration.

The sources are also available in Unicode format, used for a new corpus version in the NoSketchEngine platform (http://nosketch.korpuss.lv/#dashboard?corpname=senie_unicode) released in 2022. These Unicode files now serve as input data for normalisation of the texts into Modern Latvian orthography, which is intended for easier search and comprehension of the corpus data. This task presupposes not only transliteration, but also morphological adaptation of Old Latvian spelling to the modern one. Both procedures can be facilitated and accelerated by elaboration of certain rules of automatic conversion. The paper presents the methodology followed and describes the main types of conversion rules: 1) unambiguous graphemic correspondences, 2) positional (graphemic and morphemic)

correspondences and 3) individual (lexical) correspondences.

Our recent experience is based mostly on the texts from beginning of the 16th c. till early 1680s. These texts are characterized by the greatest amount of spelling variation, and thus they hopefully cover most of the potential problems.

Straightforward grapheme-to-grapheme conversion covers only a tiny number of cases (e.g. v > u as in vnd > und). Conversion of a combined grapheme to a single letter is quite common: ah > ā, tſch > č (opposite cases are also found: x > ks). Conversion of duplicated graphemes into one (the use of duplicated consonants to denote shortness of the previous vowel, like in pills ‘a castle’) has a number of exceptions where real duplication occurs in modern spelling (like elle ‘hell’ or cases of phonetic variants like pills – liter. pilns ‘full of’). Needless to say, one grapheme may stand for different phonemes (y – /i:/, /ij/) and one phoneme can be represented in different ways (/c/ – cz, tz, ts, ds). An important challenge is the order of rules in the series of replacements.
Every source of the first period requires an individual conversion table. There is little sense in writing grapheme-to-grapheme rules for small prayers due to idiosyncratic spellings. Even sources written by one author (such as the texts of G. Mancelius) cannot be converted with the same conversion algorithms.
The methods described will help to make early Latvian texts more accessible to scholars of the humanities.

The modernization of the Corpus of Early written Latvian is supported by the grant ‘Resources of digital humanities: integration and development’ (VPP-IZM-DH-2020/1-0001).

Federico Aurora, Andrea Alessandro Gasparini, Alessandro Palumbo

ENCODE - Bridging the <gap> in ancient writing cultures, enhance competences in the digital era

Keywords *TEI, Cultural Heritage, Digital competencies, Teaching, Linked Open Data*

Contribution short paper

Affiliation University of Oslo, Norway

Abstract The talk will present the international project "ENCODE Bridging the <gap> in ancient writing cultures, enhance competences in the digital era" (https://site.unibo.it/encode/en/). The study of ancient texts through philology, palaeography and epigraphy are increasingly embracing the use of digital tools and related methods to approach their objects of study. However, these innovations require new competences and training both for students and

researchers. Our project's main aims are thus to promote further the use of digital methods and tools in the preservation and study of ancient written texts and to establish a collaborative and shared platform for the teaching and learning of the relevant competences.

The project started in September 2020 and will last until January 2023. It is structured around a series of conferences and training workshops organized by the participant institutions (University of Bologna, Parma, Würzburg, Leuven, Hamburg and Oslo). These events are open both to member institutions and to external participants. The workshops are meant to give participants either introductory or advanced training in the use of different sets of digital methods and tools (e.g. text encoding with EpiDoc, Linked Open Data, digital critical editions, crowdsourcing, relational databases, AI) applied to the study of ancient texts (e.g. inscriptions, papyri, manuscripts, etc.) across a number of languages and cultures (e.g. Ancient Greek, Latin, Egyptian, Ethiopic, Runic inscriptions). The conferences will be an occasion for reflection and exchange of experiences, with a special focus on learning and teaching.

The main objective of all these activities, however, is to give theoretical material and practical experience to support the ultimate goals of the project:
1) The surveying of existing competences and practices in the teaching of digital methods applied to the study of pre-modern texts.
2) The formulation of a shared definition of digital competencies needed by academic staff and students involved in programmes focusing on written cultural heritage.
3) The design and testing of teaching modules, both basic and advanced, paired with guidelines and examples of best practices for their employment at other universities, making the modules customizable and transferable.

More concretely, the project will result in the creation of an online set of Open Access resources: a reviewed, reasoned and updated database of teaching modules (basic and advanced, 5-10 ECTS) – completed with detailed guidelines and an introductory MOOC – and a network platform of researchers, students and institutions.

Anda Baklāne, Valdis Saulespurēns

Detection of the target domain of conceptual metaphors in Latvian poetry: exploring the effectiveness of Word2vec and BERT models

Keywords *Conceptual metaphor, Word2vec, BERT, Latvian poetry*

Contribution short paper

Affiliation National Library of Latvia, Latvia

Abstract Within the framework of the cognitive metaphor theory, a metaphor consists of the source domain and target domain where the source domain is the figurative concept that is used in to express meaning and the target domain is the intended meaning. The automatic detection of metaphor can be approached in various ways: it can be aimed at the source domain or the target domain of the metaphor. The text can be analyzed to identify which expressions are used metaphorically (source domain) or attempts can be made to detect and interpret the intended meanings (target domain). The Contribution of this study have explored the effectiveness of Word2vec and BERT (Bidirectional Encoder Representations) models to identify the presence of target domain concepts in the corpus of 20th century Latvian poetry (1920-1999; 400 poetry books). In particular, the testing has been aimed at the conceptual metaphors related to temperature, such as STRONG EMOTIONS ARE HEAT, VITALITY IS HEAT, PASSION IS HEAT, DEATH IS COLDNESS, LONELINESS IS COLDNESS. Several source domain terms (such as ‘fire’, ‘ice’) have been manually annotated in the corpus and provide additional basis for validating and interpreting the results. It is not expected that the concepts of target domains should be detectable in the majority of poetic texts since the target meaning is often just implied, not stated, however, it is relevant to ask how often the target domain is actually expressed. Further inquiry into discerning different target domains related to the same source domain is being tested by Contribution by using the methods of text classification. In the past, the usage of topic modeling approaches have not proved to be highly effective for detecting topics in poetry, especially for the short lyric forms; the testing of naïve LDA based on bag of words approach to analyze the corpus of Latvian poetry seems to confirm this conclusion. The usage of vord2vec algorithm has provided more promising results in demonstrating that associations between broader networks of the source domain terms and some target domain terms are being formed. The most promising preliminary results have been gained by using a pre-trained BERT model which includes Latvian laguage corpus. Further gains can be achieved by fine tuning said BERT model.

Katrine Frøkjær Baunvig[1], Kristoffer Laigaard Nielbo[2]

Mermaids are birds. Text Mining N.F.S. Grundtvig's Bestiary

Keywords *Text mining; Word embedding; N.F.S. Grundtvig*

Contribution long paper

Affiliation 1: Center for Grundtvig Studies, Aarhus University, Denmark;
2: Center for Humanities Computing, Aarhus University, Denmark

Abstract In a classic study of Nuer religion British Anthropologist E.E. Evans-Pritchard explored the problem of religious symbols embedded in the Nuer metaphor 'twins are birds' (Evans-Pritchard 1974 [1956], 123-143). In this paper we will present a study concluding that not only twins but also mermaids are to be understood as birds. At least this is how they semantically behave in the lexical habitat of the influential Danish romanticist and nineteenth-century poet, pastor, and politician N.F.S. Grundtvig (1783-1872). As in the Nuer case the cause for this behavior is to be found in the symbolic structures of a religious logic.

Within the paradigm of Natural Language Processing and aided by the programming language Python this study consists of a cascade of word embeddings plotting the bestiary arising from Grundtvig's 1068 publications in their tokenized, lemmatized, 'algorithmifyed' avatar. Our interest here lies with exploring the view on animals in an Contributionhip that grew highly influential in the Danish farmers-movement of the nineteenth and twentieth centuries. How are human 'companion species' (Haraway 2003) displayed in this material? And further: What is the role ascribed to the variety of fantastic beasts also roaming the corpus?

This study is based on the digital scholarly edition Grundtvigs Værker enriched by philologists' strenuous mark-up and thus offering a clean, reliable, and flexible corpus-material open to comprehensive and hermeneutically complex explorations. In order to build the catalogue of word embeddings we have made the Grundtvig-algorithm compute the distance between a set of so-called seed terms (i.e., kalv, ko, svin, abe, and, hund, kat, havfrue, fugl, bæst, udyr, engel) and the corpus lexicon. Further: for each seed the algorithm has extracted terms with the shortest distance (primary associations) as well as terms with the shortest distance to the primary associations (secondary associations). The distance between all terms (i.e., seeds, primary and secondary associations) have been computed and terms have been connected based on their distance under a given threshold. Finally, semantic clusters have been extracted using the Louvain method (Blondel et al. 2008) and the visualization is represented generated with terms as nodes and thresholded distances as edges. Among other things, this procedure has, as mentioned, enabled us to conclude that in the eyes of an influential romanticist mermaids are birds.

Bibliography Blondel, Vincent, Jean-Loup Guillaume, Renaud Lambiotte & Etienne Lefebvre 2008 "Fast unfolding of communities in large networks", Journal of Statistical Mechanics: Theory and Experiment, 1-12.

Evans-Pritchard, Edward Evans 1974 Nuer Religion, Oxford University Press [1956], New York and Oxford.

Haraway, Donna 2003 The companion species manifesto: dogs, people, and significant otherness, Prickly Paradigm Press, Chicago.

Ümit Bedretdin, Pihla Toivanen

Supervised multi-class classification for humanities research: An article classifier using BERT and topic modelling

Keywords *computational media studies, language technology, countermedia*

Contribution Poster

Affiliation University of Helsinki, Finland

Abstract Advances in the accessibility of technology have motivated a growing interest in combining qualitative and quantitative methodologies in media studies. Current machine learning methods make it possible to gain insights from large datasets that would be impractical to analyse with more traditional methods. Supervised document classification presents a good platform for combining specific domain knowledge and close reading with broader quantitative analysis.

The poster presents experiments in training a classifier that combines BERT-based contextual embeddings with document topic information derived from a topic model. The approach is tested through a test scenario involving the classifying of Finnish countermedia articles. Our initial results using a BERT-based classifier were promising, however with clear room for improvement still remaining in both classification accuracy as well as efficiency with regard to training set size. In this poster, we thus wish to test whether the performance of the classifier could be further improved by augmenting it with a topic model. The hypothesis that features derived from a topic model could complement features derived through embeddings is plausible, because contextual embeddings contain mainly linguistic and distributional information up to the level of a couple sentences, whereas topic information allows us to introduce document-level structural information. This should be a complementary addition and useful for document-level classification tasks.

We use the Structural Topic Model that has helped researchers achieve good results in many domains of social sciences, from analysis of high court judges' tweets (Curry & Fix 2019) to comparative politics (Lucas et al. 2019).
To test our hypothesis, we combine the contextual sentence embeddings from a pretrained BERT model with topic distributions thus obtained. As test material, we use articles from the Finnish countermedia publication MV-lehti (Toivanen et al. 2021). The data represents a real classification need, considering there has been an influx of terms like "fake news" and "alternative media", that have been used to juxtapose countermedia with legacy news sources in the hybrid-media environment. There is a growing body of evidence to suggest that the implied boundary between truthful and untruthful news may not be as clear as it would seem. Countermedia outlets, instead of publishing strictly made up content, might instead introduce biases more by means of curation of information (Ylä-Anttila et al. 2021).

The dataset consists of 37 185 articles gathered from 2014 to 2018 and includes our training data, which is composed of 1000 randomly sampled and pre-annotated articles. In this work, we experiment training a classifier to classify articles into one of three categories that have previously been identified from the data using frame analysis (Toivanen et al. 2021), as well as contrast this categorization with a more simple binary classification, as well as a recall-focused strategy with manual post-filtering.

Jenny Bergenmar[1], Koraljka Golub[2], Siska Humlesjö[1]

Queerlit database: making Swedish LGBTQI literature easily accessible

Keywords *fiction, bibliographic database, metadata, LGBTQI, subject indexing, linked open data*

Contribution poster

Affiliation 1: University of Gothenburg, Sweden;
2: Linneaus University, Sweden

Abstract How can LGBTQI fictional literature become more accessible to readers and scholars? The project Queerlit Metadata Development and Searchability for LGBTQI Literary Heritage addresses this question in two ways: by the development of a thesaurus for the description of Swedish LGBTQI literature, and by building a curated bibliographical database for this material with flexible search options.

Despite the community and scholarly interest in LGBTQI literature, relevant LGBTQI literature is hard to find both for readers and researchers. Subject

indexing is underdeveloped for this topic, and subject headings have been historically inadequate and offensive. The Queerlit project will provide publicly available tools for subject indexing through the thesaurus, and facilitate access to LGBTQI literature through the database. This will make new research possible, such as gaining overviews of the development of specific themes over time, the presence of LGBTQI literature within or outside of the literary canon or in different genres, and changing ideas and perceptions concerning sexualities and gender identities.

The project has four aims:
1) To develop a subject-specific thesaurus for indexing LGBTQI literature in collaboration with KvinnSam (National Resource Library for Gender Research). We will depart from the currently largest international thesaurus for LGBTQI materials, Homosaurus structured as a linked data vocabulary. Each term will be assessed for applicability and translated to Swedish. Subject to available resources, the Queerlit thesaurus will also be mapped to other subject indexing systems.
2) To identify LGBTQI fiction in collaboration with experts on Swedish LGBTQI literature. A difficult and methodologically challenging question is how to establish criteria for inclusion: how do we distinguish between stable and variable content meanings (Campbell)?
3) Constructing a sub-database in LIBRIS (Swedish Union Catalogue) to contain bibliographic records of fictional LGBTQI literature in collaboration with the Swedish National Library and KvinnSam while using the external thesaurus to describe the material.
4) Making the sub-database available through a separate end-user search interface allowing for more specialized searches than LIBRIS does, and linking the records to relevant open data from external sources.

Work on LGBTQI within the digital humanities is still scarce, and related digital projects mainly concern digitization of material in existing collections, or creating collections of new digital material. Possibly this is a consequence of the emphasis on data mining and empirical studies within the digital humanities, which may be at odds with the human effort required to recognize and understand LGBTQI texts. However, we agree with Ruberg, Boyd, and Howe (2018) who identify subject indexing as a productive place to make queer interventions in the digital humanities.

Bibliography

Campbell, Grant D. “Queer Theory and the Creation of Contextual Subject Access Tools for Gay and Lesbian Communities.” In Feminist and Queer Information Studies Reader, edited by Keilty, Patrick, and Rebecka Dean, 290-308. Sacramento: Litwin Press, 2013.

Homosaurus, https://homosaurus.org

Ruberg, Bonnie, Jason Boyd, and James Howe. “Toward a Queer Digital Humanities.” In Bodies of Information: Intersectional Feminism and the Digital

Humanities, edited by Losh Elizabeth and Wernimont Jacqueline, 108-28. Minneapolis; London: University of Minnesota Press, 2018

Karl Berglund[1,2], Mats Dahllöf [3]

Books as muzak: Tracking streamed audiobook reading by the hour

Keywords *Readership studies, digital book consumption, audiobooks, streaming services, book history*

Contribution long paper

Affiliation 1: Department of Literature, Uppsala University, Sweden;
2: Centre for Digital Humanities Uppsala (CDHU), Department of ALM, Uppsala University, Sweden;
3: Department of Linguistics and Philology, Uppsala University, Sweden

Abstract For many people in the 2020s, especially in the Nordic countries, book reading means listening to a streamed audio file through a smartphone. But reading on a page and listening to an audiobook are distinctly different practices. The change of medium with the rise of streamed audiobooks alters where, when, and how books can be read. This ongoing shift in how books are consumed are transforming reading and publishing practices, but also the possibilities for studying reading and publishing. Access to large-scale data points on private book consumption behaviour enable scholars to answer questions that could only be speculated about earlier.

The purpose of this paper is to study temporal aspects of contemporary digital book consumption at scale and thereby addressing when-questions about digital reading that previous readership studies have nothing or very little to say about. When are people reading audiobooks, and how does it differ from ebook reading? When do book consumption peak? What are there of differences between genres, Contributionhips and book formats regarding temporal patterns? How can the answers to the (descriptive) questions above be (analytically) understood in relation to our contemporary digital book culture?

The empirical foundation is consumer behaviour data from Storytel, one of the key players globally in subscription-based digital bookselling, and the major platform in Sweden. The dataset covers consumption patterns tracked per hour for all individual users in Sweden (ca 450,000), during a one-year period (May 2020 to May 2021), and for three book segments: bestsellers in Sweden 2004–2020 (334 s), the most popular s on the Storytel platform in Sweden 2015–2020 (“beststreamers”, 64 s), and born-audio Storytel Originals (843 episodes, which equals roughly 84 novels/audiobooks in length, since 10 episodes of a season is

usually about the length of a novel). The three corpora comprise 482 s altogether.

The results show distinctly different temporal reading patterns between ebook and audiobook reading, where the latter is much more evenly distributed over the hours of the day and night. Choices of fiction, however, do not seem to affect temporal reading patterns at all. Moreover, certain types of reader clusters are identified based on their temporal reading habits, and discussed in relation to theories on digital reading. Night readers is a category of specific interest, since the dataset shows a correlation between reading quantity and night time reading.

In the discussion, the results are used to problematise the concept of reading as we know it. Platforms for streamed audio enable new kinds of habits that makes the emerging digital reading practices into something quite distant from what is generally meant when talking about reading. “Reading” audiobooks in streaming services seem to involve a whole spectrum of activities – from careful close listening to books used as muzak.

Yngvil Beyer, Andre Kåsen

NorHand: a preliminary dataset for Norwegian HTR (1825-1940)

Keywords *HTR, manuscripts, LAM*

Contribution Poster

Affiliation National Library of Norway

Abstract Cultural institutions in the LAM (libraries, archives and museums) sector hold collections of huge amounts of handwritten documents from the recent and distant past. Unlike printed material, handwritten material is often harder to read at a glimpse and domain expertise is therefore often a prerequisite. Moreover, the emerging technology for automatic text recognition (known as OCR for print material and HTR for handwriting) needs training data for both writing style and language. With the NorHand dataset the National Library of Norway contributes a large set of high quality data. All transcripts have been processed with Transkribus and both the layout analysis and the text recognition have been corrected by human evaluators.

The dataset consists mainly of diaries and letters written by 15 different writers. In total, in its current state (version 0.7), it comprises 632 000 tokens distributed in 91362 lines in 4144 pages. The writers are: Nini Roll Anker, Harriet Backer, Vilhelm Bjerknes, Kristine Bonnevie, Lagertha Broch, Camilla Collett, Hulda Garborg, Knut Hamsun, Ebbe Hertzberg, Kitty Kielland, Henrik Ibsen, Edvard Munch, Petronelle Nielsen, Amalie Skram and Sigrid Undset. The dataset will be published in the Norwegian Language Bank Resource Catalogue (https://www.nb.no/sprakbanken/en/resource-catalogue/).

The language is fairly varied since Norwegian went through quite a lot of changes in the timespan contained in the dataset and the first standardisation of Norwegian was first ratified in 1907. In order to assess the quality and consistency of the recognition and evaluation process, the dataset has continually been used to train new recognition models with the tool PyLaia (Puigcerver 2017). PyLaia is an open-source toolkit for HTR that makes use of deep learning techniques. We used it off-the-shelf i.e. without any hyperparameter tuning. Optimization of parameters will be done in the future when the dataset has reached a mature state both with respect to size and diversity. 175 epochs were trained, the final CER (character error rate) on the training / validation set was: 3.6% / 3.8%.

There are two versions of the dataset; one uncorrected HTR and one corrected (ground truth) text. We are working on some strategies for using these versions for post-processing the HTR using Transformer-based translation engines. Applying data-driven post-editing could possibly redeem some of the errors that occur in the HTR process, and this approach could be an important step towards a general HTR for Norwegian.

Bibliography Puigcerver, Joan. 2017. Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition? ICDAR17

Magnus Breder Birkenes, Lars G. Johnsen, Andre Kåsen

Aligning the past and the present: two approaches to automatic modernization of historical texts

Keywords *sequence alignment, machine translation, language change*

Contribution short paper

Affiliation National Library of Norway

Abstract The study of texts is increasingly harder as the text gets older. This is mostly due to language change, but also, in a mass digitization context the accuracy of optical character recognition (OCR) deteriorates as the text's age decreases, cf. Tanner et al. 2009. One important issue is the variable and historical spelling forms of written texts. While language variation is a general challenge for most

studies of historical sources, it is devastating to quantitative text analysis. In order to cope with such variation, development of resources and techniques that are able to map different word forms to an invariant space is needed. To illustrate, the word for woman has at least three spelling variants in 19th and 20th century Norwegian: Qvinde, Kvinde and Kvinne. Ideally these three spellings would be considered as the same lexeme in an analytical context. Our approaches to this mapping problem are collection-driven. At the National Library of Norway, basically all books published in Norway since 1519 have been digitized. Hence, establishing a historical parallel corpus of e.g. biblical texts or the novels of famous Contribution is possible. In a pilot study, we looked at the collected works of Henrik Ibsen, in an early 20th century and a late 20th century edition, where the former represents Dano-Norwegian and the latter represents mid 20th century Norwegian. With such a corpus as a starting point, we applied different alignment algorithms to induce parallel data.

Before we align words we align sentences. Reimers & Gurevych (2019) present a system for aligning sentences with Transformers, while Jalili Sabet et al. (2020) present a new system for word alignment also using Transformers. For word alignment we try different BERT models for Norwegian (see Kummervold et al. 2021 and Kutuzov et al. 2021). In combination, these two alignment systems yield parallel sentences and words.

The contribution of the present work is two-fold. With the Ibsen parallel corpus of the pilot study we have fine-tuned a Transformer-based machine translation engine as described in Raffel et al. 2020 for automatic modernization. In addition, we have a preliminary mapping of older and newer word forms i.e. a historical dictionary. In future work we will explore which of the two approaches are more favorable to quantitative text analysis.

Bibliography

Jalili Sabet et al.: 2020. SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings. In EMNLP'20

Kummervold et al.: 2021. Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. In NoDaLiDa'21.

Kutuzov et al.: 2021. Large-Scale Contextualised Language Modelling for Norwegian. In NoDaLiDa'21.

Raffel et al.: 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In Journal of Machine Learning Research.

Reimers & Gurevych: 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In EMNLP'19.

Tanner et al.: 2009. Measuring Mass Text Digitization Quality and Usefulness. In D-Lib Magazine 15.

Robert Borges

MoReDaT: A Modular Remote Data-collection Toolkit for Linguistics Research

Keywords *Linguistics, research infrastructures, remote data collection, spoken language*

Contribution short paper

Affiliation Institute of Slavic Studies, Polish Academy of Sciences, Poland

Abstract The COVID-19 pandemic effectively halted data-gathering activities in many disciplines. In linguistics, particularly the area of language documentation, researchers have been unable to conduct fieldwork without risking the health and safety of already vulnerable speech communities. Given that documentation projects generate records of highly endangered languages, the creation of primary-data sets to preserve and potentially revitalize these languages is particularly urgent. Although there is arguably no substitute for in-person fieldwork, COVID-19 has raised many questions about how to effectively employ our now widely-available technological infrastructure for generating data remotely.

Clearly technological developments introduced in response to the pandemic will continue to be useful whether or not COVID-19 ceases to hinder mobility. And at least in some research contexts, the availability of technology and know-how is sufficient to collect primary data remotely or in hybrid form. Recent/ongoing remote data collection in general linguistics can be categorized as either supervised or semi-supervised. Supervised data collection consists of "the usual" data collection activities (elicitation, translation, story telling, etc) over video-conferencing software, e.g. Zoom, (cf. Mannby 2021), or a combination of video conferencing and stimulus-display applications (cf. Leemann et al. 2020). In semi-supervised scenarios, researchers provide direction to individuals in the target-language community, who then return primary data either via existing general technological infrastructure, (e.g. WhatsApp (K. Rybka p.c. 2021)), or via dedicated software (Griscom 2020). Unsupervised data collection methods have also been used for more specific tasks (i.e. dialectology, cf. Hinskens et al. 2021; Hasse et al. 2021), where sentence reading tasks, lexical selection tasks, and matched-guise tasks were crowdsourced via mobile apps. These apps were extremely successful in terms of the amount of data gathered, however, the tasks are overly specific, and furthermore, the apps themselves are (apparently) hidden behind closed-

access licenses. This means that attempts at similar unsupervised data collection need to engage in the expensive undertaking of building infrastructure from scratch. It is clear that the need for generalized, open-access tools is opportune.

MoReDaT is just such a toolkit. It is a modular web app, developed for unsupervised remote collection of linguistic data, written primarily in Python's Django framework. It will be (by the time of the DHNB meeting) available on an Open-Source license and distributed as a fully functional application. In this talk, I will first discuss the theoretical and practical issues surrounding remote data collection in linguistics and situate the development of MoReDaT within its broader research project. I will then present/demo the main modules of MoReDaT, which have been designed (a) to replicate the typical field tasks utilized by general linguists, irrespective of preexisting knowledge on the target language, and (b) to be fully customizable in terms of the the stimuli that can be utilized. I will conclude with a discussion about the promising future of remote data collection in linguistics; not only as a reactionary measure during COVID-19, but in the face of increasingly acute environmental costs and scarce research budgets, 'going (partially) remote' will become an effective means to continually source more data.

Bibliography

Griscom, Richard. 2020. "Mobilizing Metadata: Open Data Kit (ODK) for Language Resource Development in East Africa." In Proceedings of the First Workshop on Resources for African Indigenous Languages, 31–35. Marseille, France: European Language Resources Association (ELRA). https://aclanthology.org/2020.rail-1.6.

Hasse, Anja, Sandro Bachmann, and Elvira Glaser. 2021. "Gschmöis – Crowdsourcing Grammatical Data of Swiss German." Linguistics Vanguard 7 (s1). https://doi.org/10.1515/lingvan-2019-0026.

Hinskens, Frans, Stefan Grondelaers, and David van Leeuwen. 2021. "Sprekend Nederland, a Multi-Purpose Collection of Dutch Speech." Linguistics Vanguard 7 (s1). https://doi.org/10.1515/lingvan-2019-0024.

Leemann, Adrian, Péter Jeszenszky, Carina Steiner, Melanie Studerus, and Jan Messerli. 2020. "Linguistic Fieldwork in a Pandemic: Supervised Data Collection Combining Smartphone Recordings and Videoconferencing." Linguistics Vanguard 6 (s3). https://doi.org/10.1515/lingvan-2020-0061.

Mannby, Emil. 2021. "Linguistic e-fieldwork: how the Lingfil field-methods course survived a pandemic." Presentation given at Fieldwork in Anthropology and Linguistics' event for student fieldworkers dept. of Anthropology and dept. of Linguistics and Philology, Uppsala University. https://www.lingfil.uu.se/digitalAssets/926/c_926886-l_1-k_fal_for-students_1.pdf

Kristin Charlotte Carlsson

Collaboration and the role of a metadata librarian in an e-infrastructure project at the University of Oslo

Keywords *e-infrastructure, metadata librarian, metadata, collaboration, special collections*

Contribution short paper

Affiliation University of Oslo, Norway

Abstract In 2020, the University of Oslo was granted funds to establish the project Special Collections and Databases. The project's goal is to establish a common, digital platform to facilitate archiving and cataloguing of special collections and making them accessible, as well as discipline-specific archives, locally developed databases, text, and audio-visual materials not included in the University library's catalogue.

Among two new positions created for the project was that of a metadata librarian. Part of their role is to coordinate between the needs of the pilot projects and the possibilities in the digital platform when it comes to metadata, and between the platform and the project group in order to determine which solution is the right choice looking forward. An interesting question is whether the role of the metadata librarian in a project like this also could include a more general coordinating aspect, focusing on establishing a good collaboration platform.

The material in the pilots participating in the project (two of which will be detailed in this paper) is quite different in nature and therefore requires different approaches. Some material is neither digitized nor properly organized and is in need of a proper archival structure. Other materials are already wholly or partly digitized, but carries with it other issues such as a need for a good interface for search and access, as well as clarifying issues related to privacy and copyright. A significant amount of time is therefore necessary to establish a good platform for collaboration with clearly defined and tailored roles and workflows. In addition to expertise related to metadata, knowledge of storage and archiving as well as legal expertise is therefore needed. Regular contact with the research community is also vital, in order to capture the future needs of scientists wishing to conduct research on the collection's resources.
This paper aims to give an insight into the special collections e-infrastructure project at the University of Oslo, from the perspective of the metadata librarian and their various collaborators.

Bibliography 1 https://www.ub.uio.no/om/prosjekter/spesialsamlinger-og-databaser/

Steven Coats

The Corpus of British Isles Spoken English (CoBISE): A New Resource of Contemporary British and Irish Speech

Keywords *Corpus linguistics, British English, Irish English, Scottish English, YouTube*

Contribution short paper

Affiliation University of Oulu, Finland

Abstract In recent decades, corpora of transcribed regional speech have provided material for the study of spoken English in the British Isles, giving new insights into local, regional, and national varieties (Anderwald & Wagner 2007; Corbett 2014; Corrigan et al. 2012; Kallen & Kirk 2007; Szmrecsanyi 2013). These and similar resources have proven to be tremendously useful, but in some cases, they are not large enough for the study of rare syntactic or pragmatic phenomena. This paper introduces the Corpus of British Isles Spoken English (CoBISE, https://cc.oulu.fi/~scoats/CoBISE.html), a 112-million-token corpus of 38,680 word-timed, part-of-speech-tagged Automatic Speech Recognition (ASR) transcripts, corresponding to more than 12,801 hours of video, from 495 YouTube channels of local councils or other institutions of local governance in 453 locations in England, Scotland, Wales, Northern Ireland, and the Republic of Ireland. As is the case for the Corpus of North American Spoken English (Coats 2021), many of the transcripts in the corpus are records of public council meetings (see also Coats forthcoming).

The paper summarizes the methods used for data collection, processing, and geolocation of the channels in the corpus. Because CoBISE consists of ASR transcripts, it is “noisy” data, containing errors. Nevertheless, due to its size and the preponderance of accurately transcribed forms, it can be used to extract reliable linguistic signals for a wide range of relatively frequent phenomena. Because the transcripts are from videos viewable by anyone with access to the internet, phenomena of interest can also be examined and manually verified in the corresponding videos—the paper provides an example of how this can be done for a low-frequency syntactic construction. Finally, the structure of the corpus facilitates the creation of corpora of video or audio data with a simple pipeline of download and conversion scripts, opening up the possibility for semi-automated analysis of (for example) acoustic or visual aspects of communication.

Bibliography Anderwald, Lieselotte & Susanne Wagner. (2007). FRED – The Freiburg English Dialect Corpus: Applying corpus-linguistic research tools to the analysis of dialect data. In Joan C. Beal, Karen P. Corrigan & Hermann Moisl (Eds.), Creating and digitizing language corpora volume 1: Synchronic databases, 35–53. Houndmills, Basingstoke: Palgrave Macmillan.

Coats, Steven. (Forthcoming). Dialect corpora from YouTube.

Coats, Steven. (2021). Corpus of North American Spoken English (CoNASE). https://cc.oulu.fi/~scoats/CoNASE.html.

Corbett, John. (2014). Syntactic variation: Evidence from the Scottish Corpus of Text and Speech. In Robert Lawson (Ed.), Sociolinguistics in Scotland, 258–276. Houndmills, Basingstoke: Palgrave Macmillan.

Corrigan, Karen P., Isabelle Buchstaller, Adam Mearns and Hermann Moisl. (2012). The Diachronic Electronic Corpus of Tyneside English. Newcastle University. https://research.ncl.ac.uk/decte

Kallen, Jeffrey & John Kirk. (2007). ICE-Ireland: Local Variations on Global Standards. In Joan C. Beal, Karen P. Corrigan & Hermann (Eds.), Creating and digitizing language corpora volume 1: Synchronic databases, 121–162. Houndmills, Basingstoke: Palgrave Macmillan.

Szmrecsanyi, Benedikt. (2013). Grammatical variation in British English dialects: A study in corpus-based dialectometry. Cambridge University Press.

Coppélie Cocq, Evelina Liliequist, Lacey Okonski

Protecting the Researcher in Digital Contexts

Keywords *research risks; online abuse; doxing; research ethics*

Contribution short paper

Affiliation Umeå University, Sweden

Abstract In recent years, a growing need for protecting the researchers has unfortunately become necessary as online risks such as death threats and “doxing”, or sharing information publicly for the purposes of harassment and intimidation, have become more frequent risks in relation to an increased digital landscape of anti-gender, far right extremists, and anti-science movements. Researchers now face new risks, and greater levels of risk are posed to researchers. Researchers in the humanities and social sciences are posited to be at a greater risk for these threats (Massanari, 2018, p. 2). Especially targeted are researchers whose work, and/or whose public identity is norm breaking (e.g., ethnicity, minority identity, sexual identity, political

activism, etc.) or challenges white male supremacy, colonialism, heteronormativity and in other ways critically study power structures.

The growing need for developing resources to protect researchers has been emphasized by e.g. the latest ethical guidelines of the Association of Internet Researchers, stating for instance that an "essential measure is that institutions develop policy detailing support procedures for researchers experiencing online threats or harassment related to their work" (AoIR, 2019:11). Here, we are taking a first step toward identifying unsafe research situations, as well as resources and strategies for preventing and protecting humanities researchers with the aim of developing such a policy. By improving safety and support, entities such as universities, departments, and research groups can ensure that researchers do not drop funded lines of research for safety reasons.

While surveillance from the alt-right can intimidate researchers (Massanari, 2018), giving researchers the resources they need to safely conduct research on topics such as gender, race, climate, and politics has far reaching implications for societal and political spheres more generally and for research ethics and methodologies more generally. For example, if researchers do not have proper protections in place and avoid researching topics that put them at risk they may also miss opportunities to conduct work that informs public debate and challenges harmful narratives/paradigms (Yelin & Clancy, 2021). Alternatively, if researchers are enabled to publish risky projects within a research collective and with the institutional support of the university they may be able to spend more time contributing to research and less time mitigating risk factors.

We wish to address these matters and discuss consequences both on an individual level in terms of well-being and academic careers, as well as consequences for the academic society in general when researchers/research gets attacked. We will also suggest strategies researchers can use to protect themselves and each other.

Alexander Conroy

Quantifying conceptual history?

Keywords *Distant reading, literary studies, conceptual history, theoretical reflections*

Contribution short paper

Affiliation University of Copenhagen, Denmark

Abstract Digital Humanities have already adapted theoretical points and analytical perspectives from certain parts of the linguistic turn, namely discourse analysis.

Although not as extensive as discourse studies, conceptual history has received some digital attention recently. This paper examines the theoretical conditions and advantages of combining insights from conceptual history and digital methodology from computational text analysis.

Conceptual history in the Koselleckian tradition operates with analytical tools that seem to fit semantic oriented distant readings easily, such as semantic fields, synchronicity and diachronicity. Further, the empirical ambitions of Geschichtliche Grundbegriffe, the principal work of conceptual history, evidently surpassed the methodological capability as the project progressed. Therefore, future endeavors of conceptual history could benefit from the expanded scope of digital and quantitative methods.

However, the methodological transition from close reading to distant reading is not only a question of changing the methods of research. It is first and foremost a theoretical matter and must be treated as such with the purpose of aligning the methodological steps in computational text analyses with the theoretical framework of conceptual history. The key is to understand how the digital operationalization of existing historical concepts relates to concepts as such, i.e., in a metatheoretical sense, as well as to recognize the epistemological conditions for these operations. Within traditional conceptual history, two theoretical premises are important to acknowledge:
1. Concepts are essentially ambiguous; unlike words, the meaning of concepts cannot be defined, only interpreted.
2. The conceptual historian is (re)constructing past realities via concepts rather than giving an account of the past itself.

The latter entails critical reflections on the historical study as such. It could be argued that the element of construction is more evident in distant readings than close readings and that detailed accounts of how a concept is operationalized computationally hold the potential for strengthening this ontological awareness within conceptual history. The former, on the other hand, raises some theoretical issues. How does the output of for example topic models or word embeddings relate do concepts as such? How do you distinguish words from concepts in distant readings? And how do you determine the boundaries of one concept to another? To support these theoretical reflections, I will provide provisional examples from my research in the concepts of destiny within the Modern Breakthrough of Scandinavia.

Mats Dahllöf

Quotation and Narration in Contemporary Popular Fiction in Swedish – Stylometric Explorations

Keywords *fiction, dialogue, quotation, style*

Contribution short paper

Affiliation Uppsala University, Sweden

Abstract A fundamental feature of many genres of fiction is the alternation between a narrative frame (NF) and quoted inset (QI) dialogue. Both formal and representational features distinguish NF and QI segments. This paper is an explorative study on the stylistic differentiation between frame and inset material in recent commercially successful fiction in Swedish. There are mainly two orthographic options as regards this distinction: Explicitly enclosing inset segments within quotation marks is one. Using an initial dash to indicate utterance display is another, in which case frame and inset material typically alternate in a way not made explicit by the orthography. The corpus behind the present study comprised 450 novels. In order to deal with dash orthography data (135 books), we trained a multilayer perceptron classifier to tell NF and QI material apart. We relied on the fact that native quotation mark text can be converted to annotated dash orthography data, which can then be used for supervised training and validation. A small-scale manual evaluation on the texts we aim to analyze, yielded an accuracy score around 95%. In order to explore the stylometric relations between NF and QI components in the novels, we looked at a selection of basic grammatical features. A characterization of each feature was made by means of recording the fraction of works in which the relative frequency of the feature is higher in QI than in NF. This summarizes how Contribution tend to “use” that feature to create a contrast between NF and QI. Another way to examine how the NF and QI styles are related is to apply a correlation test. We saw, for instance, that QI material in 100% of the books are denser in auxiliary verbs, second person pronouns, and interjections, while NF segments in all or almost all cases are denser in nouns, adjectives, third person pronouns, and prepositions. We could also observe that e.g. noun density in NF and QI correlate in a strong way. The same holds for adverbs and cardinal numerals. This suggests that books and Contribution exhibit stylistic tendencies which affect both narrator and the characters, as far as the kind of fiction we have studied go.

Milda Dailidėnaitė[1,2], Valts Ernštreits[1], Gunta Ķļava[1]

Latvian prefixes in Livonian and digital tools to sort them out

Keywords *Livonian, Latvian, morphology, language contacs, corpus linguistics*

Contribution long paper

Affiliation 1: University of Latvia, Latvia;
2: University of Tartu, Estonia

Abstract

Statistical methods and neural networks can do wonders for languages with huge amounts of available data but can be challenging even in such cases. As we know nowadays most people speak English as a second language and it is possible that a very significant part of the linguistic data in English is produced by non-native speakers. This can be solved because many native English speakers are available. This issue is way more complex when dealing with endangered and critically endangered languages, where most or even all the available speakers use the language at a second capacity. Distinguishing between inherent features, borrowings, mistakes, and cases of code-switching is very challenging and statistical data can be insufficient or even unapplicable. In 2020 we carried out a study on verbal prefixes of Latvian origin in Livonian focusing on their frequency. Verbal prefixes are not an inherent feature of the Finnic languages but are attested in Livonian. The frequency ranges from 0,78% in mid 19th century, up to almost 11% and a pattern emerged: every time the linguistic community was scattered, the frequency of the prefixed verbs grew. Usually, they are considered to be borrowings, even though they are confirmably not obligatory, and lack stability and regularity. We argue, that statistical approach is not adequate in this case and that the need for a different methodological framework when dealing with endangered languages also means that the usual corpus based research is not sufficient. It might be necessary to investigate the backgrounds of the informants in order to understand how past events might have affected their language capacity and use. Research of such phenomena could largely benefit from digital tools that would allow for more sophisticated data processing.

Collecting and processing such data manually would prove to be very challenging. For this reason UL Livonian institute is currently working on a system containing integrated databases where texts can be cross-referenced with the information of informants, location, morphological information, etc. The system will provide the possibility to cross-reference the data from various databases, thus enabling the researcher to filter out the most relevant data on a whole new level, that is one could chose lexical or morphological elements not only based on their stem or morphology, but also based on the location they were used in, the informant they were used by and other criteria.
We will present a case study comparing our previous study of the Latvian prefixes in Livonian, which was mostly based on manually collected data and how such study can be carried out using our new system, thus presenting the possibilities of the system too. We will also discuss how the methodological framework used for some specific study can be applied (or modified and then applied) for a different study.

Jānis Daugavietis

Building Latvian DH tools and resources: rationale and typical practices

Keywords *Latvian DH scene, web survey, DH resource building, DH resource builders*

Contribution short paper

Affiliation Institute of Literature, Folklore and Art, University of Latvia, Latvia

Abstract One of the most important stages in creating a good DH tool / resource is the development process itself. So far, the research has focused on exploring a ready-made tool, probably relying heavily on the 'ideal digital humanities project' model, which consists of four basic components: content, users, management, dissemination (The LAIRAH Digital Humanities Checklist). In practice, this often means researching different user groups, the professional users or the experienced or novice lay users (Walsh et al., 2016). Relatively often humanities professionals have also been studied as users of these tools and resources (see Gibbs & Owens, 2012; Green & Courtney, n.d.; Warwick, 2012 etc.). Although some of them are the builders of such resources themselves, they have been little studied in this role.

We already know quite a lot about different user categories, including such as crowdsourcers and researchers, but for various reasons, tool and resources builders have remained little researched. From rare studies we know, for example, that minority of tool developers actually conduct usability tests (Thoden et al., 2017, p. 1); "only about half of tool developers considered the number of users that adopted a tool as an indicator of success, only about a third ran usability studies, and a disappointing 14% conducted surveys about the tool" (Gibbs & Owens, 2012, p. 1). We know quite a bit about the motivation and typical practices of DH tool / resource builders in developing these tools.

This paper analyzes the data of the web survey and expert interviews of Latvian DH tools builders. Typically, these are the two groups - the representatives of humanities and the programmers, which seem to remain in the shadows (or stands above all else), when researching functioning of DH resources, although they play a key role in the creation and development of these tools. The aim of this paper is to critically evaluate various aspects of their involvement and participation in different stages of DH tools and resources building. Starting with motivation to get involved in resource development, ending with their own assessments of this process and end result.

The survey is based on closed questions analyzed in the spirit of the quantitative approach, but it also includes a number of open-ended questions, which also allow for a more reflexive analysis. We used the following breakdown of the DH resource creation and maintenance process: data use issues; collaboration and communication (incl. interdisciplinarity); planning and

project management; awareness and outreach; resources; technology (Poole, 2017). In addition to the web survey, semi-structured qualitative interviews were used in the research project. The study uses a critical digital humanities approach (Berry & Fagerjord, 2017). In addition to instrumental-technocratic research questions (on best practices and typical difficulties), it also raises ideologically charged questions of values and attitudes.

Katrin Dennerlein, Michael Huber

Reports on modelling dramatic metadata. With examples of the communicative relevance of female playwrights in the second half of the 18th century

Keywords *history of german drama, womens writers, metadata modeling, wikibase, networks*

Contribution long paper

Affiliation JMU Würzburg, Germany

Abstract The purpose of digital literary studies is not only to examine the canon of literary works, about of which we already know quite a lot, but also to consider new works. If we want to learn more about which of the many books that are now digitally accessible were significant in their time and in what way, we need more analyzable data on the handling and evaluation of literature. Such metadata has so far been provided in the form of printed bibliographies, but these cannot easily be made digitally usable. In this paper we present a digitization project on the "Bibliography of theatrical periodicals from 1750-1800" (Bender/Bushuven/Huesman) and some subsequent analyses on the communicative presence of female playwrights from this period. The female Contribution and their works are thereby thematically and structurally located in the discourse on drama and theater for the first time.

We first discuss the data modeling problems that arise from the fact that the bibliography was designed exclusively for analog use to retrieve individual persons, works, theaters, troupes, etc. and requires considerable knowledge of drama and theater history. We then report on the process of digitization with OCR4all which performs with over 99% accuracy after training. Then we discuss the relation to the most advanced data model of this area the Swiss Performing Arts Data Model (SPA, https://github.com/sapa/spa-specifications) and justify our choice of Wikibase as a database framework. While it was obvious that a graph database for later analysis and linking to other projects in the field is the only future-oriented solution, the advantages of Wikibase lie in its linkage to Wikidata and authority records of the national libraries.

An important problem of the project is the considerable effort of Datacleaning which has to be done by researchers with deep knowledge of the research area. We present the impact of to two innovations yet to come: The plans of the German National Library (DNB), to use Wikibase for their authority records "Gemeinsame NormDatei" (GND) will make information exchange between our project and the GND, and therefore normalisation of and Contribution names, easier. The other innovation is the introduction of "Federation" to the wikibase stack, which will help exchange data with Wikidata and related Wikiprojects. All aspects of the project are demonstrated on the example of dramas written by women. We show our solutions for the identification of female Contribution by means of certain markers, then present the astonishing finding that compared to their previous consideration in the history of drama of something 1-3%, a mention rate of female Contribution of 15% can be observed. Then we analyze in which number of different and relevant theatrical periodicals the women Contribution and their dramas are mentioned. Finally, we want to investigate whether there is a connection between the thematization of certain emotions and the gender of the Contribution in the theater periodicals. This is possible due to thematic registers of the digitized bibliography.

Philip Diderichsen[1], Jens Bjerring-Hansen[1], Dorte Haltrup Hansen[1],
Ross Deans Kristensen-Mclachlan[2]

Mending Fractured Texts. A heuristic procedure for correcting OCR data

Keywords *19th century literature, fraktur, OCR correction, Tesseract*

Contribution short paper

Affiliation 1: University of Copenhagen, Denmark;
2: Aarhus University, Denmark

Abstract In this paper we present an evaluation pipeline comparing different methods of Optical Character Recognition (OCR) of 19th century printed fraktur (gothic/blackletter) as well as a correction pipeline, which combines re-OCRing and language technology. The work has been carried out at the University of Copenhagen in relation to a research project involving digital explorations of a corpus of some 900 Danish and Norwegian novels from 1870 to 1899, totalling app. 50 million words. Roughly 25 % of these are printed in the traditional fraktur font, which was almost totally dominating in the beginning of the 19th century. These texts are important culturally, since they represent mostly forgotten, popular novels, however they pose technical and methodological challenges in terms of processing the text from printed page to digital corpus. In order to provide the best possible material for digital literary analysis as well as more linguistic studies, we designed a handcrafted OCR correction pipeline for the fraktur part of the corpus consisting of several different heuristic correction steps, with reference to a gold standard. The first step is a

preprocessing step which takes care of obvious and unambiguous OCR errors. In the second step, we align our primary OCR output candidate (the output from Tesseract using the “Fraktur.traineddata” pretrained OCR model) with several other OCR output candidates and perform context-sensitive correction with reference to these. Especially the Danish “æ” and “ø” characters can be successfully recovered with reference to the Danish, non-fraktur “dan.traineddata” Tesseract model. Finally, in the third step, we employ the SymSpell algorithm (https://github.com/wolfgarbe/SymSpell) to perform spelling correction backed by a word form dictionary hand-crafted from various relevant sources. The pipeline yields an improvement in word error rate from about 11% (89% correctly recognized word forms) to about 3% (97% correctly recognized word forms).

Alina El-Keilany, Thomas Schmidt, Christian Wolff

Distant Viewing of the Harry Potter Movies via Computer Vision

Keywords *Computer Vision, Film Studies, Distant Viewing, Object Detection, Emotion Recognition*

Contribution long paper

Affiliation Media Informatics Group, University of Regensburg, Germany

Abstract Distant viewing, in analogy to distant reading, has been established as term to describe large-scale analysis of visual media like movies with computational methods. The project we present is situated in the same line of research. We present an exploratory case study in film studies for the application of various computer vision (CV) methods for movies of the fantasy genre, namely the movies of the Wizarding World franchise (eight Harry Potter movies and the two first movies of the Fantastic Beasts series). Our goal is to (1) examine the benefits and problems of the methods for quantitative movie analysis in the fantasy genre, (2) investigate if the methods can uncover specific characteristics of certain movies and (3) if we can uncover diachronic developments of metrics across single or multiple movies.

We apply the following image-based CV methods: Color-, contrast- and brightness analysis, object detection, location/scene detection (meaning if the scene of the image is indoor or outdoor), age-, gender- and emotion recognition (the last three are all based on facial information). For each method we use established CV-libraries and state-of-the-art machine learning (ML) models for Python (e.g. Detectron2 for object detection). All methods are used on the image level; thus, we extract one frame per second of each of the

movies and use this corpus of over 77,000 images as representative sample for our analysis.

The analysis of color/brightness distribution shows that the movies become on average darker and less bright which is in line with the content of the movies. Concerning the distribution of objects, the most frequently detected objects are quite homogenous across movies: persons, ties, chairs and books. The location detection shows that the movies mostly play indoor, however with significant distribution differences for certain movies. The average age for all movies is quite similar around 38 with no significant changes across the movies. Gender recognition detects males as significantly more frequent; the distribution does however become more equal with the more recent movies. The most frequent emotions are sadness and anger. While we did identify significant differences considering emotions between movies, we identified the biggest benefits for this method in visualizing emotion progression throughout a movie to interpret the emotional arc of a movie. We intend to discuss the results in more detail for the presentation.

Overall, we were indeed able to find significant differences between movies and interesting progressions of certain metrics across movies. However, while we did not perform an exact evaluation, we identified several problems with the chosen methods and applications since most of the ML-models are trained on material of the social media domain and are not optimized for fantasy movies. For example, the object detection predicting fantasy creatures as dogs or horses or all face-based methods having issues with faces that are not looking directly into the camera. We plan annotation studies for some of the examined metrics to perform domain adaptation and optimize methods to the challenges of the specific use case of fantasy movies.

Antoinette Fage-Butler, Kristian Hvidtfelt Nielsen, Loni Ledderer, Marie Louise Tørring, Kristoffer Laigaard Nielbo

Exploring public trust and mistrust relating to the MMR vaccine in Danish newspapers using computational analysis and framing analysis

Keywords *newspapers, diachronic approach, information dynamics, framing analysis, trust, MMR*

Contribution short paper

Affiliation Aarhus University, Denmark

Abstract The aim of this paper is to investigate how Danish media framed the MMR vaccine debate as a matter of public trust or mistrust. Our results, based on computational analysis of the information dynamics of 231 newspaper articles from 2001 to 2019 and subsequent qualitative framing analysis provides additional information about MMR vaccination coverage in the three major Danish national newspapers, Politiken, Berlingske and Ekstrabladet. We used an LDA model to train article-level dense low-dimensional representations and explored the information dynamics using Nielbo et al.'s [1, 2] approach to change detection in news-based information signals. In addition, we used Entman's [3] approach to identify and analyse frames of trust and mistrust in MMR vaccination. We found that the Danish MMR debate followed patterns of novelty and resonance that typify the expected dynamics of news reporting by legacy new media when news is not catastrophic or shocking [2]. In support of this finding, the framing analysis showed that all three newspapers consistently promoted vaccines as safe and valuable for society throughout the period. Drawing on interdisciplinary perspectives from cultural studies, science studies, public health, computational humanities and media studies, this study presents a methodologically innovative approach to studying historical and near-real time framing of (mis)trust in vaccination in newspaper articles. Recent debates about the safety of Covid-19 vaccines underline the importance of quantifying and qualifying vaccine discourses and paying attention to legacy media's overall agenda-setting role.

Olof Falk

Genre- and subject-based access to literary fiction in digital library catalogues: a brief overview

Keywords *fiction, cataloguing, classification, indexing, metadata*

Contribution short paper

Affiliation University of Borås, Sweden

Abstract In the age of digital libraries and digital library catalogues, the potentials of providing access points to literary fiction based on content aspects - such as genre or subject - seem greater than ever before. However, complexities found in the properties of texts in this category (particularly texts considered high literature) make content-based description signifitancly challenging, often causing cataloguers and cataloguing theorists to opt for other principles of Authors , and consequently resign from developing access points that could arguably benefit many categories of library users. As it seems, these challenges has also caused research and development in this area to stagnate over the last decennia (with some notable exceptions), arguably leaving a challenging, but intriguing, set of loose ends in the theories and practices of digital library cataloguing. Through a literary review, this paper aims to investigate and detail the current state of research and practice concerning digital cataloguing of literary fiction. The paper will provide a brief overview of the classic concepts of subject indexing and classification, and progress into a review of more recent discussions, including discussions of likely reasons why this area is, at present, underresearched. The paper also discusses why collaborative, interdisciplinary efforts may contain keys to unlocking the potentials to solve this problem, rather than the efforts of any solitary discipline. The paper departs from problems relating to classic LIS subdomains, such as classification and subject indexing, as well as more recent subfields such as information retrieval, and ventures further on into subjects of interest to the digital humanities, such as literary studies, literary history, digital archiving, and digital cultural heritage preservation (and access). The paper also discusses how improved practices in digital cataloguing with respect to aspects of literary content may amplify studies of key interest to the digital humanities, for example, by potentially equipping literary corpora with access to authoritative, content-describing metadata to support large-scale literary or bibliographic analyses.

Elena Fernández Fernández

Social Acceleration: An Empirical approach using Computational Text Analysis and Newspapers in Spanish (1988-2018).

Keywords *Social Acceleration, Science and Technology in Society*

Contribution Poster

Affiliation University of Zurich, Switzerland

Abstract The Practice of Conceptual History, by Reinhart Koselleck, explores the idea that there is a direct relationship between technological advancements and an acceleration in the social construction of time. This poster presents the work on this topic by Fernández Fernández et al. in "Measuring the Acceleration of the Social Construction of Time using the BOE (Boletín Oficial del Estado)". The Contribution introduce a quantification of this theory by measuring information density and information variety of narratives in a BOE (Boletín Oficial del Estado) datasate of thirty years (1988-2018). Using Quantitative Narrative Analysis, they define a narrative unit as a triplet of Subject, Verb, Object (SVO), and they define information density (ID) as the ratio of narrative units per words per year. Afterwards, they quantify the different contexts of narratives to measure information variety (IV) by constructing a network of semantic closeness from trained word embeddings. This poster shows an increased IV and ID over the observational time, indicating more and more facts being reported. The results will show evidence of an accleration of the social construction of time.

Rosemarie Fiebranz, Daniel Löwenborg

18th-century village environment 3D visualized – the heritage site Ekeby hamlet

Keywords *3D visualization, 18th century, Sweden, rural history, interdisciplinary*

Contribution long paper

Affiliation Uppsala University, Sweden

Abstract Our paper will present and discuss a unique new, source-based, and complex 3D visualization of a dense village environment in the 1770s. We intend to demonstrate its great pedagogical value for intermediation, interpretation, and use of history. Furthermore, we want to open up for a problematization of difficulties, complications, and possible disadvantages of the method, in its creation, application, further development, and similar complex, source-based visualizations.

The farm buildings from the 19th century are still preserved in the hamlet of Ekeby near Uppsala. Since the 16th century, the farms' returns were allocated to the Archbishop of Uppsala. The farmers remained tenants under similar conditions until the mid 20th century. As a building monument and unique village environment, it is a popular heritage site. However, pedagogically and practically, it is a challenge to convey (to both the general public and researchers) how the hamlet's buildings have been structured and changed over the centuries. The sight of the now preserved, dense, 19h C buildings is fascinating and "obscures the view" for creating concepts of the older, disappeared buildings.

The research project Kring ringgatan focuses on the site's comprehensive history from the oldest period to the 20th century. People's living circumstances in and around the village, the spatial conditions, and their use are studied interdisciplinary with contributions from history, archaeology, and name history. We combine standard historical methodology with map analyses, GIS applications, and 3D reconstructions with tools from modern game design. The result is a detailed 3D model of the settlement with explanatory annotations, which users can explore interactively on the Sketchfab platform. The model can also be embedded on websites with more detailed information about the underlying research and the interpretations and considerations behind the reconstruction.

A detailed house inspection protocol from 1777 formed the basis for an accurate 3D visualization of the hamlet's then about 40 buildings, from the smallest privies to barns and residential buildings. Building techniques and dimensions are well described in the source. The indications of the houses' locations on the village plot are often more challenging to interpret, with approximate latitudes and incomplete information about their mutual location. The close collaboration between historians, building historians, and a graphic/game designer made the visualization possible, in a process where other sources about the village's history also contributed. Contextual knowledge of the local rural and social patterns, together with a good familiarity with the environment's topography, distances, and vegetation, enabled advanced holistic interpretations that we managed to convey visually in an innovative way.

By making the 3D visualization of Ekeby available on the web, we contribute to the use and understanding of the cultural heritage and promote the exploration and communicating of its history. It offers an easily accessible, in-depth study of the site's history for visitors and interested audiences over longer distances. The visualization can facilitate history teaching and learning and be a resource and starting point for further research.

Rikard Friberg von Sydow

“Take these broken links” - Twitter, the Q-drops and the collapse of a digital ecosystem

Keywords *Q, QAnon, content analysis, Twitter, interlinked*

Contribution Poster
Affiliation Södertörn University, Sweden

Abstract The QAnon-movement, a conspiracy theory and political movement who gained fame in the storming of the Capitol in Washington D.C in the beginning of 2021, see the messages from the source “Q” as their most important source of information. “Q” has posted short messages on different message boards (4Chan, 8Chan et cetera) since November 2017. The messages comment on contemporary political events and claim to foresee political development. [1] Who the source, “Q” is, is not known, although speculations exist on the internet. At least we know that the person or persons making the posts seems to want us to believe that they have insight into the backstage of the political scene. [2] In the beginning of 2021 I downloaded the 4953 messages from Q, usually called the Q-drops, as raw text and as pictures. This has given me the possibility to do different types of content analysis of the messages using tools as awk and grep of the Unix and Linux operating systems. [4] Now I want to use this material to analyze the core utility of the internet – its core purpose according to the hypertext pioneer Ted Nelson; “the cells interlinked within cells interlinked” to use Vladimir Nabokov vibrant zukunftvision from the postmodern poetry of Pale fire. [5]

The analysis I want to present in this poster focus on the relation between the Q-drops and Twitter-content. Preserving the Q-drops using the technique I have described above saves all the content created by Q, but not content that has been linked from external platforms. Twitter is one of these major platforms and also a platform that during the US Election 2021, and in the aftermath of the Capitol-riots, purged accounts connected to the followers of the QAnon-movement. [5] How many Twitter-links in the Q-drops have survived and what is the consequences for interpretation of the material without the links that now are dead? Is it possible to see a pattern regarding which type of links, regarding types of accounts etcetera, that still are alive and working? The analysis I am suggesting could help us understand digital ecosystems and our possibilities to understand them when a movement within that ecosystem has been deplatformed and the only witness of their former content are broken links.

Bibliography ”Operation Q” https://operationq.pub/

LaFrance, Adrienne (2020) The Prophecies of Q - American conspiracy theories are entering a dangerous new phase, The Atlantic. [2020-01-01]

https://www.theatlantic.com/magazine/archive/2020/06/qanon-nothing-can-stop-what-is-coming/610567/

Grep Manual Page https://man7.org/linux/man-pages/man1/grep.1.html

Awk Manual Page https://man7.org/linux/man-pages/man1/awk.1p.html

Pam, Andrew (2015) “Intertwingled inspiration” in Dechow, D & Struppa, D Intertwingled, The Work and Influence of Ted Nelson, Zürich: Springer Open.

Nabokov, Vladimir (2011) “Pale Fire”, Berkeley: Ginko Press.

Burns, Katelyn (2021) Twitter purged thousands of QAnon-spreading accounts, including some of the conspiracy’s most prominent backers, Vox, https://www.vox.com/policy-and-politics/2021/1/9/22222074/twitter-purge-qanon-prominent-backers-flynn-powell

Mats Fridlund

Digitizing Humanities Research in the GLAM: Institutionalizing Digital Humanities at the National Library of Sweden’s KBLab, 2017-2021

Keywords *GLAM, digital labs, digital history, text mining, language technology*

Contribution short paper

Affiliation University of Gothenburg, Sweden, Sweden

Abstract This study contributes to the recent history of digital humanities in Europe through a study of the establishment and practices of the digital lab KBLab of the National Library of Sweden (KB). Through the institutionalizing in 2021 of the KBLab as a permanent part of KB it became among the first of such national digital library labs worldwide and thus a pioneer digital institution within the digitization of the GLAM sector and within digital humanities scholarship internationally. The presentation outlines this recent history of the institutionalization of the KBLab 2017-2021.

The origin of the lab was the commissioning in 2017 of a KB-report on similar library labs internationally which after its publication in 2018 led to the KBLab being founded as an internal project at the KB in 2019. During its first two years the lab set up an Authors and an infrastructure in the KB that made possible for researchers to access and utilize the collections of the KB in a digital data format. By 2021 the KBLab were supporting 18 research projects within primarily humanities, social science and language technology. Several of the

digital resources the KBLab was developing for researchers was after its development adapted and implemented in KB's normal library services and thus participating in furthering the digitization of a major national library within the international GLAM sector. In addition, KBLab developed and in 2020 made freely accessible its Swedish language model KB-BERT of written Swedish which was followed in 2021 by its Swedish language audio model KB-wave2vec of spoken Swedish. Both of these models had been made possible by training them on KB's vast digital collections of written and spoken Swedish. KB-BERT was widely adapted for applications by researchers, commercial companies and authorities which was estimated have treduced labor and operations costs of Swedish authorities with tens of millions euros.

In 2021 an external review of KBLab's activities was commissioned that was conducted by the paper's author and which was presented in June 2021. This evaluation was to discuss "the results and effects" of the KBLab as well as provide "recommendations for its future activities". Partly based on its results it was decided to make KBLab into a permanent part of the KB Authors in September 2021. Following that the KBLab was together with eleven other Swedish academic DH centres and units part of the successful HUMINFRA project proposal to become a national research infrastructure funded by the Swedish Research Council, one of the first such Swedish national research infrastructures within the digital humanities.

The paper provide a history of the formative period of the lab and its various activities with a focus and discuss the positive as well as critical views on the lab's activities among its stakeholders. uses primary and secondary sources, including 35 semi-structured interviews with representatives of the KBLab's stake holders within academia, the KB, GLAMs and research councils who included the KBLab's actual and potential users among scholars and GLAM representatives.

Bibliography

Matras, Yaron 2002. Romani: A Linguistic Introduction. Cambridge: Cambridge University Press.

Filip Ginter[1], Harri Kiiskinen[2], Jenna Kanerva[1], Li-Hsin Chang[1], Hannu Salmi[2]

Deep Learning and Film History: Model explanation techniques in the analysis of temporality in Finnish fiction film metadata

Keywords *film history, deep learning, model explanation, text classification, NLP*

Contribution long paper

Affiliation 1: TurkuNLP, Department of Computing, University of Turku, Finland;
2: Department of Cultural History, University of Turku, Finland

Abstract We demonstrate the application of a deep-learning -based regressor, on a case study of predicting movie production year based on its plot summary. We show how the Integrated Gradients (IG) model explanation method can be used to attribute the predictions to individual input features and compare these to human-assigned attributions. Our purpose is to provide an insight into the application of modern NLP methods in the scope of a digital humanities research question, and test the model explanation techniques on a problem that is easy to understand, yet non-trivial for both humans and machine learning algorithms alike.

We find that the model clearly outperforms non-expert human annotators, being able to date the movies well within the correct decade on average. We also demonstrate that the model-assigned attributions agree with those assigned by humans, especially for correct predictions.

Aleksandrs Gorbunovs

Peculiarities of e-learning objects that attract the learner: Towards model development through eye-tracking

Keywords *area of interest, eye-tracking, gaze data, fatigue prevention*

Contribution short paper

Affiliation Riga Technical University, Latvia

Abstract Eye tracking tools are used in many application fields to find out information system users preferences and interests. In educational domain reseachers are eager to implement each new possibility to get know more about students activity and behavioural patterns within learning management system in order to improve learning process efficiency. For teaching staff it would be useful to observe learners activities within digital learning environment, make their gaze data records, analyse them, and come forward with improvements. Eye

tracking technology can help education professionals and instructional designers to recognize weak chains in their digital learning objects and even in a whole Curriculum. Learners' fatigue signs, attention to details on the computer screen and connectivity with peripheral vision issues – that all play an extremely important role in improving the efficiency of the learning process and in the better perception of learning objects.

Covid-19 pandemia brings both to educators and students the new challenges caused by development of digital learning objects, as well as leading of classes in virtual education environments. It might be said that the pandemia highlighted cruicial issues in education and distance learning process which were becoming more and more digital and sophisticated. Moreover, teacher's role becomes more influencial and important in organizing and providing of distance learning process. Lecturer's virtual presence and communication with students may play a cruicial role to motivate, engage and retain the students. This presentation offers an insight into author's recent findings in this field, showing an impact of virtual presenter's appearence features in videolecture on student's engagement. Eye-tracking pilot studies here show the suggested type and time periods of lecturer's apparition and visibility on computer's screen, as well as activities, i.e. gestures, made by the lecturer which attract learner's attention again. Further in this presentation the author proposes also learners' attention attracting and fatigue prevention model. Ethical and privacy aspects while eye-tracking studies will be argued as well, including also the issues to persuade and motivate learners turning on eye-tracking equipment.

Tamás Grósz[1], Noora Kallioniemi[2], Harri Kiiskinen[2], Kimmo Laine[3], Anssi Moisio[1], Tommi Römpötti[3], Anja Virkkunen[1], Hannu Salmi[2], Mikko Kurimo[1], Jorma Laaksonen[4]

Tracing Signs of Urbanity in the Finnish Fiction Film of the 1950s: Towards a Multimodal Analysis of Audiovisual Data

Keywords *multimodal analysis, object detection, automatic speech recognition, film history, media analysis*

Contribution long paper

Affiliation 1: Department of Signal Processing and Acoustics, Aalto University, Finland;
2: Department of Cultural History, University of Turku, Finland;
3: Department of Media Studies, University of Turku, Finland;
4: Department of Computer Science, Aalto University, Finland

Abstract

This paper traces signs of urban culture in the Finnish fiction films of the 1950s, through computational methods, by drawing on multimodal analysis of audiovisual content. It contributes to the development of digital tools for the study of audiovisual heritage. The Finnish National Filmography includes 208 feature films released between 1950–1959. For the study of modernization, the transformation of Finnish society, it is noteworthy that Finland urbanized rather late: prior to the 1960s most of the population still lived in the countryside. The 1950s has however been identified as a decade of profound changes, characterized by increasing urbanization. It was also the most productive decade of the studio era in Finland: cinema addressed and reflected the evolving concerns of the audience.

Our approaches to automatic analysis of media content include aural and visual object recognition and speech recognition. We have concentrated on particular features that epitomized urbanity, including visual objects, such as forms of transportation (cars, horses) and sounds (rural and urban sounds, speech). For the visual part, we have applied visual object detectors, which are able to recognize 80 common everyday objects that typically appear in photos and videos. Based on the scores and frequencies of these recognitions, we observed in the visual contents of the movies quantitative changes that took place in the course of the 1950s.

Audio processing is applied in two ways. First, we processed the movies with neural audio event detection models to find the sounds of relevant objects. Secondly, since only a few movies had subs, we applied automatic speech recognition to generate transcriptions, which are used to analyze the spoken content of the films. Our models reveal an increasing frequency of urban sounds during the inspected decade.

The paper demonstrates that aural and visual object recognition as well as speech recognition can successfully be applied in film historical analysis. Obviously, fiction films were not direct reflections of social life; they were sometimes commentaries of contemporary issues but there were also, for example, historical films that did not portray the period of their making. Still, the overall results support the idea that Finnish filmmakers fuelled the imagination of urban life in the 1950s, paving the way for modern technologies and pushing signs of rural life gradually aside. Methodologically, our experiment can be further enhanced to be more sensitive for historical objects, sounds and speech.

David Håkansson, Carin Östman, Sara Stymne, Johan Svedjedal

How fiction made the Swedish language modern

Keywords *Digital literary studies, literary dialogue, language change, distant reading, literary stylistics*

Contribution Poster

Affiliation Uppsala Universitet, Sweden

Abstract The breakthrough for the novel came during the 1830s in Sweden, and the growing number of novels by Swedish Contribution during the following decades has often been characterized as an important source for the renewal of the Swedish language century (Engdahl, 1962, p. 169). It has previously been established that fiction was important for the renewal of the Swedish language around the turn of the 19th century, but the focus has been on a few famous Contribution and works, such as Röda rummet by August Strindberg, which has been said to mark boundaries between different periods in the history of the Swedish language (Thelander, 1988). The role of fiction in general has however mainly been overlooked and little is known about the language and impact of literature at large, during this period. An important fact to consider is the volume and variety of the literary production, from historical and escapist narratives to stark, realist novels, investigating different social levels of the contemporary society.

We present a cross-disciplinary project, which started in 2020, as a collaboration between Literary Studies, Scandinavian Languages and Computational Linguistics at Uppsala University. The overarching goal of the project is to perform a large-scale investigation of Swedish fictional prose, with the focus on giving a more complete view of the Swedish language and communication history. One sub goal is to investigate the language in literary dialogue, as opposed to literary narrative, and compare the development of new linguistic patterns to those in other types of contemporary texts. The time period in focus is 1830-1930, which is a period of much change in the Swedish written language, largely driven by the goal to move the written language closer to the spoken language.

In a first set of practical experiments, we have focused on automatically separating literary dialogue from narrative, which will enable a large-scale study of thousands of works from Litteraturbanken, a collection of Swedish literary texts. As a first step we have annotated a corpus, SLäNDa, for speech, narrative and other types of elements such as letters and thoughts (Stymne and Östman, 2020). In this material we have investigated the usage of modern versus older forms of a set of function words, like inte versus icke (’not’), and found that they in general occur earlier in dialogue than in the narrative. We

have also started developing classifiers in order to identify speech segments and speech tags automatically with a focus on works where this is marked either by dashes, or not at all, as compared to the simpler case where quotation marks are used. We find that such classification is indeed aided by typographical markers, but that we can considerably improve classifiers for challenging cases by stripping away dialogue markers from the training data. In future work we will further improve our classifiers, and use them to automatically separate dialogue and narrative in order to investigate when different linguistic features occur in each type of material.

Bibliography

Engdahl, S. (1962). Studier i nusvensk sakprosa: några utvecklingslinjer. Uppsala University.

Stymne, S. and Östman, C. SLäNDa: An Annotated Corpus of Narrative and Dialogue in Swedish Literary Fiction. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC'20). Pages 826-834.

Thelander, M. (1988). Nu eller aldrig. Perspektiv på 110 års nusvenska. In Nysvenska studier 67. Pages 5–14.

Mika Hämäläinen[1], Tanja Säily[1], Daniela Landert[2]

Data-Driven Neologism Mining in a TV Corpus

Keywords *neologisms, TV corpus, subs, data mining*

Contribution short paper

Affiliation 1: University of Helsinki, Finland;
2: University of Basel, Switzerland

Abstract New words emerge in language all the time, and sometimes they become a part of the language for a long time, while sometimes they disappear from use as soon as they appeared. Following our previous methods in historical data (Säily et al., 2021), we focus on neologisms in a more contemporary setting. Our aim is to study the emergence and use of neologisms in the TV Corpus, which contains 325 million words of subs (Davies, 2021).

Due to the massive size of the corpus, studying neologism candidates by hand would be a time-consuming task. Therefore, we apply a filtering approach to create an initial list of neologism candidates. First, we extract the publication year of the TV show episode or movie where each lemma in the corpus first appears. This gives us a list of the earliest attestation of every lemma in the corpus. We found that comparing this list to the earliest attestations in the Oxford English Dictionary (OED Online, n.d.), and considering the words that appear in our corpus the same time or before their recorded earliest attestation in the OED potential neologism candidates does not yield enough results, unlike in our previous studies with historical data (Säily et al., 2021). For this reason, we use a large corpus called Corpus of Historical American English (COHA) (Davies, 2012) to do this filtering. We thus compare the earliest occurrences of words in the TV Corpus to the earliest occurrences in COHA producing a list of words that appeared earlier in the TV Corpus than in COHA. This list of candidates will then be gone through manually by carefully studying each occurrence of a potential neologism.

The use of novel vocabulary in television series has been studied by e.g. Bednarek (2018: Chapter 9). We aim to scale up this research by using a significantly larger corpus and automating comparisons with dictionary and corpus data. By comparing the TV Corpus with COHA and the OED and by utilizing the metadata associated with them, we are able to analyse the diachronic development of neologism use in English-language television series as well as register variation in their frequency, types, functions and semantics. As an example, science fiction series often seem to use words related to technological innovations (e.g. biodome in our data), and in some series neology may act as a characterization device (Reichelt, forthcoming).

Bibliography Bednarek, M. (2018). Language and Television Series: A Linguistic Approach to TV Dialogue. Cambridge University Press.

Davies, M. (2021). The TV and Movies corpora: Design, construction, and use. International Journal of Corpus Linguistics.

Davies, M. (2012). The 400 million word corpus of historical American English (1810–2009). In English Historical Linguistics 2010: Selected Papers from the ICEHL 16.

OED Online (n.d.). Oxford University Press. http://www.oed.com/
Reichelt, S. (forthcoming). Innovation on screen. Marked affixation as characterization cue in Buffy the Vampire Slayer.

Säily, T., Mäkelä, E., & Hämäläinen, M. (2021). From plenipotentiary to puddingless: Users and uses of new words in early English letters. In Multilingual Facilitation.

Fredrik Hanell, Pernilla Jonsson Severson

Netnography: two methodological issues and the consequences for teaching and practice

Keywords *netnography, community, consociality, methodology, teaching*

Contribution short paper

Affiliation Linnaeus University, Sweden

Abstract As part of a transnational project focused on creating Open Education Resources (OERs) on selected digital methods and fostering learning experiences by taking data from the past into future stories, the Contribution are currently developing an OER on netnography. Robert Kozinets, who coined the term in the 1990s, recently described netnography as offering a recipe book with clear directions for doing qualitative social media research (2020). Designing this OER, we have identified two pertinent methodological issues of netnography that have been debated during recent years: the need to shift focus from “community” to “consociality” (Perren & Kozinets, 2018) and the issue of active versus passive approaches (Costello, McDermott & Wallace, 2017). Using these two methodological issues as a starting point, this paper outlines our understanding of netnography. It provides examples of consequences for how netnography can be taught and practiced in action. Consociality is more about contextual fellowship (what we share) than the identity boundary (who we are) associated with communities. While this position holds merit, online communities still exist (and warrant consideration), and consequently, we argue for two possible points of departure for

conducting netnographic investigations:
1) community-based netnography, using the notion of community, focused on interactions characterized by (lasting) communal ties and practices;
2) consociality-based netnography, using the notion of consociality, focusing on interactions characterized by (fleeting) connections in contextual fellowships.
These two points of departure frame the nature of the phenomenon of study in slightly different ways, leading us to the debate concerning active and passive approaches in netnographic studies. Costello, McDermott, and Wallace (2017) problematize a certain preference for "observational" or "non-participatory" approaches. Such passive approaches include unobtrusive observations of interactions in a specific social setting. Active approaches include processes to generate elicited material through interactions (such as interviews) between researcher and participants and the writing of field notes.

The critique of passive approaches echoes how a key strength of netnography has historically been described as providing ethnographically thick descriptions of online interactions through the intense and sustained involvement of the researcher in the daily life of the participants (Kozinets, 2010). However, passive approaches are useful to help us navigate vast amounts of digital data and social sites and possibly gain a higher representativity and reduce the risk of bias (Kozinets, 2020). Therefore, we propose that for community-based netnography, it is advisable to engage mainly in active approaches to engage with participants of a community over time. For consociality-based netnography, passive approaches such as selecting and archiving online traces can be enough to conduct a netnographic study. Still, active approaches such as taking field notes should be considered.

Two cases with practical assignments are discussed in relation to these methodological considerations together with insights for teaching and netnographic practice. In the first case, students are invited to investigate a digital community of their own choosing that they know well. The second case introduces students to an accessible online tool suitable for learning about fundamentals of Social Network Analysis (SNA) while studying consociality using data from Twitter.

Trond Haugen

"The Peder Rafn-Project". Reading a 16th & 17th Century Collection of Danish-Norwegian and German Broadsheet ballads in Transkribus.

Keywords *Peder Rafn, Transkribus, Broadside ballads, scholarly editing, 16th & 17th Century*

Contribution short paper

Affiliation National library, Norway, Norway

Abstract In 1936 the University Library in Oslo acquired a 16th & 17th collection of broadsheet ballads (skillingstrykk) end Peder Rafns visebok, after the recipient of the collection. This unique 1030-page long collection of ballads is now accessible in digitized format at the National Library of Norway (https://urn.nb.no/URN:NBN:no-nb_digibok_2008012413001). During the summer of 2021 a team of researchers have joined forces with the DH-lab at the National Library to read, proof-read, and train the artificial OCR-intelligence of the program Transkribus on this printed material. This paper will present the rewarding methodological, and philological outcome of "The Peder Rafn-Project".

The most immediate result of the project will be a proof-read, searchable version of Peder Rafns visebok, which eventually will be connected to the digitized version of the ballads. The language data of this material will be accessible to digital analyses in open-source software for scientific computing; like Jupyter Notebook. It will possibly yield a wide variety of Danish-Norwegian poetry-, ballad-, and reformation language-studies.

The training of the AI-based OCR of Transkribus on Peder Rafns visebok, based on the meticulous proof-reading provided by researchers at the National Library of Norway, will serve as a crucial reference point for further automatized reading of 16th, 17th, and 18th century publications in Denmark and Norway, as well as for further training of Danish-Norwegian text recognition in digitized material.

The speed of the project, thanks to the AI-training of the text recognition, and transcription in Transkribus, has a significant impact on scholarly editing. Establishing the principles of proof-reading for 17th century publications in an early stage of the project, the philologists were forced to face the fundamental challenges of 17th century textual criticism (how do we transcribe ligatures, special characters, virgules, abbreviations, and typographically overrunning lines in these texts?).

Last, but not least, the proof-read transcription of Peder Rafns visebok will serve as the basis for a commented edition of a unique collection of 81 Danish-Norwegian, and 23 German broadsheet prints. Textual commentaries will

broaden our understanding of the history of broadsheet printing in Denmark-Norway. Historical commentaries will survey the status of broadsheet ballads in post-reformation Europe. Studies of printing techniques, and imagery will reveal the transnational nature of this popular cultural material. Musical commentaries will study the earthly, and spiritually edifying ballads' relation to the 16th & 17th century graduals, and book of hymns. Hopefully, it will also display the fascinating common musical repertoire of Northern Europa at the time, connecting the small print houses of Denmark-Norway to German and Dutch printing houses of the 16th & 17th century.

Raphaela Heil[1], Fredrik Wahlberg[2]

Machine learning based image restoration of archival images

Keywords *machine learning, digital manuscript restoration, archive digitization, image degradation*

Contribution long paper

Affiliation 1: Department of Information Technology, Uppsala University, Sweden;
2: Department of Linguistics and Philology, Uppsala University, Sweden

Abstract Substantial parts of the image material of today's digital archives are of low quality, creating problems for automated processing using machine learning. These quality issues can stem from a multitude of reasons, ranging from damaged originals to the reproduction hardware. Modern machine learning has made automatic "restoration" or "colourization" readily available. Curators and scholars might want to "improve" or "restore" the original's quality to create engagement with the artefacts. However, a fundamental problem of the "restoration" process is that information must always be added to the original, creating reproductions with a synthesized extended realism.
In this paper, we will discuss the nature of the "restoration" or "colourization" process in two parts. Firstly, we will focus on how the restoration algorithms work, discussing the nature of digital imagery and some intrinsic properties of "enhancement". Secondly, we propose a system, based on modern machine learning, that can automatically "improve" the quality of digital reproductions of handwritten medieval manuscripts to allow for large scale computerized analysis. Furthermore, we provide code for the proposed system. Lastly, we end the paper by discussing when and if "restoration" can, and should, be used.

Ulrike Henny-Krahmer[1], Robert Hesselbach[2]

Computational stylistic analysis of literary aesthetics in Roberto Bolaño's 2666

Keywords *Roberto Bolaño, 2666, Latin American Literature, Spanish, Stylometry*

Contribution short paper

Affiliation 1: Universität Rostock, Germany;
2: Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

Abstract Roberto Bolaño's novel 2666, published posthumously shortly after the author's death in 2004, is considered one of the most important literary texts of the last twenty years, both for Spanish-language and Latin American literature and for world literature (cf. Herralde 2005; López Merino 2009; Rössner 2007: 513-514; Birns/De Castro 2017). In addition to its complexity of content, the novel stands out for its stylistic variation, which is initially perceived in a subjective way when reading it. This is most evident in the fourth part ("La parte de los crímenes"), when Bolaño describes the murders of women in the town of Santa Teresa with unsparing directness and in a more reportorial style, which can be understood as an allusion to the real events in the Mexican city of Ciudad Juárez (cf. Ferrari 2012: 113-158; Witthaus 2015).

The subject of this contribution is to test these perceived differences by means of a computational stylistic analysis, following the paradigm of "microanalysis of style variation" (Hoover 2017; Krautter 2018). Instead of analyzing a large corpus of texts, the aim is to detect intratextual stylistic signals that are characteristic for the different parts of Bolano's novel, using quantitative methods. The analysis is based on a broad definition of style as any property of text which can be observed by formal features (Herrmann et al. 2015: 44). First, a set of stylistically relevant linguistic and literary features is determined (frequencies of most frequent words (MFW), sentence lengths, vocabulary richness, proportion of direct speech etc.). The five parts of 2666 are divided into smaller segments of equal length and the features are collected for each text segment. It is then checked whether the segments belonging to the same part of the novel are stylistically closer to each other than to segments of other parts. In the case of MFW, the stylistic similarity of the segments is calculated using a Delta measure (Evert et al. 2017) and the segments are clustered hierarchically, using the tool stylo (Eder et al. 2016). For the other types of features, statistical tests are used to decide if differences in the segments belonging to the different parts of the novel are significant or not.

In doing so, the theses are tested (1) whether there is any part of the novel at all that differs from the other parts in quantitative stylistic terms and (2) whether the aesthetics of the fourth part of 2666 can be described by a distinctive set of stylistic traits, differing from the other parts. The paper thus aims to offer a computational stylistic analysis of the work, which in (literary)

scholarly research has so far been examined primarily from a thematic, motivic, and intertextual perspective. The application of computational stylistic methods, which involve linguistic features as well as literary concepts, to Bolaño's novel 2666, fosters the interdisciplinary exchange within the literary and linguistic sub-disciplines of Romance studies. It also adds an interesting case of a microanalytic study of style variation to the Digital Humanities.

Bibliography

Birns, Nicholas and Juan E. De Castro. 2017: "Introduction: Fractured Masterpieces." In: Roberto Bolaño as world literature, ed. by Nicholas Birns und Juan E. De Castro. New York: Bloomsbury Academic: 1-19.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: a package for computational text analysis." R Journal, Vol. 8 (1): 107-21. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

Evert, Stefan, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2017. "Understanding and Explaining Delta Measures for Contributionhip Attribution." Digital Scholarship in the Humanities, Vol. 32 (Issue suppl_2): ii4-ii16. DOI: https://doi.org/10.1093/llc/fqx023.

Ferrari, Sebastian. 2012. Imagining the Inoperative Community: Documentary Aesthetic in Roberto Bolaño and Alfredo Jaar. PhD thesis, University of Michigan.

Herralde, Jorge. 2005. "2666: Datos editoriales." In: Jorge Herralde. Para Roberto Bolaño. Bogotá: Villegas Ed.: 57-59.

Herrmann, Berenike J., Karina van Dalen-Oskam, and Christof Schöch. 2015. "Revisiting Style, a Key Concept in Literary Studies." Journal of Literary Theory, Vol. 9 (No. 1): 25-52. DOI: https://doi.org/10.1515/jlt-2015-0003.

Hoover, David L. 2017. "The microanalysis of style variation." Digital Scholarship in the Humanities, Vol. 32 (Issue suppl_2): ii17-ii30. DOI: https://doi.org/10.1093/llc/fqx022.

Krautter, Benjamin. 2018. "Quantitative microanalysis? Different methods of digital drama analysis in comparison." In: Digital Humanities 2018: Book of Abstracts / Libro de resúmenes, ed. by Jonathan Girón Palau and Isabel Galina Russell. Mexico City: Red de Humanidades Digitales: 225-228.

López Merino, Juan Miguel. 2009. "Bolañismo: 2005-2008." Iberoamericana, Vol. 9 (No. 33): 191-200.

Rössner, Michael. 2007. Lateinamerikanische Literaturgeschichte. Stuttgart/Weimar: J.B. Metzler. 3. Auflage.

Witthaus, Jan-Henrik. 2015: “Biografía negativa en ‘La parte de los crímenes’ de Roberto Bolaño.” In: Roberto Bolaño. Violencia, escritura, vida, ed. by Ursula Hennigfeld. Vervuert, Madrid: 65-81.

Isto Huvila, Olle Sköld, Lisa Börjesson

Documenting and making sense of digital research processes: findings from an international survey of archaeologists

Keywords *paradata, archaeology, data creation, data reuse, surveys*

Contribution long paper

Affiliation Uppsala University, Sweden

Abstract Data-intensive research in digital humanities disciplines requires a thorough understanding of the data used in the research. Earlier surveys of researchers representing several humanities and social science disciplines have repeatedly reported the importance of understanding and conveying the context of research data as a key antecedent of its (re)usability in secondary work. Besides publications and findings, research data is increasingly seen as an outcome that makes a difference, extends the impact of research beyond the first primary findings and observations, and effectively can put research in action long time after it was first conducted. A part of the context a researcher needs to understand that has not been studied to a considerable extent so far, is how the data was created and how it has been manipulated.

The aim of this presentation is to report preliminary findings of an international survey of archaeologists conducted in 2021 on what information archaeologists who are or have been using different types of data need to know about the data to (re)use it effectively, and what archaeologists who have published data for (re)use consider it to be important for others to know about their data. The focus of the presentation is on findings relating to needs regarding knowledge about data creation and processing.

The data was collected using a web survey directed to archaeologists who had been creating and/or using data – understood in a broad sense – in their research. The study is a part of the ERC-funded research project CAPTURE on the documentation of data production and use. The survey was distributed using a wide range of primarily online channels and personal contacts. The analysed data consist of (N=) 90 responses. The statistical analyses reported in the presentation were conducted in R 4.0.3 and using qualitative open-ended coding in NVivo 1.5.

The analysis shows that the respondents considered contextual and processual information—paradata—essential for successful data (re)use. Paradata is used

for understanding multiple aspects of the data but the exact information needed depends on the type of data and how it is used. On the basis of the findings, the presentation discusses feasible strategies to capture and produce relevant paradata for diverse user needs.

I-P

Jonas Ingvarsson[1], Daniel Brodén[1], Lina Samuelsson[2], Victor Wåhlstrand Skärström[1], Niklas Zechner[1]

Between the interpretative and algorithmic: mixed methods and literary criticism

Keywords *text mining; digital epistemology; language technology; discourse analysis; mixed methods*

Contribution short paper

Affiliation 1: University of Gothenburg, Sweden;
2: Mälardalen University

Abstract The New Order of Criticism (2020–2023) is a mixed-methods project, combining interpretative and algorithmic approaches to the study of literary criticism. The project expands on a prior study ('The Order of Criticism', Samuelsson 2013), a sociological and Foucauldian (1971) discourse analysis of book reviews from the years 1906, 1956 and 2006. The current project re-examines the results from the original study through the use of computational tools, language technology and big data. At the same time we explore the compatibility between humanities methodologies and machine-based analysis (Piper 2018; Underwood 2019).

The project primarily actualizes four methodological and technical challenges:
•A re-examination, using digital tools, of the dataset used in the original study, i.e. manual transcripts of literary criticism from a systematic selection of newspapers and journals from the years 1906, 1956 and 2006 (approx 700 texts).
• An exploration of all available literary criticism from the chosen years, with the twofold ambition of: a) comparing the results with the original analysis, and; b) training the machine in actually identifying literary criticism from a large set of digitized texts, drawing on the vast newspaper database of the National Library of Sweden (KB).
• An additional study of literary criticism from 2016, combining traditional and computational methods, in the process studying the possible synergies between the two approaches.
• An epistemological reflection on the differences and similarities between

close and distant reading methods, in the processes of exploring the material, and in the production of knowledge.

The paper presents the these four challenges, discussing the project's theoretical and methodological cornerstones. The research lines up with previous studies stressing the role of digitization within the humanities not only as a matter of tools or gadgets, databases and digital objects, but also as a concern for the organizing, production and dissemination of knowledge (see Liu 2014; Ingvarsson 2021; Berry & Fagerjord 2018; Dobson 2019). Specifically, the aim is to discuss early experiences and results from the interdisciplinary approach utilized by the project, that is a collaborative process where digital and traditional researchers are in a continuing dialogue where ideas, methodologies, and their instantiation in tools, are reciprocally tested and discussed (Rockwell & Sinclair 2016). In essence, we ask: How can traditional approaches inform digital methods? How can insights from working with digital tools inform traditional scholarship?

Gerth Jaanimäe

Challenges of normalizing historical texts written in a morphologically rich language

Keywords *NLP, historical texts, corpus linguistics, text normalization.*

Contribution short paper

Affiliation University of Tartu, Estonia

Abstract Converting the historical texts from old spelling system to contemporary spelling system, also known as normalizing, can be challenging in itself due to the fact that more often than not the rules of spelling were not yet fully established. Another issue lies in the spelling variations often contained in those writings.

Estonian, which belongs into Finno-Ugric language family and on which this research is based on, is a morphologically rich language, meaning that many different word forms can be created. This however poses a new set of challenges in the normalizing process, as training data will inheritably get more varied and thus it would be more problematic to cover sufficient amount of vocabulary within it.

Another issue that can occur is that some of the words normalized can create forms which are homonymous with another word, which may cause falsely recognized lemmas for a given word. Automatic detection of these errors can

be however extremely complicated, mainly because these words are often morphologically correct and the sentences formed by them can also be in accordance with the rules of Estonian syntax.

The dataset that is used in this research consists of parish court records written in the 19th century. These texts were written mostly in Estonian and provide a valuable insight into the way of life, relationships and the language that was used colloquially during this time period. Some of these texts were written in old Estonian writing system, some in modern and a little portion in the so called transitional writing system. Also they contain a sizeable amount of dialectal variation.

These varieties make them extremely interesting from the linguistic point of view, however at the same time make them more difficult to normalize.
In this presentation we will discuss the issues described above and present the initial results of applying the character level statistical machine translation method for normalizing the Estonian texts written in the 19th century.
The experiments are performed by dividing the texts into different training and test sets according to the dialectal variation.

Also we assume that more training data will help us obtain better results. As conversion from contemporary spelling system to the old spelling system is much simpler than the reverse process and can be done with relatively few hand crafted rules, the idea to create the additional artificially created data for the training process will be tested. The old texts that are closest to modern Estonian are converted to the old spelling system and then added into the train set. As the initial training data is scarce we expect some improvement in the results while using this so called additional silver standard.

Anne Järvinen, Eetu Mäkelä

Detecting and analysing news flow dynamics and their changes in 20 years of Finnish news

Keywords *news journalism, named entity recognition, change detection, computational media studies*

Contribution Poster

Affiliation University of Helsinki, Finland

Abstract News media has been vastly influenced by digitalization in the beginning of the 21st century. Digital news is not constrained by the temporal and spatial limitations of traditional print media. The changes in the news media platform arguably allows for more extensive coverage on a specific news event as well as

a broader variety of news. However, at the same time, the resource pool of journalists remains limited, needing to accommodate both the changing content creation schedule, as well as an increasingly hybrid, dynamic and interlinked general media environment.

Thus overall, it is not clear how the dynamics of news work have changed from the mainly print-oriented culture of the early 2000's to the primarily online oriented 2010's. In this study, we use four Finnish legacy media news from 1998 to 2018 to shed light on a particular aspect of these dynamics: we focus on investigating whether there are differences in the way news flows react regarding 'hot topics', breaking news and developing scandals. Our hypothesis is that the increasing interlinkedness of the current online-first and immediate media space is to increase the homogeneity of news coverage, forcing outlets to cover the same central topics and publish the same news repetitively to gain attention, in effect both prolonging their coverage, as well as casting other topics more into the shadows.

To investigate this hypothesis, our approach necessitates the creation of two computational tools or indicators: 1) a detector for hot topics and 2) a means by which to detect the diversity of discussion at a particular point in time. Novelty events in news journalism have a tendency of involving political actors. In addition, extracting named entities (NE) has shown promise in identifying trending news. In this poster as a first approach, we start from identifying a novelty news event from the news flow by extracting named entities. We then compare their prevalence in all news flow to see how different political actors are present in the news flow. Persistent appearance in the news flow would indicate a more general news topic, whereas abrupt peaks of prevalence would suggest a novelty news event. At the same time, as a complementary approach, we aim to utilize distributional methods such as contextual word embeddings, FinBERT and topic features derived from topic models such as LDA, in combination with entropy-based measures of diversity to both identify different qualitative periods in the flow of news, as well as to characterize them for analysis.

Heidi Jauhiainen

Encoding Hieroglyphic Texts

Keywords *Encoding, hieroglyphic texts, tool*

Contribution short paper

Affiliation University of Helsinki, Finland

Abstract With the help of data science, researchers in the humanities can look at large amounts of data at once and find regularities that they might not otherwise see. In order to use digital methods, the texts to be examined must be in machine-readable form but the lack of such text corpora hinders the digital

study of ancient Egyptian texts. OCRing hieroglyphic texts would produce machine-readable corpora, but currently, there is no annotated texts in the same handwriting for training the method. There are also about 1000 hieroglyphic Unicode characters, but they are outside the Basic Multilingual Plane and are not correctly handled by commonly used software applications. In hieroglyphic text, a sign can be next to, above, or over another and two or more signs can be nested. In order to to maintain the information on the signs and their places relative to each other, Egyptologists are using encoding to represent hieroglyphic texts when preparing them for publication in printed form. The encoding uses letter-number combinations from the Gardiner list, a standard reference list for Ancient Egyptian hieroglyphs. To increase the number of machine-readable hieroglyphic texts, I have chosen to use Manuel the Codage encoding, which is the preferred encoding scheme in Egyptology. Ancient Egyptian texts will be encoded in JSesh, an open-source hieroglyphic editor. The aim is to publish annotated texts in structured form and a tool is being build to turn the binary format files produced in JSesh into such machine-readable form. In this paper, Gly2Mdc 1.0, which extracts and cleans the encoding from the binary file, is introduced.

Tommi Jauhiainen, Jussi Piitulainen, Erik Axelson, Krister Lindén

Language Identification as part of Text Corpus Creation Pipeline at the Language Bank of Finland

Keywords *text corpus, language identification, fin-clarin*

Contribution short paper

Affiliation University of Helsinki, Finland

Abstract The Language Bank of Finland hosts text corpora originating from Finland. Two of the most used ones are the Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland (KLK-fi, http://urn.fi/urn:nbn:fi:lb-2016050302) and the Suomi24 Sentences Corpus 2001-2017 (suomi24-2001-2017-korp-v1-1, http://urn.fi/urn:nbn:fi:lb-2020021803). The over 5 billion tokens of the KLK-fi corpus originate from the Finnish magazines and newspapers starting from 1820, which the National Library of Finland has digitized. The Suomi24 Sentences Corpus contains over 4 billion tokens collected from the various discussion forums of the Suomi24 social networking website from 2001 to 2017. The Language Bank of Finland (https://www.kielipankki.fi/language-bank/) has received considerable additions to both corpora. Currently, we are creating new versions of them. As part of this process, we are debuting language identification as part of our corpus creation pipeline. As a language identifier, we are using our recently

published HeLI-OTS software (http://urn.fi/urn:nbn:fi:lb-2022021303). HeLI-OTS is an off-the-shelf language identifier equipped with language models for 200 languages. As a method for language identification, the program uses the HeLI method developed at the University of Helsinki. This method is currently state-of-the-art for language identification between many languages, as evidenced by its first place on the ULI-178 track of the Uralic Language Identification (ULI) 2021 shared task (https://aclanthology.org/2021.vardial-1.1.pdf). We are identifying the language of each sentence, and it will be possible to use the resulting information as part of queries in the Korp user interface (https://korp.csc.fi/korp/). In this paper, we investigate the results and the quality of the language identification process. We were especially interested in seeing how the relatively low OCR quality of the oldest part of the KLK-fi collection will affect language identification when using an off-the-shelf program like HeLI-OTS. The oldest part of the KLK-fi collection and many parts of the Suomi24 corpus contain Finnish written dialectally or otherwise differing from the standard written Finnish. Many of the documents in a collection considered monolingual are, in fact, multilingual or written in a language other than previously thought. We present the preliminary figures and look at some of the exciting cases this process has brought to our attention.

Ellert Þor Jóhannsson, Finnur Ingimundarson

Describing inflectional patterns of nouns in Old Icelandic

Keywords *Morphology, Inflectional database, Old Icelandic*

Contribution short paper

Affiliation The Árni Magnússon Institute for Icelandic Studies, University of Iceland, Iceland

Abstract The Database of Old Icelandic Inflections (DOII) is a project at the Árni Magnússon Institute for Icelandic Studies (SÁM) at the University of Iceland with the goal to describe the inflectional patterns of Old Icelandic. DOII uses the same structure as the Database of Icelandic Morphology (DIM), which is an already developed digital resource (cf. Bjarnadóttir 2012). The linguistic data comes from A Dictionary of Old Norse Prose (ONP), a historical dictionary of the medieval language of Iceland and Norway (cf. Johannsson & Battista 2016). The first phase of the project focuses on simplex nouns. ONP lists 2,800 simplex nouns with more than 10 citations, which give a broad representation of the nouns system and can be processed in the DOII as an initial step to describe the inflectional patterns of Old Icelandic.

In this paper we will present the project. First, we account for the background of the project, how the inflectional system of Old Icelandic or Old Norse has traditionally been described, (e.g. by Iversen 1990), and why such a project is

warranted. We then give an overview of the inflectional system and discuss the two main components of the project, namely the structure of the computer model and data we need to process. We also discuss the benefits of our approach as well as normalization principles. We then describe the long-term vision of the project before we go into more detail of the initial phase and the methods used to input the morphological information about simplex nouns. We briefly mention some of the challenges and issues we have had to consider and how we have chosen to approach them. Finally, there are some conclusions based on the work so far.

Bibliography

Bjarnadóttir, Kristín. 2012. The Database of Modern Icelandic Inflection. LREC 2012 Proceedings: Proceedings of "Language Technology for Normalization of Less-Resourced Languages", SaLTMiL 8 –AfLaT 2012, 13-18.

Iversen, Ragnvald. 1990. Norrøn grammatikk. 7. utg. revidert ved E.F. Halvorsen. Oslo: Tano.

Jóhannsson, Ellert Þór & Simonetta Battista. 2016. “Editing and Presenting Complex Source Material in an Online Dictionary: The Case of ONP” í Tinatin Margalitadze & Georg Meladze. (eds.) Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, 6-10 September 2016, Tbilisi, 117-128.

Marko Jouste[1], Jukka Mettovaara[1], Petter Morottaja[1], Niko Partanen[2]

Archive infrastructure and spoken language corpora for Saami languages in Finland

Keywords *Saami studies, Aanaar Saami, research infrastructure, language technology, corpus linguistics*

Contribution short paper

Affiliation 1: University of Oulu, Finland;
2: University of Helsinki, Finland

Abstract There are three Saami languages spoken in Finland: Aanar Saami, North Saami and Skolt Saami. For these languages, there are archived multimedia materials from over one hundred years. Various linguistic, folkloristic and ethnomusicological materials have been stored in the Saami Culture Archive of University of Oulu. New materials are also actively collected, especially in connection with language revitalization work, and to support language teaching, planning and research (for information on language revitalization work in Saami context, see Olthuis et al. 2013).

The materials archived in the Saami Culture Archive are closely connected to

other work done at the University of Oulu, including indigenous cultural work, teaching and research. As these materials originate from an indigenous Saami culture, there are several specific questions that need to be addressed, especially with regard to access and cultural information in the materials. At the same time, there is value in making the materials as accessible as possible for the communities themselves, ideally online, so that they can be used to their full potential in language teaching, planning and maintenance.
In our solution, the materials are carefully described and analysed by specialists in these languages and cultures, many of whom are also native speakers. We use a manual tagging method where personally identifiable and sensitive information is marked, and can be restricted or removed from later derivations. These versions are made available in the Language Bank of Finland. This way, parts of the archived materials can be used as language learning tools, or as example sentences e.g. for dictionary infrastructure, while also taking into account the cultural integrity and sensitivity of the data. We use national computational infrastructure that is secure and allows for continuous refinement of the materials. What is specific in our approach is the extensive use of modern language technology in data management and validation, combined with a close connection to the needs of language communities.
In our study we evaluate the time spent in different phases of our workflow, providing thereby concrete numeric guidelines for similar future projects. We also analyse and test in detail the sensitive information tagging method which we have designed, with the goal of being able to estimate how much of the tagged content is actually necessary to pseudonymize and anonymize. Similar tagging methods have been used before (Partanen et al. 2020), but concrete estimations of how well they function are extremely important. We also provide accurate statistics about the size of the resulting corpus, and describe it in a manner that will directly benefit the new users, both in digital humanities and other fields of research.

Bibliography

Olthuis, M.-L., Kivelä, S., and Skutnabb-Kangas, T. 2013: Revitalising Indigenous Languages: How to Recreate a Lost Generation. Bristol: Multilingual Matters.

Partanen, N., Blokland, R., & Rießler, M. 2020. A pseudonymization method for language documentation corpora: an experiment with spoken Komi. In Proceedings of the 6th International Workshop on Computational Linguistics of Uralic Languages, January 10-11 2020, Vienna, Austria.

Kati Kallio[1,2], Maciej Janicki[2], Eetu Mäkelä[2], Mari Sarv[3]

Recognizing intertextuality in the digital corpus of Finnic oral poetry

Keywords *Intertextuality, variation, similarity recognition, text reuse, oral poetry*

Contribution short paper

Affiliation 1: Finnish Literature Society, Finland;
2: University of Helsinki, Finland;
3: Estonian Folklore Archives, Estonia

Abstract While the digital corpora enable new perspectives into the variation and continuums of human communication, they often pose problems relating to implicit biases of the data and the limited reach of current methods in recognizing similarity in linguistically complex data, especially in small languages.

The digital corpus of historical Finnic oral poetry in alliterative tetrameter is characterized by significant poetic, linguistic and orthographic variation. At the extreme, a word may be written in hundreds of different ways. The current corpus comprises 189 189 poetic texts in six Finnic languages recorded in 1564–1957 by 5287 recorders. It has a long curation history and significant bias towards some genres, poetic forms, and regions the collectors preferred.
In this poetic tradition, an idea is typically expressed with several parallel, partly alternative poetic lines or motifs, and similar verse types may be used in different contexts. A manual attempt to find all the occurrences of some widely used expression or motif in the corpus is an unreachable task. While the digital tools – starting from simple queries to more advanced methods – make it possible to aim at wider intertextual analyses, some part of relevant material is, typically, not reached. It thus becomes central to estimate the amount and quality of the relevant data that, with different methods, is not recognized.

Here, we discuss two strategies of mapping intertextuality in the corpus: 1) proceeding with text queries and 2) recognizing similar poetic lines computationally, based on string similarity. We compare these approaches with one another, and then proceed to compare the results they yield with existing type index and the results of early 20th century manual research. While the methodological and theoretical foundations of this type of research no longer hold, and while our further interest is in the intertextuality and variation rather than in the problematic concept of poem type as such, parts of earlier analyses may be used in evaluating the performance of digital approaches.

Andra Kalnača, Tatjana Pakalne

Assigning meaning to novel productively formed complex words in actual language use: a case of the Latvian agentive suffix -tāj-

Keywords *novel naming units, productive derivation, disambiguation, context, agentive nominalizations*

Contribution short paper

Affiliation University of Latvia, Latvia

Abstract Any large sample of a reasonably derivationally rich language will have a substantial number of novel productively formed, i.e. non-lexicalized and non-dictionary-listed, complex words. Such novel naming units by themselves, i.e. as context-free pairings of form and meaning (realizations of a word-formation pattern) are often semantically underspecified. This is due to the very nature of productive derivation as a generic means for satisfying specific naming needs. In many instances, a need to name something, for which no established word exists, occurs in a specific context and is, therefore, tied to the specific semantic and syntactic structure of a sentence. The 'general derivational meaning' of a word-formation pattern is in that case specified/ supplemented with additional semes to fit the specific naming requirements. Thus, the exact intended meaning is assigned to a productively formed novel complex word, first, in speech production and, subsequently and independently, in speech perception.

While human speakers usually have no difficulty in interpreting novel productively formed complex words based on context, common sense and general knowledge, in digital language processing a large portion of novel productive derivation would require some sort of disambiguation/meaning specification solution.

Latvian has a rich system of productive word-formation featuring, primarily, suffixation and prefixation, and compounding (Kalnača, Lokmane 2021). The agentive nominalization suffix -tāj- is one of the most productive category-changing non-transpositional derivational suffixes in Latvian. The 'general derivational meaning' of the word-formation pattern V+-tāj-→N is broadly defined as 'the possessor of a property expressed in relation to an action' (Soida 2009, Nau 2013; typologically on action nouns see Štekauer et al. 2012, Roy&Soare 2013, Alexiadou 2017).

In this study, we identify a) the axes along which the 'general derivational meaning' is narrowed down/ specified in actual novel derivatives in specific contexts (lemma frequency 1 to 10 in the LVK2018 corpus), b) the relevant contextual clues, base verb properties, e.g.:
ANIMACY: human, animal, microorganism; group of individuals, Authors ; machine, technology, etc.;

(1) veicinā-tāj-s ‘promoter, facilitator (may refer to any of the above)’
SPECIALIZATION: occupation, specialized tool, machine, technology vs. non-specialized performer of an action, etc.;
(2) uzlabo-tāj-s ‘enhancer, additive, improver with object’, rakstī-tāj-s ‘writer'
EVENT: event-related vs. non-event-related;
(3) turē-tāj-s ‘holder (may be event or non-event related)’
SEMANTIC ROLE: agent, experiencer, etc.;
(4) meklē-tāj-s ‘seeker, searcher’ – slimo-tāj-s ‘smb who is ill’
TYPE OF ACTION: single, specific action vs. repeated, habitual actions, prolonged state/ status based on actions vs. hypothetical action.
(5) staigā-tāj-s ‘walker (may refer to any of the above)’

Bibliography

Alexiadou, A. 2017. Nominal Derivation. The Oxford Handbook of Derivational Morphology. Lieber, R., Štekauer, P. (eds.). Oxford: OUP.

Kalnača, A., Lokmane, I. 2021. Latvian Grammar. Riga: ULP.

Nau, N. 2013. Latvian agent nouns: their meaning, grammar, and use. Baltic Linguistics 4, 79–13.

Soida, E. 2009. Vārddarināšana [Word-Formation]. Riga: ULP.

Štekauer, P., Valera, S., Körtvélyessy, L. 2012. Word-formation in the world’s languages: A typological survey. Cambridge: CUP.

Roy, I., Soare, E. 2013. Event related nominalizations. Categorization and Category Change, Cambridge Scholars Publishing, 123–152.

The Balanced Corpus of Modern Latvian 2018 (LVK2018). http://www.korpuss.lv/id/LVK2018

Almazhan Kapan[2], Suphan Kirmizialtin[1], Rhythm Kukreja[1], David Joseph Wrisley[1]

Fine Tuning NER with spaCy for Transliterated Entities Found in Digital Collections From the Multilingual Arabian/Persian Gulf

Keywords *Named Entity Recognition, Gulf Studies, Colonial Archives, Transliterated Names, custom models*

Contribution short paper

Affiliation 1: NYU Abu Dhabi, United Arab Emirates;
2: NYU Shanghai, China

Abstract

Searchable, transcribed cultural heritage text collections have become an important part of digital GLAM. With the democratization of handwritten text recognition (HTR) platforms, this trend of studying and reusing more texts from archives will no doubt continue. The situation presents an ethical dilemma for computational study, however, since archival materials, particularly those of an intercultural nature, or those written in metropolitan languages about the colonized world are poorly served by text processing methods. Processes like named entity recognition (NER), which allow for the semi-automated annotation of transcribed archives, hold much promise for network and spatial analysis of the sources of the historical humanities, and yet state-of-the-art NER systems (e.g. NLTK) are generally trained on English-language corpora with a metropolitan, media focus and, thus, do not accurately identify non-English entities and labels. They do not offer much in terms of tag customization to move beyond western cultural notions of an entity, and NER is further complicated by inconsistent transliteration practices prevalent in Orientalist scholarship.

Indeed, initiatives exist which begin to tackle the inequities within the field of digital textual analysis. For example, the ongoing #NewNLP workshop (https://newnlp.princeton.edu/about/) is fostering the development of language resources for various world languages. We would like to underscore, however, that the inequities are still present for NLP approaches to English corpora historical in nature and originating from non-metropolitan environments. Our paper reports on collaborative work attempting to close this gap for collections in English, containing transliterated names coming from Arabic-speaking or adjacent Muslim cultures. We present an approach to extracting a variety of named entities (NE) from unstructured historical datasets from open digital collections dealing with a space of the informal British empire—the nineteenth- and early twentieth-century Persian Gulf region. The sources are largely concerned with people, places, tribes and transactions in the region, yet models in state-of-the-art NER systems function with limited sets of tags and they do not capture many of the entities of interest to the historian and do not perform well with entities transliterated from other languages. We build custom spaCy-based NER models trained on domain-specific annotated datasets: the correspondence ledgers of the British Colonial Residency in Bushire and the encyclopedic Gazetteer of the Persian Gulf, Central Arabia and Oman attributed to John Gordon Lorimer. We also extend the set of named entity labels provided by spaCy and focus on detecting entities of non-Western origin, particularly from Arabic and Farsi. We test and compare performance of the blank, pre-trained and merged spaCy-based models and suggest further improvements. Our study makes an intervention into thinking beyond Western cultural notions of the entity in digital historical research.

Heidi Karlsen[1], Lars Johnsen[2]

A Digital Discourse Analysis of the Norwegian National Library's Collections - The Idea of a Feminine Essence in the First Half of the Twentieth Century

Keywords *Literary studies, historical studies, sentiment analysis, topic modeling, discourse analysis*

Contribution long paper

Affiliation 1: BI Norwegian Business School/University of Oslo, Norway;
2: Norwegian National Library

Abstract This paper presents a methodology for identifying discourse. The methodology we propose is threefold: developing search strings, applying the search strings to the target corpus and analyzing the captured material. We have identified discourse on women in the first half of the twentieth century in books, newspapers and journals in the Norwegian National Library's digitized collections.

Our methodological approach consists of topic modeling, situation modeling, sentiment analysis, concordance analyses and collocation analyses. We will show how we have combined these methods in order to prepare the clusters of words that function as our search strings, apply these strings to the target corpus and analyze the captured passages.

The documents based on which we have developed our search strings consist of selected novels and journals from the target period. We have used topic modeling (LDA, clustering) on the non-fictional texts and situation modeling as well on the novels (Piper 2017).

We approach discourse as the identification of statements (Foucault 1969). A discursive statement is an epistemological object. Its function depends on its relation to other statements in an archive. Furthermore, it does not come across through one and only one combination of linguistic signs. This is one reason why it is challenging to operationalize the (foucauldian) discursive statement for digital methodology.

However, sub-corpus topic modeling in combination with a bag-of-word approach and measures of Jaccard similarities can be applied to capture a concentrated material from digitized text collections, suitable for identifying discourse (Karlsen 2020). In this presentation we argue that digital methodology also can be used to analyze the captured material. More specifically, we show that our methodology has enabled us to identify statements on women in the first half of the twentieth century in several tens of thousands passages in the Norwegian National Library's digitized collections.

To identify discourse in the captured material, we have mapped patterns in ways of speaking. Using structured topic modelling, we have trained a topic model in order to study in what contexts one speaks, how these ways of speaking function and what status the subjects who produce discourse on women have — including their societal position and gender. We have added confounding factors to the topic model to analyze the sentiment encoded in passages, distributions of Contribution and variations in topics across the captured passages. Furthermore, insights browsers have allowed us to study the prevalence of topics across the entire corpus.

Bibliography

Foucault, Michel. 2014 [1969]. L'archéologie du savoir. Paris: Gallimard.

Karlsen, Heidi. 2020. “A discourse Analysis of Woman's Place in Society 1830-1880 through Data Mining the Digital Bookshelf.” Dissertation, University of Oslo.

Piper, Andrew et al. 2017. "Studying Literary Characters and Characters Networks (character, social networks, literature, natural language processing, text mining)". https://dh2017.adho.org/abstracts/103/103.pdf

Roberts, Molly, Brandon Stewart, and Dustin Tingley: https://github.com/bstewart/stm

Tangherlini, Timothy and Peter Leonard. 2013. “Trawling in the Sea of the Great Unread: Sub-corpus Topic Modeling and Humanities research". Poetics 41(725-749).

Maria Kazakova

The Proposal of an Algorithm for the Studies in Humanities based on TikTok-material

Keywords *TikTok, ethnocultural stereotypes, web-scraping, comparative analysis, research algorithm*

Contribution short paper

Affiliation HSE University, Russian Federation

Abstract In this proceeding, we will discuss an algorithm for collecting and using TikTok-data in the humanities. We will illustrate it via two studies. The first one is a comparative study of ethnocultural stereotypes; the second one compares the understanding of the term “powerful women” in the English-speaking and the Russian-speaking context. Scientific publications based on the TikTok material are still scarce, and most of them focus on the sociopolitical and cultural impact

and possibilities of TikTok as a platform.

The empirical basis of both studies is a collection of TikTok publications for the years 2020-2021, automatically collected (web-scraped) via our own API in Python. The collection consists of four data types: 1) tags, 2) video captions and text stickers in videos, 3) automatically recognized speech texts, and 4) comments. While we have been writing the code for the scraping, we have also compared different automatic speech recognition tools. To process the empirical basis, we have used quantitative methods of computational linguistics, text content analysis, network analysis, and data visualization techniques. The innovative feature of the empirical basis of the study is that we have included into the analyzed collection the videos created using the #Stitch editing function, while in other studies on the material of TikTok publications the content for analysis was selected only by hashtags, user profiles or randomly.

The first study traces the formation and reinforcement of the ethnocultural stereotypes about the United States and New Eastern countries. The term “New East” refers to a cultural region that includes Eastern Europe, the Balkans, Russia, and Central Asia.

We conclude that on Tiktok, Americans as an ethnic community are rather discussed by other countries of the world, while New Easterners are willing to tell the world about themselves and to be finally heard on a global stage. Thus, there are more heterostereotypes about Americans and autostereotypes about the New Eastern cultures on this platform.

By March 2021, no scholarly articles on the ethnocultural stereotypes on TikTok had been found among scientific publications in Russian or English languages. Our study takes the first step in this direction by analyzing relevant TikTok publications using the theory of ethnocultural stereotypes.
In the second study, we analyze the content of videos with the hashtags #powerfulwomen and #сильныеженщины (translated analog in Russian). We conclude that English-speaking videos with this hashtag often revolve around the characters of movies and TV series, the topics of inner power, minority empowerment, and witchcraft. The Russian-speaking videos also use TV characters as role models, but not only American and British series are popular, but also Turkish series such as “Muhteşem Yüzyıl” have their impact on the population. Also, famous figure skaters are considered to be powerful women. The female politicians, as in the American context, are nowhere to be found in Russian-speaking videos.

Joonas Kesäniemi[1], Mikko Koho[1,2], Eero Hyvönen[1,2], Esko Ikkala[1]

Using Wikibase for Managing Cultural Heritage Linked Open Data Based on CIDOC-CRM

Keywords *Wikibase, CIDOC CRM, Knowledge graph, Cultural heritage data, Linked Data*

Contribution poster

Affiliation 1: Semantic Computing Research Group (SeCo), Aalto University, Finland;
2: HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland

Abstract While thousands of Linked Data datasets of Cultural Heritage (CH) have become openly available in the Linked Open Data (LOD) cloud and elsewhere, an evermore serious challenge is how to manage knowledge graphs when the underlying ontologies and metadata evolve over time (Koho et al., 2018). This paper presents a new practical approach to the problem, based on the collaborative Wikibase platform, the underlying technical foundations of Wikidata. This paper contributes to the state-of-the-art by showing how to utilize the Wikibase model not only for data publishing but for the management of cultural heritage documentation in an event centric way that is compatible with the CIDOC CRM standard and ontology for CH data (Doerr, 2003).

In order to facilitate event centric modelling of CH data with native linked data management tools, our work presents a practical method for organizing content in a local Wikibase instance. Our solution consists of a knowledge base initialization procedure, method for populating Wikibase's data model with CH data, and tools for exposing content as CIDOC CRM based RDF. In addition to the flexible data model, the underlying Mediawiki platform adds support for many important management features such as authentication and change history.

Stemming from its collaborative nature, Wikibase's data model is designed to model statements about items of interest and their references. It also allows for every statement to be qualified with one or more statements, similarly to RDF reification. Duplicate, supporting, and conflicting data are all welcomed if they contain reference to a proper source. This statement level provenance model provides a basis for management of CH datasets. CIDOC CRM is an extensible reference model designed to handle the complexities of describing CH information. However, its event-based approach requires significant conceptual and technological adjustments for people and systems accustomed to object-centered solutions.

The presented solution enables memory institutions and research projects to create and maintain interoperable and reusable CH data without the need to comprehend the intricacies of the CIDOC CRM model. The proposed method

and associated tools are evaluated through a series of usability studies conducted as part of a project with the National Archives of Finland, where the goal is to establish a Wikibase based maintenance platform for a dataset in the military history domain.

Bibliography

Doerr, M. (2003). The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine 24(3), 75–75.

Koho, M., Ikkala, E., Heino, E. and Hyvönen, E. (2018). Maintaining a linked data cloud and data service for Second World War history. Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection, 11196, pp. 138-149. Springer.

Heikki Keskustalo[1], Laura Korkeamäki[1], Kimmo Kettunen[3], Elina Late[2], Sanna Kumpulainen[1]

"What is Missing from our Palette?": Methodological Learning Experiences from a Digital Humanities Research Project

Keywords *historical research, methodology, task-based information interaction, user-oriented research, data-oriented research*

Contribution long paper

Affiliation 1: Tampere University, Finland;
2: Federation of Finnish Learned Societies, Finland;
3: University of Eastern Finland, Finland

Abstract In 1969 Allen Newell defined "method" in his seminal paper in terms of its problem statement, procedure to deliver solutions, and justifications [1]. Digital humanities (DH) uses various born-digital and digitized materials as research data. The use of these materials is examined with various research methods in both data-oriented and user-oriented research settings. Traditional test collection-based information retrieval experiments allow comparing the effectiveness of matching methods based on relevance information known beforehand, without involving any human. Differently, in simulated work tasks human participants are introduced into the setting to perform simulated tasks defined by the researcher. This kind of setting provokes different types of research questions, such as does improved ranking actually increase the utility of a system from the participant's point of view. To answer this, methods and techniques that are different to the previous ones are required - such as observing and interviewing people. Last, consider studying historians performing actual research work in a naturalistic setting. This scenario empowers the researcher to ask even more complex questions, such as what kinds of barriers of interactions are there when the historians cooperate in a multi-disciplinary research group. In this paper, we study and explicate these

issues from the methodological point of view. Our study is based on our shared experiences as a research team in a recent four-year research project in the domain of digital humanities studying researchers' information interactions with historical data in particular. Focusing on six individual studies we present a metasynthesis of the methodological approaches used to answer different types of research questions, which share the common goal of understanding and eventually learning to support the use of DH research data. We utilize the task-based information interaction evaluation model of Järvelin et al. [2] as our framework to structure discussion, as it helps emphasize the key activities of a person interacting with the information access systems, i.e., activities of planning, searching, selecting, and working with information items, and synthetizing and reporting. With the goal of improving our self-understanding, we first systematically tabulate the properties of our methodological selections discussed, in terms of the division of Newell, i.e., domain (and problem statement), procedure, and justifications. We then discuss the necessity of using research methods together in a complementary manner to answer complex research questions. Each method is limited in its own right, which necessitates the applying of methods together to become conscious of the real needs of DH users to better serve them.

Kimmo Kettunen

Geographic Space in Pentti Haanpää's Novel Korpisotaa – where does the War Happen?

Keywords *Pentti Haanpää, semantic tagging, keyness, corpus methods, literary analysis*

Contribution short paper

Affiliation University of Eastern Finland, Finland

Abstract Pentti Haanpää (1905-1955) was one of the most important Finnish Contribution in the first half of the 20th century. His short stories and novels describe many times life in the north-western part of Finnish countryside, but his collected works also include many other themes. Among his works are five books, three novels and two short story collections, that describe either military life or war. His first war novel Korpisotaa describes the Finnish Winter War of 1939–40. Haanpää wrote the novel based loosely on his own war experiences for a competition of a best winter war novel arranged in 1940 by Prentice-Hall together with the Finnish publisher Otava; the novel won third place in the competition. The novel is generally considered as the first realistic war novel published in Finland (Haanpää, 1999; Karonen, 1985, 1999), and its reception was favorable in general (Koivisto, 1998: 272).

In this study we focus on analysis of geographic space in Korpisotaa. We use a digital version of the novel to be able easily search all the relevant space and location words in the novel. The methods we use in the study are familiar from linguistic corpus studies, and they have been used to some extent also in literary studies (e.g., Bondi and Scott, 2010; Fischer-Starcke, 2010). Besides common methods like keyness and frequency counts we can benefit from a lexical semantic tagger of Finnish (Kettunen, 2019; Kettunen and La Mela, 2021). Usage of the tagger systematizes finding of geographic space words in the novel and in the comparison texts and enables us to perform keyness counts for semantic word groups instead of single words.

Our analysis concentrates on three different semantic classes in the semantic schema of our semantic tagger: Z2 (geographical names), M7 (places), and W3 (geographical terms). Our results show that i) exact names of geographical locations (Z2) are mentioned seldom in the novel and they do not name locations of the war ii) words in the two locational classes W3 and M7 in the novel describe either Finnish natural landscape, civilization, or space. Mentions of political space are not very frequent, but they exist (border, homeland/ country). The most frequent two words in the semantic keyness classes W3 and M7 are metsä ('forest') and maa ('earth'/'ground'/'country'/ 'soil'). Forest is one of the main scenes of the war and it is described both as threatening and protective. Many times, forest seems also endless. Maa is a polysemous word with four main meanings. Part of them relate to the country or state, but mainly maa is used in its concrete meanings of ground and soil.
Generally, our work contributes to the usage of corpus methods in literary analysis. Even for a novel length, availability of a digital version of the text helps detailed analysis very much. Usage of a semantic tagger of Finnish brought available a more general level of analysis than plain words. Keyness analysis showed its strength in the analysis of the novel both on the level of words and semantic classes.

Mikko Koho[1,2], Heikki Rantala[1], Eero Hyvönen[1,2]

Digital Humanities and Military History: Analyzing Casualties of the WarSampo Knowledge Graph

Keywords *military history, linked data, digital humanities, data analysis, data visualization*

Contribution short paper

Affiliation 1: Semantic Computing Research Group (SeCo), Aalto University, Finland;
2: HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland

Abstract WarSampo – Finnish Second World War (WW2) on the Semantic Web collects, integrates and harmonizes data about Finland in WW2 and publishes the resulting Knowledge Graph (KG) as Linked Open Data (LOD) [1]. The data is available via an open SPARQL endpoint and a public web portal (https://sotasampo.fi/en) for searching, browsing, and analyzing the data through nine different perspectives. This paper extends our earlier publications on WarSampo by showing how the LOD service and the portal can be used in Digital Humanities data analyses.

This paper shows how various prosopographical phenomena can be highlighted and visualized. For example, we show 1) how the data can be used to visualize geographical variance in the enlisted to officer ratio, 2) how the perished soldiers' places of domicile correlate with their mortality, social class, and place of burial, and 3) how occupation affected the likelihood of surviving from prisoner-of-war camps.

The core dataset of WarSampo is the casualty register of the National Archives of Finland, containing detailed information about all 94,700 people killed in action in Finland during WW2. This data is enriched with interlinked datasets of, e.g., military units, war diaries, wartime photographs, historical places, historical maps, and war-time events. WarSampo also contains a person register of 4200 Finnish prisoners of war and 5600 notable individuals from other data sources (Wikipedia, etc.). An occupation ontology links to the international HISCO classification, providing information about the soldiers' social class.

WarSampo Portal contains tools for simple prosopographical data analysis of the person registers. Accessing the SPARQL endpoint directly enables further analysis and retrieval of data for external tools. Just SPARQL and result set visualizations already enable answering complex questions about history. For example, one can study social stratification by examining the social classes of the perished soldiers and its changes over time during the war, or seeing how age correlates with social class.

With the enhanced possibilities for information retrieval and data grouping attained from the harmonization and reconciliation of metadata values with rich ontologies, the possibilities for answering humanities-driven research questions are greatly increased. Exploiting the new possibilities requires understanding about the data provenance, Semantic Web technologies and computational data analysis, as well as domain knowledge of military historical

research, thus making it an interesting case for interdisciplinary Digital Humanities research.

The WarSampo KG enables getting new insights into WW2, the arguably most devastating catastrophe in human history. The ontology and data infrastructure of WarSampo can be further extended with new data to enable digging even deeper into the societal research questions which interest many military history scholars today [2].

Bibliography [1] Koho, M., Ikkala, E., Leskinen, P., Tamper, M., Tuominen, J. & Hyvönen, E. (2021). WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. Semantic Web, 12(2), pp. 265-278.

[2] Biddle, T. D., & Citino, R. M. (2015). The role of military history in the contemporary academy. Foreign Policy Research Institute Footnotes, 1–6. White paper. https://www.fpri.org/docs/society_for_mil_hist_whit_paper.pdf

Heikki Kokko[1], Tuula Pääkkönen[2]

Translocalis Project: Making of the new digital cultural heritage from the forgotten 19th century historical material

Keywords *cultural heritage, digital humanism, digital history, digital resources, data modeling*

Contribution short paper

Affiliation 1: Tampere University, Academy of Finland Centre of Excellence in the History of Experiences;
2: University of Helsinki, The National Library of Finland

Abstract The phenomenon of letters to newspapers developed into a nationwide and pervasive culture of local letters in the mid-1800s Finnish-language press. A characteristic feature of this culture was that the readers' letters published in the press were written in the names of local communities. Thus, the writer of the letter claimed to represent the entire local community. This interaction between different locations via the press transformed local into societal and societal into local. The culture of local letters had decisive influence on the development of Finnish society, nationalism and civic society in the nineteenth century. However, as a cultural heritage this material had drifted into oblivion. The Translocalis Project that was awarded the major cultural project grant of the Alfred Kordelin Foundation (2021-2023) collects and researches this forgotten cultural heritage of 19th century Finland and creates a new digital cultural heritage from it. The project is implemented in collaboration with the

Academy of Finland Centre of Excellence in the History of Experiences (HEX) and the National Library of Finland. The objectives of the project are:

- Collect all the local letters to the Finnish-language press from the era of 1775-1885
- Publish this database as the part of the digital collections of the National Library of Finland
- Produce the first comprehensive historical research of the phenomenon of local letters.
- Prepare the follow-up project that seeks to collect the letters from the turn-of-the-19th-century press digitally in collaboration with the computational scientists.

Currently, there are over 70,000 local letters that have been manually collected from the digitalized newspaper material by HEX with the tools provided by the National Library of Finland. During the collection work, the letters have been enriched with metadata, such as the place of writing and the named writers of the letters. The database is fully optical OCR-recognized; thus, it enables the methods of digital humanities.

In 2023 the Translocalis database will be open-access published as a part of the digital collections of the National Library of Finland. The database will be published in the presentation system as part of digitized materials and provides search capabilities for the metadata gathered and the OCR-recognized text content of the letters to enrich and structure the clippings' database. The National Library of Finland oversees the technical implementation of the online service. Furthermore, the historical research that is based on the Translocalis database will be published in 2023. Thus, it will work as the "operational manual" that provides the historical context to the digital Translocalis Database.

Our presentation focuses on presenting this project from the different the perspectives of its participants. We seek collaborative partners to our future DH-project in which the letters of the late 1800s and early 1900s are collected automatically by the methods of computational science.

Mogens Kragsig Jensen, Jakob Povl Holck, Evgenios Vlachos

Structuring the Past using Text Mining. Occupational Perspectives on Dansk Folkeblad 1838 and 1840.

Keywords *OCR, text-mining, visualization, 19th century Gothic letters, Danish history*

Contribution poster

Affiliation University Library of Southern Denmark

Abstract Digital Humanities have gained momentum as an academic field. Using modern computational methods and tools on traditional humanities materials in libraries and archives, in combination with software for analysis and graphical display, has enabled exceedingly advanced text mining initiatives.
Through the means of digitization, after applying optical character recognition (OCR) technology to scanned texts, certain types of information that were previously challenging and time-consuming to work with may suddenly be easier studied and disseminated within the Humane Sciences.

This poster presents a pilot-study in which we demonstrate how text analysis and visualization techniques can be utilized to extract, structure and present information on local Danish professions among the subscribers of Dansk Folkeblad (Danish Peoples Magazine, DF). Specifically, we investigate the island of Fyn (Funen) with a focus on the years 1838 (387 subscribers, 425 occupations) and 1840 (366 subscribers, 413 occupations). The lists were published once a year and bound annually. We utilize Transkribus for the OCR part, and the R statistical computing environment for the visualizations. Interesting findings are not limited to the following: main occupations where these of a “Sognepræst” (pastor), “Kjøbmand” (merchant/grocer) and “Proprietær” (gentleman farmer).

DF was a weekly magazine published 1835-1848 by Selskabet for Trykkefrihedens rette Brug (The Freedom of the Press Society). The society was well-known and important in the common debate in the period, and gathered some support among the leading circles, not only in Copenhagen, but in all parts of the country, and thus DF reached out to those interested and active in the public debate. Featured subjects were, besides freedom of the press, current public information and politics. Even though the association was large and important in this the last period of the Danish absolutism, it gradually became more political in its activities, and did, in consequence, lose many members, which led to its dissolution in 1848, the year before the first Danish democratic constitution, which stipulated freedom of expression.
A new focus on the subscribers lists with the use of Digital Humanities techniques can serve as an example of the possibilities in working with the business distribution of the subscribers.

Clelia R. LaMonica

Using online legal databases in English for Specific Purposes

Keywords *English for Specific Purposes, Legal English, Discourse Analysis, Digital datasets*

Contribution short paper

Affiliation Uppsala University, Sweden

Abstract This paper examines the use of digital resources to integrate authentic legal texts and materials into courses in English for Specific Purposes for students of Law and Economics, with an aim to engage Swedish university students with English in practice and better equip them for professional communication. This is especially important given the linguistic situation in Sweden, wherein many international Authors and companies are housed, and English use is widespread among governmental Authors for public communication. Even domestically based Authors are exposed to international regulations and EU law, often discussed using English as a legal lingua franca.

Using digitally available legal archives and databases provides instructors with a wealth of material to illustrate differing discourse structures and language use in context. Corpus analysis can then be further implemented in the classroom to illustrate points of grammar and communication, but additionally gained by the students as a tool they can use to develop their English proficiency outside of the classroom. Here, results of a preliminary case study involving interviews and surveys with Swedish students from a course in English for Specific Purposes for students of Law and Economics are examined, based on their reflections on assignments and course instruction using discourse and corpus analysis. Examples of course exercises and materials using such digital datasets are provided and discussed.

It is common practice for students of Law to be offered one optional course in English for Specific Purposes which covers the fundamentals of English grammar and general structures surrounding written and spoken communication. These often follow a traditional grammar-based curriculum, using a textbook and exercises to teach the rules and conventions, along with some degree of specificity towards a particular discipline, such as law or business. However, legal professionals are then confronted with a variety of text types, styles, and international conventions, in addition to legal jargon specific to the register, which may differ greatly from the English to which have been otherwise exposed. The complex nature of legal discourse and genres necessitates a broad view and varied exposure in instruction (Bhatia, 2017).

The use of authentic documents in Legal English curricula in the United States has also previously been proposed as a means of encouraging close reading of real legal texts; it furthermore facilitates the discussion of texts' linguistic effectiveness and "empowers students to criticize legal texts" while enabling them to "craft language to achieve a desired discourse message" (Hoffman, 2011: 1). It is argued here that this approach would benefit international law students, such as those in Sweden, and it is thus important that effective materials are chosen from digitally available resources and exercises are designed, e.g. from samples of EU and case law, UK parliamentary datasets, and other texts from the legal profession such as official correspondences.

Rafael Leal[1], Heikki Rantala[1], Mikko Koho[1,2], Esko Ikkala[1], Minna Tamper[1],
Markus Merenmies[3], Eero Hyvönen[1,2]

WarMemoirSampo: A Semantic Portal for War Veteran Interview Videos

Keywords *Linked Open Data, Named entity recognition, Named entity linking, Military history, War veterans*

Contribution short paper

Affiliation 1: Semantic Computing Research Group (SeCo), Aalto University, Finland;
2: HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland;
3: The National Archives of Finland

Abstract WarMemoirSampo is a Linked Open Data (LOD) resource of Finnish Second World War (WW2) veteran interview videos, as well as a semantic portal for easy access to them. It hosts a collection of video interviews of Finnish WW2 veterans, mostly reminiscing about their lives during and after wartime. Rough transcriptions of the interviews have been provided, which form the basis of the textual information presented in the portal and the metadata extracted from them. The system is being realized using the Sampo model and by enriching the videos with related information from the WarSampo and Wikidata knowledge graphs. WarMemoirSampo.

This work addresses the key technical challenge of extracting semantic linked data from the textual descriptions of videos: it describes a system in which natural language text is processed in order to produce a harmonized knowledge graph (KG) -- including named entities, keywords and lemmas -- with timestamp information. In order to achieve this, we created an RDF graph featuring data for the interviews as well as the interviewees. The main building blocks of the graph are coarse summary notes written by the interviewers, alongside their corresponding timestamps. However, the timestamps lack precision, since they may be repeated over multiple notes. These circumstances led us to our basic unit of data for the interviews: a group of notes that share the same timestamp, corresponding to a given stretch of the interview. They are of varying length, but typically several minutes long. The textual contents are enriched semantically using NLP techniques and knowledge extraction, resulting in new metadata about mentioned named entities - e.g., people, places, and events - and keywords generated via a pre-trained subject indexing tool.

The WarMemoirSampo portal is implemented using the Sampo-UI framework, which enables faceted search, exploration and analysis of the interviews. It is possible to identify interviews from specific interviewees and/or based on mentions of places, persons or subject matters (keywords) of interest. The results of the search take the user to the relevant parts of the video interviews, so that the veterans can be heard in their own voices. A semantic recommender system provides the user with links to related interview snippets present in the database, as well as additional information in WarSampo. More features are planned for the future: the links to the entities found in the texts will be integrated into the text itself instead of being listed separately. An event detection tool is to be developed which extracts event information using times and places mentioned in the interviews. Moreover, when Finnish speech-to-text technology advances to the point that everyday dialectal speech can be automatically and reliably transcribed, the same tools could be used on the transcriptions, resulting in richer and more accurate metadata.

Peter Shaff Leonard

Rummet som icke är: Generative models and latent space in visual collections.

Keywords *Machine learning, generative adversarial networks, art history, visual collections*

Contribution long paper

Affiliation Yale University, United States of America

Abstract Advances in convolutional neural networks have given scholars new methods of analyzing large-scale collections of images, such as paintings or photography. Computer vision methods based on deep learning have enabled scholars to find macro patterns of visual similarity, in a form of “distant looking” that mirrors text mining approaches to literary studies.

Emerging alongside these analytic capabilities are generative possibilities: the use of cultural heritage collections as the raw materials for digital hallucination, or algorithmic dreaming. Methods such as Generative Adversarial Networks pit two algorithms against each other, in a high-speed digital contest over what “real” paintings or photographs are. When properly trained, and with enough ground truth for evidence, one result of these network pairs is a generator that can produce entirely novel images, which pass superficial inspection by humans.

Often referred to as “deep fakes” in popular journalism, images produced from

these machine learning models are steadily advancing in verisimilitude and sophistication. This paper considers generative models trained on several large Scandinavian and Northern European visual collections, including 18th Century satirical prints from the British Library, the Flora danica copper prints of the 1760s, and negatives from the studio of pioneering female photographer Lina Jonn in turn-of-the-century Lund. Are the artificial artifacts produced from these trained models the “criminal sibling of criticism”, or can they be put to use in considering the boundaries and limits of representation of Nordic visual culture? What are the ethical questions surrounding visual collections as "ground truth" for the production of novel artifacts? And what active roles can humanists play in guiding the development of these techniques?

Bibliography

Grafton, Anthony, and Ann Blair. Forgers and Critics: Creativity and Duplicity in Western Scholarship. 2019.

Karras, Tero, et al. “Training Generative Adversarial Networks with Limited Data.” ArXiv:2006.06676 [Cs, Stat], Oct. 2020. arXiv.org, http://arxiv.org/abs/2006.06676.

Petri Leskinen[1,2], Heikki Rantala[1], Eero Hyvönen[1,2]

Analyzing the Lives of Finnish Academic People 1640–1899 in Nordic and Baltic Countries: AcademySampo Data Service and Portal

Keywords *Linked Data, Data Analysis, Network Analysis, Cultural Heritage, Digital Humanities*

Contribution long paper

Affiliation 1: Semantic Computing Research Group (SeCo), Aalto University, Finland;
2: HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland

Abstract This paper shows how the newly published Linked Open Data (LOD) service and semantic portal AcademySampo can be used for Digital Humanities (DH) research. The original primary data, based on ten man years of digitization work, covers a significant part of the Finnish university history based on the student registries in 1640–1852 and 1853–1899. They contain biographical descriptions of 28000 students of the University of Helsinki, originally the Royal Academy of Turku. AcademySampo also sheds light to the academic history of Sweden and Baltic countries through their shared history with Finland in the larger Swedish Empire. The Finnish student registries have been widely used by genealogists and historians by close reading. We argue that by using AcademySampo unprecedented new possibilities for DH research are now enabled: the underlying knowledge graph can be accessed and analyzed using Semantic Web technologies and tools and with the ready-to-use data-analytic

tools of the portal. Examples of data-analysis are presented by using the AcademySampo system for studying migrations of students in Finland, Sweden, Russia, and Estonia, history of student nations, inheritance of occupations and social classes, lengths of family lines of students, and network analyses of students. Related analyses have been made before using biographical dictionaries but not for academic history and student registries.

Thea Lindquist, Erik Radio

Metadata Tools for Bibliographic Data Science

Keywords *Metadata, bibliographic data science, documentation, digital humanities, European history, academies*

Contribution short paper

Affiliation University of Colorado Boulder, United States of America

Abstract Bibliographic data science provides new approaches to understanding historical contexts and corpora through the use of metadata as source material. Researchers at the University of Colorado Boulder have initiated a project drawing on this approach to investigate the impact of the Fruitbearing Society, the first and largest academy of arts and letters in 17th-century central Europe, through the large-scale analysis of historical bibliographic data. Undertaking this work requires a variety of programming tools and methods to wrangle and process this data. Further, identifying various data sources requires careful consideration of the potential rights issues that may accompany the aggregation of disparate data sources. Through this work we aim to inform our understanding of the Society and the 17th-century central European publishing landscape, as well as provide openly available research outputs in the form of data, code, and documentation. This article provides a brief overview of the project's background and aims and then examines the processes and tools used to scrape, script, and store the metadata in ways that facilitate bibliographic data science.

Eetu Mäkelä, Pihla Toivanen

Analyzing the representation of politicians in the media – results and methodological concerns

Keywords *political journalism, representation, media logic, data complexity*

Contribution long paper

Affiliation University of Helsinki, Finland

Abstract In early 2021, our research group collaborated with the Finnish journal Suomen Kuvalehti on an article examining how Finnish members of parliament were represented in four major news media in the year 2020 (Suomen Kuvalehti 2021). Based on our initial analyses, it looked like we were on to a major scoop. The right-wing populist Finns party seemed less well represented in commercial media, while being artificially propped up by the Finnish public broadcaster YLE. Female representatives also seemed to be left much more to the sidelines. However, the more we dug into our data, the more complex the picture became. After multiple rounds of methodological evaluation, improvement and reflection, both of these findings ended up melting away. What we found instead was that the publicity of members of parliament was highly predicted by the office they were holding. Only by controlling for this variable could parties successfully be compared. Here, we did find some differences between e.g. how comparable representatives from the Finns party and the centre-right National Coalition Party figured in the media.

Further, an interesting overall dynamic appeared: when looking at individual representatives, their portion of news coverage varied widely and featured distinct patterns: some were present relatively constantly due to their political position, while others featured only at particular times, such as when flagged in a scandal. Looked at on a daily level, reporting volume thus looked extremely event-driven, as seems natural. When looked at as a whole, however, the number of news stories featuring members of parliament stayed remarkably stable through the year. We hypothesize this to be determined basically by there being a fixed number of reporters covering the political beat. Thus, the dynamic is such that while the overall output stays relatively the same on a weekly basis, its content changes widely in response to the events of the day. Beyond these and other findings, our paper will also outline the process by which we were able to obtain them, starting from crawling the relevant news from the websites of the media outlets (Toivanen & Mäkelä 2021) to multiple rounds of increasingly elaborate statistical analyses. Particularly, we will focus on the many checks and balances we had to employ in this process to ensure that we were measuring what we wanted to measure, and were not getting bit by either faulty data or other interacting phenomena apart from the ones we wanted to measure.

Bibliography Suomen Kuvalehti (2021, May 21). Vuosi valokeilassa: Kuka sai medialta huomiota? Kuka jäi varjoon? Suomen Kuvalehti selvitti tutkijoiden kanssa, miten kansanedustajat näkyivät neljässä suuressa uutismediassa vuonna 2020. [A year in the spotlight: Who got the attention of the media? Who was left in the shadow? In co-operation with researchers, Suomen Kuvalehti analyzed how Finnish members of parliament were represented in four major news media in the year 2020] Otavamedia. https://suomenkuvalehti.fi/jutut/kotimaa/politiikka/nain-isot-mediat-antoivat-

nakyvyytta-kansanedustajille-sk-selvitti-tutkijoiden-kanssa-kuka-kerasi-huomion-ja-kuka-jai-varjoon-vuonna-2020/?shared=1183317-08c94dca-4

Pihla Toivanen, & Eetu Mäkelä. (2021). hsci-r/finnish-media-scrapers: Scrapers for extracting articles from Finnish journalistic media websites. Zenodo. https://doi.org/10.5281/zenodo.4604909

Johan Malmstedt

Collecting Silences: A comparative analysis of silence in Swedish radio from P1 and P3, 1980 - 1989

Keywords *Swedish Radio, Media History, Audio analysis, Silence, Public Service Broadcasting*

Contribution short paper

Affiliation Umeå University, Sweden

Abstract This article proposes a methodological approach to the analysis of large-scale radio archives. By collecting and measuring silences in broadcasting, the analysis explores stylistic differences within the Swedish media monopoly in the 1980s. This period marks the last decade of the Swedish public service monopoly and offers rigorous data on the development of mass media without deregulated economic competition. In 1993 Sweden became one of the last European countries to allow commercial radio, nevertheless, during the decade prior, competition was introduced from within the monopoly. At the time, Swedish radio featured several separate channels. This analysis focuses on the longstanding flagship channel P1, and its main competition, P3, which was considered a more youth-oriented alternative. Prior research has stressed the independent and different character of these two channels, subscribing to a depiction of a harmonically diversified media system. By analyzing the very audio data preserved from the broadcasting, it is possible to explore such stylistic divergences, as well as similarities, on a quantitative level. What unique features can be detected by studying the very signal, and are there homogenous tendencies beyond these differences?

In order to the approach this question, I propose a simple method: collecting silences. By means of basic signal processing, the amount of silence in broadcasting can be extracted and comparatively studied. Once data on the pauses and gaps in the broadcasting has been extracted, both the total amount and the number of occurrences can be analyzed. Moreover, the method can be applied both to the overall structure and on separate sections within the broadcasting, to achieve a more granular understanding. In doing this, it is possible to achieve new knowledge on the style and pace of the broadcasting,

and how these matters changed throughout the decade. The aim is both to render new insights into Swedish media history and to suggest new ways for digital humanities to integrate the vast potential of audio analysis.

Jani Marjanen, Antti Kanner, Eetu Mäkelä

Using a bigram model for semantic change to study Finnishness in early newspapers

Keywords *ethnonyms, conceptual change, historical newspapers, bigrams*

Contribution long paper

Affiliation University of Helsinki, Finland

Abstract Our paper analyzes the historical understanding of Finland and Finnishness as it was expressed in newspapers published in the late eighteenth century and the early nineteenth century. As the period saw the decimation of the Swedish Kingdom and establishment of the Grand Duchy of Finland within the Russian Empire, a change in language use can be expected, but the changes occurring are rather fine-grained and difficult to detect without a systematic and transparent charting of the data. This paper suggests a method based on the analysis of bigrams to study this type of semantic change. Many existing methods are designed to navigate massive amounts of linguistic data and do well in solving computational tasks, but are not always a good match for the kind of historiographical questions such as ours. More broadly, we establish that the application of Finnish can in principle refer to the categories of language, geography, nationhood, statehood, and territoriality. Our analysis shows that especially the categories of language and state underwent a gradual shift in the period from the eighteenth to the nineteenth century.

Benjamin George Martin

Digital Analysis of Global Debates: Text Mining the UNESCO Courier, 1948-2011

Keywords *digital text analysis, conceptual history, digital history, international relations, public humanities*

Contribution short paper

Affiliation Uppsala University, Sweden

Abstract This paper will present my ongoing effort to advance our knowledge of global conceptual history by applying methods of digital text analysis to an international text corpus: UNESCO's official journal Courier. Founded in 1948, Courier's self-described mission was to "promote UNESCO's ideals, maintain a platform for the dialogue between cultures and provide a forum for international debate." More than almost any other twentieth-century publication, Courier had global aspirations and global reach. At its high point in the 1970s-80s it featured articles from prominent intellectuals across the globe published in 35 languages with an overall distribution of over 1.5 million copies. It was, moreover, one of very few publications available on both sides of the Iron Curtain.

Working with developers and researchers at Humlab (Umeå University), I build on insights from the digital analysis of newspapers and other print media to develop a suite of methods for this corpus, including (for now) quantitative analysis of word frequency and word co-occurrence trends and topic modeling. In this paper, I will present our approach to Courier and outline preliminary results, focusing on our use of topic modeling. We use this method in particular to explore the impact of decolonization, asking how Courier's content changed with the arrival of the "third world" in the 1960s, and to what degree the journal's topical foci reflected broader political and ideological trends.

A goal of the larger project of which this study is a part ("International Ideas at UNESCO", inidun.github.io) is to put digital humanities into action by working with UNESCO to make our curated corpus and some of our analytical tools accessible to the public through an interactive platform. The paper's final section will present the status of this effort and discuss some of the challenges involved in bringing it to fruition.

Antonina Martynenko

Reading the unreadable: towards formal distinctions between 19th-century Russian women's poetry and one failed hoax

Keywords *stylometry, poetry, gender*

Contribution short paper

Affiliation University of Tartu

Abstract While stylometry is commonly used for Contributionhip attribution, this study aims to apply its techniques to identify stylistic sources of a poetic hoax "The Works by Anna Smirnova" (1837, in Russian, unknown author(s)). The book constitute an example of a mismatch between the imaginary 1830s poetess' portrait and the language means used to imitate her writings, in particular, "men's" sentimental poetic language of the 1790s driven to the point of absurdity.

55 poems written in Smirnova's name are hardly readable. Being very long, they are deliberately made incoherent and meaningless to a reader. The poems can be described as randomly mixed poetic clichés that makes the hoax similar to the bouts-rimés game, where a poem is written according to a list of pre-selected perplexing rhymes that often results in the text's absurdity.
Although usually stylometric frequency-based approaches assumed to lose some of the information kept in literary texts, these techniques might have an advantage over human readers in this case, where the text was designed to be meaningless.

The similarity between "Smirnova" and other Contribution was studied though clustering experiments examining probable Contributionhip signal and thematic similarity to 31 poets and 6 poetess active between 1790 and 1840. For more precision each author's corpus was divided into decades; most of the women's poems were digitized for the first time for this study.
The hypothesis is that the hoax's style is significantly different from any women's poetry of the 1830s, regardless the usage of a young poetess' image. On the formal level, its archaic metrical form (iambic hexameter) may already point to the hoax's sources in the sentimental poetry of the end of the 18th century. For the main experiment I calculated Burrow's delta between 250 most frequent words and built hierarchical clusters according to Ward's method; built on 100 dendrograms consensus tree shows that the hoax's author is not close to any poets of the 1830s, but creates a stable cluster with the most-known male poets of the 1790s and 1800s. The association of the hoax with the sentimental poetry remains stable with different number of MFW. During the presentation this outcome will be discussed in more details and compared with classification experiment (SVM). The resulting analysis will demonstrate that the hoax was a failed attempt to imitate women's poetry while taking the "material" from outdated "men's" sentimental style.

In a broader context, the study aims to show how computational methods can help to study associations between gender and literary style, particularly, in cases of parodic style imitations. At the same time, this historical case allows to examine to which extent lexical features are useful while working with deliberately distorted texts which author remains unknown.
(For preliminary results, see: https://github.com/tonyamart/smirnova_hoax)

Bibliography Mānušs, Leksa, Neilands, Jānis & Rudevičs, Kārlis. 1997. Čigānu-Latviešu-Angļu un Latviešu-Čigānu Vārdnica. Riga: Zvaigzne ABC.

Inés Matres

Practices of looking from the photo archive for a postphotographic age

Keywords *photographic heritage, collection digitization, postphotographic condition, digital literacies*

Contribution short paper

Affiliation University of Helsinki, Finland

Abstract In this paper I take up a challenge addressed by photo archivist Joan Schwartz to "consider how collection digitization affects visual meaning making in the digital-born world" [1]. I approach this complex question by differentiating two factors that conflate in this. On the one hand, the postphotographic condition in which our use and consumption of images today is placed: a global context of visual abundance where manipulation and appropriation are reinforced and where the spectator is often ignorant of cultural references [2,3]. All this require wary practices of looking, inquisitive of production and circulation contexts [4]. On the other hand, we need to consider the way photographic heritage is presented to us in the digital-born world, accessible through digital libraries and online galleries, where its archival and database logic [5] add layers of meaning to the photographic image.

Drawing from interviews with photographic curators involved in digital projects and photo elicitation in photo archives in Finland, I inquire about practices of looking [6] that derive from tasks of collecting, sampling, digitizing, cataloguing and interpreting photographs. Analysing this chain of practices, a continuum of mediation emerges in which photographs are "put to work" that is not just determined by the digitisation process. This chain include decisions made at the time of collecting, hierarchies of access created when selecting takes for digitization, or during the cataloguing process that translates photographs into textual networked content (thus becoming searchable and visible). As a result, digital photographic heritage does not necessarily present authorial intention or lay out the network of relationships concurring it the formation of photographs as much as we would think. These practices can blur the bond

that explain why photographs were made, collected or have reached us, which contributes to think of photographic heritage in similar terms as the postphotographic.

In her same address, Schwartz observes that meaning is greatly determined not by obvious attributes captured in the photographic image (and often in its metadata) but by the questions they inspire in the beholder. Considering this and returning to the initial question, when photographs are digitised and transformed into searchable objects, they acquire a form of presentation (materiality) more adequate to inquiry, and allow to find answers to our questions. However, they also requires the spectator to engage with images beyond their content onto other documents, biographies, and publications, similarly as an archivist would. Thus, practices of looking from the archive (we could say literacies) are much needed in a postphotographic time.

Bibliography

[1] Schwartz, J. (2019) "'In the Archives, a Thousand Photos That Detail Our Questions': Final Reflections on Photographs and Archives." In: Bärnighausen, J. et.al.: Photo-Objects on the materiality of photographs and photo archives in the humanities and sciences

[2] Fontcuberta, J. (2015) The Post-Photographic Condition.

[3] Ritchin, F. (2009) After Photography

[4] Mitchell, W. J. (1994) The Reconfigured Eye : Visual Truth in the Post-Photographic Era.

[5] Manovich, L. (2001) Database as symbolic form

[6] Sturken, M. and Cartwright, L. (2009) Practices of looking: An introduction to visual culture

Haralds Matulis, Sanita Reinsone, Ilze Ļaksa-Timinska

Automatic Detection of Dates in the Corpus of Diaries

Keywords *date detection, corpus analysis, crowdsourcing, digitization, hand-written texts*

Contribution short paper

Affiliation Institute of Literature, Folklore and Art of the University of Latvia, Latvia

Abstract Diary writing tradition is a complex phenomenon. Forms and styles of how personal diaries are written can differ even within one notebook of a single author. However, it can be assumed that the date at the beginning of a daily record is a formal element that distinguishes diaries from other forms of personal autobiographical reflections in written form. Dates are important formal elements that keep diaries structured and aligned with the narrated time.

This paper deals with the automatic detection of dates in a corpus of digitized hand-written diaries in Latvian and methodological challenges seeking a solution on how to automatically distinguish dates denoting specific date entries from other mentions of dates appearing in the text. The pilot corpus of diaries was built by the Institute of Literature, Folklore and Art, University of Latvia.

Exploration of how consistently dates are designed in 40 personal diaries of different lengths (consisting of a few records up to diaries written for 55 years) that are included in the pilot corpus, written by Contribution of different age, educational and social backgrounds in the Latvian language from 1917 to 2021, leads to a conclusion that date notation is a creative practice and depends on an author's taste, habits and probably also emotional mood. Within one diary, there can be up to 16 different types of how dates are fixed. There are also no regulations on where a date should be located. Most often, it can be found at the beginning of a record, while sometimes it can be placed at the end or even integrated into the record. Sometimes, a date of a record can be deduced in the context of other records only.

Dates are important to carry out a diachronic analysis of diaries and compare metrics across different Contribution. Slicing every diary into entries for single days could give a useful research perspective. One day is a semantically meaningful time unit for analysis, and it also coincides with the text splitting system used by the majority of diaries Contribution.

The challenge of finding dates in the diary corpus consisted of two parts: (1) to find all dates occurring in the corpus. The style of dates notation in diaries is oftentimes elliptic, sometimes obscure and varies widely even within one author. (2) to find all metadates – i.e., dates depicting start for the diary entry

of a particular day – and to effectively distinguish metadates from other dates and numerical data in the corpus.

The most frequent placement of the metadate followed the following convention: empty line before an entry of a new day, metadate on a new line, diary entry for that day on a new line. Although, there were numerous irregularities to this schema of metadate placements which were crucial to be taken into account to correctly pre-process corpus that will be discussed in more detail in the full paper.

Florian Meier

Towards Contributionhip Attribution in the Trykkefrihedsskrifter: A Stylistic Analysis of the Danish Freedom of the Press Writings' Main Writers

Keywords *Contributionhip attribution, trykkefrihedsskrifter, stylistic analysis, text mining*

Contribution poster

Affiliation Aalborg University Copenhagen, Denmark

Abstract Contributionhip attribution, i.e. finding the true author of a text for which Contributionhip is disputed or unknown, is usually performed on a comparative basis. This means that an Contributionhip model compares a text's style representation with the style of possible author candidates. Finding possible candidates can be done in various ways. One option is that domain experts with knowledge about a collection are able to select possible candidates based on experience from close-reading. However, if a collection is of considerable size, close-reading all texts is extremely time-consuming. Moreover, if the number of texts of unknown Contributionhip is vast a pre-selection of texts with similar style is needed. In both situations, computational approaches for creating stylistic profiles of Contribution and texts can be helpful to narrow down the number of potential author candidates and texts to be considered in machine learning-based Contributionhip attribution experiments.

The Trykkefrihedsskrifter, the danish press freedom writings, is a collection of pamphlets published and collected during the 1770s in the kingdom of Denmark-Norway. The publication of these short texts was made possible through the abolition of censorship by Johan Friedrich Struensee, the de-facto regent of Denmark and Norway at that time. In these pamphlets, Contribution could for the first time freely and without restrictions discuss recent events and

topics like religion and the church, economy and trade, societal conditions or patriotism. To better understand the collection and study idea generation and spread at that time, knowing who wrote the texts is of high importance. However, about half of the collection is of unknown Contributionhip and the number of Contribution shows a very long-tailed distribution with more than 100 Contribution having written only a single text.

In this paper, we present a stylistic analysis of the top-10 Contribution of the danish freedom of the press writings with the aim of supporting researchers in the selection of author candidates and texts of unknown Contributionhip for Contributionhip attribution experiments. In a stepwise text mining approach, we characterize and compare these Contribution via (1) measures of vocabulary richness and lexical diversity, (2) distance-based approaches and (3) dimensionality reduction techniques. A special focus will be put on Martin Brun who is the author of most books with known Contributionhip, as it is very likely that he is the author of other pamphlets, due to his high productivity. This work will be accompanied by a website presenting interactive visualizations supporting researchers in the candidate selection task.

Aleksi Nicolas Moine

An Exploration of the Mythical Networks of Northern Karelian Incantations: Epistemological and Methodological Issues

Keywords *network analysis, incantations, folklore texts*

Contribution poster

Affiliation University of Helsinki, Finland

Abstract The practice of incantations was still part of everyday life in northern Karelia in the 19th century. Incantations mostly had a therapeutic or protective function. The charmer could for instance use incantations to heal a sick or wounded human body, to help find a partner, or to protect the cattle. In order to so, the charmer negotiated with non-human agents and forces, who all formed a form of social network that entangled also the human agent. Early folklore collectors and researchers were particularly interested in mythical knowledge of charmers and singers and collected, thus, a large corpus of incantations along other genres of oral poetry. These texts, varying from a few lines to a few hundred lines, have later been published and digitised in the Suomen Kansan Vanhat Runot (SKVR) collection.

Thus, the use of digital tools for the analysis of this collection of folklore texts is possible. In this paper, I propose a method of studying Northern Karelian

incantations based on social network analysis. I work on a limited, and yet thick, corpus of around 500 texts that were collected in the parish of Ilomantsi between 1816 and 1939. More precisely, I examine epistemological and methodological issues related to the analysis of this specific historical material. The extreme linguistic variation makes the computational analysis challenging and the nature of the texts needs to be taken into account.

I draw from the ontological turn in anthropology and consider the relationships between humans and non-human agents to be as meaningful as social relationships. Indeed, all these agents coexist in the same world, which is constructed through words and deeds. The incantatory practice has a communicative function, and it reasserts every time the relationships between agents, giving meaning to the world in a moment of crisis. My aim is, thus, to have a better understanding of the conception of the human body upon which the efficiency of charms relies, through an understanding of the network of agents. Who are the most called agents, and for what situations? What are the closest relationships between agents? The description of the network is necessarily partial due to the nature of the corpus, but gives an idea of the themes and texts that informants considered important. This has a huge heuristic value and brings some perspectives for a comparative analysis of folklore genres.

Liisa Maria Näpärä

Connecting researchers and digital collections at the National Library of Finland

Keywords *cultural heritage collections, research, national libraries, cooperation*

Contribution short paper

Affiliation National library of Finland, Finland

Abstract The National Library of Finland (NLF) and researchers have been collaborating for many years. However, during the last couple of years developing the research services has been of closer interest. Regarding the emergence of data-driven research and digital humanities, there have been several focus points that have required attention in order to be able to serve researchers. More information about the current state of the multidisciplinary researchers' needs has been acquired by collecting information from researchers and following international examples from library labs. Hereby, the new research service model was defined at the NLF. Despite the international examples, the lab is not the preferred term. In the Finnish language, the lab does not seem to have

a fitting translation in this context. Preferred terms are data services or research services depending on the perspective and emphasis.

Still, a lot of piloting and more learning by doing are needed as awareness of researchers' actions with the digital collections are increasing. For example, researchers' requirements for data they use in research are only at the beginning to slowly shape up. On the other hand, understanding all the possibilities for data that the NLF offers for research are accumulating recognition. Mutual understanding between researchers and the National Library of Finland is the key to engage researchers with its collections and bring digital humanities into action. Thus, collaboration with researchers and mutual knowledge sharing actively continues as copyrights remain a complicated subject.

The paper will provide insights into the research project pilots that have been going through the newly developed process at the NLF. It brings together experiences working in collaboration with researchers and cultural heritage research. Additionally, experienced issues in the process are considered. For example, data that is extracted from the collections of the NLF and used in the research process are taken into a closer look. Data biases as well as data FAIRness at the different stages of the research process, are noteworthy aspects.

Michael Neiß

3D laser scanning as a tool for artefact studies

Keywords *3D modelling, 3D laser scanning, artefact studies, archaeometallurgy, Viking studies*

Contribution short paper

Affiliation Uppsala University, Sweden

Abstract Due to the flood of digital innovations, it becomes increasingly difficult for archaeologists to monitor current technological trends and to integrate them into the framework of our discipline. Therefore, it seems more than natural to create synergies with experts from other disciplines. My paper will provide examples from a few colloborative pilot studies, which explored the utility of 3D modelling as a tool for analyzing objects of copper alloy and casting moulds from the Viking and Medieval world. Our results prove that 3D modelling is not only a valuable tool for iconographic documentation, artefact reconstruction and production analysis. It also paves the way for novel and creative research

strategies. For instance, 3D-laser scanning on debris from an early workshop site at Ribe led to a re-assessment of the chronological relationship between different production phases. This produced a substantially clearer image of this rare episode of metal casting and motivated a re-evaluation of the scale and time frame of the workshop activity. In consequence, 3D-model-based analysis appears to be useful in answering traditional questions in Viking Age artefact studies, and may also inspire innovative approaches when combined with other methods. Material culture is the cardinal source of archaeology. As such, it needs to be recognised as a natural starting point for theorizing research. A continuous re-evaluation of old finds, based on current methodological concepts, is therefore essential. And yet, the 'postmodern turn' that struck archaeology during the 1990s and onwards led to a significant downpriorisation of Viking artefact studies. This triggered a downwards spiral as archaeologists tried to extract new knowledge from ever more out-dated research literature, to the point that they started to handle material culture as illustrations for various theoretical arguments, rather than as a source material which has a story to tell. The new ’material turn’ of the last decade in combination with innovative digital strategies seem apt to put material studies back into their rightful place.

Seraina Nett, Rune Rattenborg, Carolin Johansson, Gustav Ryberg Smidt, Jakob Andersson

Here, There, and Everywhere: A Global Heritage Perspective on Cuneiform Culture

Keywords *Cuneiform, Cultural Heritage, Spatial Analysis*

Contribution long paper

Affiliation Uppsala University, Sweden

Abstract Counting a conservative half a million individual texts, the cuneiform corpus ranks among the largest discrete bodies of written sources from the ancient world. Cuneiform texts encompass a noticeably diverse range of genres, are found across most of the Middle East, and traverse some three millennia of written human history, from c. 3,000 BCE to the 1st century CE. Because of its immense size and extreme temporal and spatial spread, no attempt has ever been made to map and analyse this corpus in full.

The three-year research project Geomapping Landscapes of Writing (GLoW) based at Uppsala University and generously funded by Riksbankens Jubileumsfond, aims to produce an updated, global survey of cuneiform inscriptions in collaboration with existing digital text catalogues and open

access data repositories. Harnessing GIS-aided spatial analysis coupled with digital humanities research tools, the project explores our newfound technological ability to accurately capture, assess, and quantify the material imprint of this immense corpus. In doing so, we aim to make available an updated suite of attribute, spatial, and temporal metadata resources for free dissemination and reuse, as well as dedicated studies of corpus composition, linguistic landscapes, and the materiality of texts.

This paper will review the work of the project to date and offer some preliminary insights on the assembled dataset in the context of current trends in cuneiform studies. Coupling a suite of geospatial indices recently assembled by the project with text metadata, we will furthermore present an explorative global perspective on the relationship between the archaeological origins of cuneiform inscriptions and the museums and collections around the world that now hold these artefacts. Highlighting currents in the discovery, acquisition, and dissemination of the earliest written heritage of many modern nations of the Middle East, we conclude by evaluating and discussing the implications for understanding and approaching the cuneiform corpus as cultural heritage now and in the future.

Kristoffer Nielbo[1,2], Rebekah Baglini[1,2], Andreas Roepstorff[2]

Information Decoupling as a Pandemic Signature in News Media

Keywords *Newspapers, Pandemic Response, Change Detection, Adaptive Filtering, Information Theory*

Contribution long paper

Affiliation 1: Center for Humanities Computing Aarhus, Jens Chr. Skous Vej 4, Building 1483, 3rd floor, DK-8000 Aarhus C, Denmark;
2: Interacting Minds Centre, Jens Chr. Skous Vej 4, Building 1483, 3rd floor, DK-8000 Aarhus C, Denmark

Abstract As the first wave of Covid-19 virus spread across the world, content alignment of news stories could be observed both within and between media sources. Initially, news stories were interspersed with news coverage of other events, but as the virus spread across the world, news media focused almost exclusively on the pandemic. From the perspective of cultural dynamics, the Covid-19 pandemic provides a natural experiment that allows us to study the effect of a global catastrophe on the the dynamics of news media's information. While news media are neither unbiased nor infallible as sources of events, they do reflect preferences, values, and desires of a wide socio-cultural

and political user spectrum. As such, news media coverage of Covid-19 functions as a proxy for how cultural information systems respond to unexpected and dangerous events.

Previous studies have shown that variation in newspapers' word usage is sensitive to the dynamics of socio-cultural events [1, 2], can be used to model effects of change [3], and can accurately capture thematic development indicative of the evolution of cultural values and biases [4-5]. Recently, a set of methodologically related studies studies have applied windowed relative entropy to text representations in order to model information novelty as a reliable content difference from the past and resonance as the degree to which future information conforms to said novelty [6, 7]. Recent studies have found that successful social media content show a strong association between novelty and resonance, and that variation in the novelty-resonance association can predict significant change points in historical data [8, 9].

In this study we expand upon studies of novelty and resonance in cultural dynamics by modeling change in printed news media during the initial phase of Covid-19. Specifically, we propose an empirically derived principle of News Information Decoupling, which can explain how the information flow in news media responds to events of a catastrophic nature.

Bibliography

[1] J. Guldi, The Measures of Modernity, International Journal for History, Culture and Modernity 7 (2019) 899–93.

[2] M. Kestemont et al, Mining the Twentieth Century's History from the Time Magazine Corpus, in: Proceedings of the 8th LaTeCH 2014.

[3] P. Bos et al, Quantifying Pillarization, Proceedings of the 3rd HistoInformatics Workshop (2016) 10.

[4] M. Wevers, Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990, in: Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, 2019.

[5] M. Wevers et al., Tracking the Consumption Junction: Temporal Dependencies between Articles and Advertisements in Dutch Newspapers, DHQ 14 (2020).

[6] A. T. J. Barron et al, Individuals, institutions, and innovation in the debates of the French Revolution, PNAS 115 (2018).

[7] J. Murdock et al, Exploration and Exploitation of Victorian Sciencein Darwin's Reading Notebooks, arXiv:1509.07175 (2015).

[8] K. Nielbo et al., Trend Reservoir Detection: Minimal Persistence and

Resonant Behavior of Trends in Social Media, Proceedings of Computational Humanities Research 1 (2021).

[9] E. Vrangbæk et al, Composition and Change in De Civitate Dei: A Case Study of Computationally Assisted Methods, Studia Patristica (2021).

Lars Oestreicher[1], Jan von Bonsdorff[2]

From Visual Forms to Metaphors - Targeting Cultural Competence in Image Analysis

Keywords *Multi-modal machine learning, High-level image content, Visual metaphors, Cultural competency, Pictorial conventions*

Contribution short paper

Affiliation 1: Department of Information Technology, Uppsala University, Sweden;
2: History of Arts Department, Uppsala University, Sweden

Abstract Image analysis has taken a large step forward with the development within Machine Learning. Today, recognizing images as well as constituent parts of images (faces, objects, etc.) is a relatively common task within machine learning. Currently, it is even possible to perform real time item recognition (not training), using a RaspBerry Pi with a standard camera (M.Sc. Thesis reference). However, there is still a big difference between recognizing the content of a picture and understanding the meaning of the image.
In the current project we have set out to take an interdisciplinary approach to this problem, including Art history, Machine Learning and Computational Linguistics. Current approaches pay large attention to the details of the image when trying to describe what's in the picture, resulting, e.g., in that smiling faces will support the interpretation of the image as "positive" or "happy", even of the picture itself is a scary scene, where the smiles are increasing the scare in the scene. Other problematic issues are metaphors, sarcasm and pragmatic interpretations, e.g., enabling humoristic interpretations of a picture. In short, we hope to be able to identify the agency of the pictures, i.e., what the pictures want to tell us on a higher level of interpretation. For this purpose we intend to combine different machine learning approaches, including the combination of semantic models for image analysis, such as CLIPS (by openAI), with Natural Language Processing models, similar to those used for sentiment analysis, e.g., of movie reviews.

We have randomly scanned issues of the richly illustrated monthly and weekly journals of the Swedish magazines "Veckojournalen" and "Bonniers Månadstidning" from the 1920's to the 1950's, altogether 180 issues. Each journal contains an average of 70 usable images, which would make up a dataset of ca. 12,600 images.

The first problem we want to tackle during this first phase is finding the most effective method for handling a large number of both annotated and unannotated images. Connected to this first problem is also to investigate the (semi-)automated extraction of our image-and-text datasets.
The second problem we will deal with is how the annotations of an image need to be prepared so that the ML system can determine, not only what is shown, but what meaning the image is attempting to purvey. We ask ourselves what formations in the picture would be optimal for the learning process of the neural networks, and believe that agency descriptions, visual narratives and metaphors are highly relevant.

Thus, our objective for the first phase of the project is to investigate the capacity of the multi-modality models' for recognizing such high-level image content as for example context, agency, visual narration, and metaphors.
In the paper we will further describe our current approach, the general ideas behind it, and the methods that will be used. We will also try to problematize the task at hand, looking at possible difficulties on the way.

Patrik Öhberg, Daniel Brodén, Mats Fridlund, Victor Wåhlstrand Skärström, Magnus P. Ängsal

A Unifying or Divisive Fear? Terrorism in Swedish Public Opinion and Parliamentary Motions 1986–2020

Keywords *terrorism studies, language technology, parliamentary data, survey data, politcians*

Contribution long paper

Affiliation University of Gothenburg, Sweden, Sweden

Abstract This paper contributes to the digital history of political terrorism through a study of the concerns regarding the threat of terrorism and political extremism among the Swedish public and Members of the Parliament (MPs). The aim is to explore the intersection of public opinion as expressed in national survey data and data on political action in the form of motions by Members of the Parliament, focusing on trends over time and the significance of political sympathies. Our study points to the impact of the attacks in the USA in 2001 and the Drottninggatan truck attack in 2017 on both the public's anxiety and the MP's activity. It also shows the significance of political sympathies in this context with both citizens and MPs on the right being more concerned with the issue of terrorism than those on the left, while the pattern is repeated but in reverse when it comes to concern about political extremism. Through the analysis we highlight the benefits of combining parliamentary data and survey data as well as the importance of the parliamentary context itself, in exploring

the relationship between public opinion and MPs activity on terrorism. An argument underlying the investigation is that the analysis of parliamentary data should be grounded in the context of the institutional and historical framework in the political system.

Emily Sofi Öhman

SELF & FEIL: Reusing emotion lexicons for multilingual emotion detection in interdisciplinary projects

Keywords *Emotion analysis, multilingual, lexicon, annotation, sentiment analysis*

Contribution short paper

Affiliation Waseda University, Japan

Abstract This project introduces two emotion lexicons: SELF (Sentiment and Emotion Lexicon for Finnish) and FEIL (Finnish Emotion Intensity Lexicon). The lexicons use annotation projections from EmoLex (Mohammad, 2010, 2013, 2017) with carefully edited translations. Although data-driven solutions are increasingly popular in applied language technology and even digital humanities projects, lexicon-based approaches are still used and can even be more appropriate and more useful for certain types of projects (Öhman, 2021). To my knowledge, this is the first comprehensive sentiment and emotion lexicon for Finnish. It is freely available on GitHub.

Many lexicons for emotion detection exist, but these are mostly in English. To create SELF and FEIL the lexical items in EmoLex and the related intensity lexicon were re-translated and systematically edited to mitigate the effects of the context-free single-word translations produced by Google Translate. All the adjustments resulted in an overall lexicon size of 12,448 entries in Finnish for SELF and 7291 entries for FEIL, roughly a 10% overall reduction in lexicon size from the originals.

The most common correction was when the target word for two semantically related words in English was translated into the same word in Finnish but an alternative translation could not be found (>40%), a further 35% were similar corrections but an alternative translation was possible. Other common corrections included the mistranslation for sense (birch to birch tree instead of flogging) and specificity (emaciated to laihtunut instead of riutunut). There were also a number of instances where the part-of-speech had changed or the cultural connotation was different.

The lexicon has successfully been used on several interdisciplinary projects.

This includes the analysis of emotion intensity in political manifestos (Koljonen et al., forthcoming), computational literary analysis (Öhman & Rossi, forthcoming), and diachronically tracking attitudes towards health authorities during the COVID-19 crisis (Pääkkönen et al., forthcoming). SELF and FEIL were shown to perform well on fine-grained tasks in particular where a rich, fairly informal, emotive lexicon is common.

This approach to lexicon creation by re-using existing annotation work to lower the cost of creation has been shown to be a reliable way of conducting lexicon-based emotion analysis, especially when used in conjunction with other natural language processing and statistical tools. Furthermore, the methods employed in the creation of this lexicon can be used as a guide to creating similar lexicons in many other languages with far fewer resources required compared to creating such lexicons from scratch or for training datasets for machine learning. The next steps include expanding the lexicon with domain-specific annotations and revisions, as well as increasing the number of languages it is available in.

Bibliography

Koljonen, J., Ohman, E., Ahonen, P., & Mattila, M. Forthcoming. “Strength and Intensity of Sentiments and Emotions in Party Manifestos: Finland 1945-2019.”

Mohammad, S. and Turney, P., 2010, June. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text (pp. 26-34).

Mohammad, S.M. and Turney, P.D., 2013. Crowdsourcing a word–emotion association lexicon. Computational intelligence, 29(3), pp.436-465.

Mohammad, S. and Bravo-Marquez, F., 2017, September. WASSA-2017 Shared Task on Emotion Intensity. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 34-49).

Pääkkönen, J., Öhman, E., Laaksonen, S-M. Forthcoming. Unconventional Communicators in the COVID-19 Crisis.

Öhman, E. & Rossi, R. Forthcoming. Affect and Emotions in Finnish Literature: Combining Qualitative and Quantitative Approaches.

Öhman, E. 2021. The Validity of Lexicon-based Emotion Analysis in Interdisciplinary Research. In Proceedings of ICON’21 workshop NLP4DH.

Teemu Tapani Oivo

Environmental knowledge, media producers and technology, in Finnish-Russian border regional media flows

Keywords *Environmental journalism, digital journalism, knowledge visualization*

Contribution short paper

Affiliation University of Helsinki, Finland

Abstract Media sources play an essential role in public opinion formation on energy and waste-related knowledge: the way how media narrate a particular environmental event influences people's informed decision-making. The epistemology of journalism revolves around their ability to conduct knowledge-work and present consistently reliable, factual, and valuable information for the citizens. In a recent reaction to the modern environmental problems, Latour (2018: 23) pointed to the core of the problem: 'Facts remain robust only when they are supported by a common culture, by institutions that can be trusted, by a more or less decent public life, by more or less reliable media'. As a member of a larger research project "Flowision", I explore how the mediatized knowledge on fossil and renewable energy and waste is produced and distributed by Finnish and Russian media and how technological tools in the media influence the visibility of energy and waste-related knowledge. What are the interrelations of social actors and technology development of the relevant media field? My conference paper introduces work in progress in this project. In a preliminary case study, I have selected news events in Finnish-Russian border regions to examine their distribution through electronic media outlets. To elaborate on epistemology and representation of media outputs on energy and waste-related news events, I seek to learn how the producers deal with technological factors in their work. Overall, this study seeks to highlight combinations of material-semiotic assemblages in media producers that support mediatized representation and public understanding of current energy and waste issues in Russia and Finland.

Arttu Oksanen[1,2], Minna Tamper[2], Eero Hyvönen[2,3], Jouni Tuominen[2,3,4], Henna Ylimaa[5], Katja Löytynoja[5], Matti Kokkonen[5], Aki Hietanen[6]

A Tool for Pseudonymization of Textual Documents for Digital Humanities Research and Publication

Keywords *pseudonymization, anonymization, data protection, named entity recognition*

Contribution poster

Affiliation 1: Edita Publishing Ltd.;
2: Aalto University, Dept. of Computer Science;
3: University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG);
4: University of Helsinki, Helsinki Institute for Social Sciences and Humanities (HSSH);
5: Statistics Finland;
6: Ministry of Justice in Finland

Abstract Much of Cultural Heritage (CH) texts and legal documents of interest to a wider audience, such as interviews of people and court decisions, contain sensitive personal data. This makes it difficult to publish and use them for research given the EU General Data Protection Regulation (GDPR), unless personal data contained is disguised. However, manual pseudonymization or anonymization of the documents is a costly and time-consuming procedure. This paper presents the Anoppi tool and web service for automatic and semi-automatic pseudonymization of Finnish documents. Utilizing both statistics and rule-based named entity recognition methods (NER) and morphological analysis, Anoppi is able to automatically or semi-automatically pseudonymize documents written in Finnish preserving their readability and layout. Anoppi is the first pseudonymization tool developed for Finnish; some related works in other languages will be discussed.

The tool consists of a language analysis component (LAC) and a web-based user interface (WUI). LAC performs the actual automatic pseudonymization by carrying out NER and disambiguation on the document while WUI allows the user to modify the result of the automatic pseudonymization. The pseudonymization is done by replacing the names with grouped sequential identifiers, such as ‘person A’ or ‘place B’. By performing morphological analysis on the original documents the software is also able to inflect the generated pseudonyms correctly to improve the readability of the pseudonymized text. LAC aims to maximize recall in the named entity recognition phase as it is easier for the user to delete suggested entities than manually pick new entities from the text, if LAC did not automatically recognize them.

Evaluation of Anoppi shows promising results in locating the names of persons, Authors , places, and different types of identifiers of specific form. The service is in pilot testing in the Ministry of Justice of Finland for pseudonymization of Finnish court decisions in order to make them available on the Web and for

data analysis in the forth-coming public LawSampo data service and portal for publishing and studying Finnish legislation and case law. Future work focuses on identifying things such as rare diseases or unique job s that make reidentification of people straightforward.

Eljas Oksanen[1,2,3], Heikki Rantala[3], Jouni Tuominen[4,2,3], Michael Lewis[5], David Wigg-Wolf[6], Frida Ehrnsten[7], Eero Hyvönen[3,2]

Digital Humanities Solutions for pan-European Numismatic and Archaeological Heritage Based on Linked Open Data

Keywords *Semantic Computing, ontologies, pan-European archaeological data harmonisation, numismatic/archaeological culture heritage, museum collections management*

Contribution short paper

Affiliation 1: Department of Cultures, University of Helsinki, Finland;
2: Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland;
3: Semantic Computing Research Group (SeCo), Aalto University, Finland;
4: Helsinki Institute for Social Sciences and Humanities (HSSH), University of Helsinki, Finland;
5: Portable Antiquities Scheme (PAS), British Museum, UK;
6: Römisch-Germanische Kommission des Deutschen Archäologischen Instituts, Germany;
7: National Museum of Finland, Finnish Heritage Agency, Finland

Abstract This paper presents the interdisciplinary research project DigiNUMA that investigates challenges and solutions in data management and dissemination of pan-European Cultural Heritage. The project develops a new model for harmonising national and international archaeological datasets for Digital Humanities (DH) analysis as well as public dissemination through Linked Open Data (LOD). DigiNUMA answers challenges and opportunities created by the digitisation of society:
1) The need for digital solutions in Cultural Heritage management stemming from the vastly increased amount of information generated by the public, with particular reference to the growing number of archaeological finds recovered by metal-detecting and other public finders in European countries including Finland and the UK.
2) The pan-European need to develop an internationally operable and

harmonised LOD infrastructure for using Cultural Heritage data from different countries in research.
3) Increasing the accessibility of Cultural Heritage data among different audiences, including outside the scientific community.

The project develops an infrastructure for transnational Cultural Heritage data management and dissemination based on ontologies extracted from the classifications and typologies used for describing Cultural Heritage artefacts. DigiNUMA examines the potential offered by data harmonisation strategies in developing digital heritage services. Ontological work will be complemented by user experience research on international public heritage portals, in order to develop optimal solutions for structuring and disseminating heritage data. DigiNUMA extends the FindSampo framework [1], Sampo model (https://seco.cs.aalto.fi/applications/sampo), and the new FindSampo system (see https://loytosampo.fi) into a transnational technical solution for Cultural Heritage data management and dissemination. As a case study, the rich and complex numismatic data from Finland (Finnish National Museum) and the UK (British Museum and FitzWilliam Museum) are investigated. The functionality of the resulting data models will be tested through intensive DH analyses (e.g., Geographic Information Systems analysis, multivariate statistical analyses, spatial network analysis) with a focus on Viking Age numismatics (AD 800–1150). During this historical period coin circulation was particularly international, creating socio-economic links between countries in Europe and beyond, and underlining the necessity to bring numismatic data (like other archaeological material) together from trans-national sources in order to better appreciate world-historical large-scale patterns in economic growth and monetisation.

DigiNUMA produces (1) a Nomisma.org-conforming ontology of Viking Age coin data; (2) new internationally relevant research on coin use and contacts between monetised and non-monetary societies during the Viking Age; (3) theoretical research on models suitable for international heritage services for numismatic data, with direct relevance for all archaeological data. The project collaborates with other pan-European LOD data harmonisation projects, especially ARIADNEplus (https://ariadne-infrastructure.eu/) (all archaeological data) and Nomisma.org (specifically numismatic data).

Bibliography

[1] Eero Hyvönen, Heikki Rantala, Esko Ikkala, Mikko Koho, Jouni Tuominen, Babatunde Anafi, Suzie Thomas, Anna Wessman, Eljas Oksanen, Ville Rohiola, Jutta Kuitunen and Minna Ryyppö: Citizen Science Archaeological Finds on the Semantic Web: The FindSampo Framework. Antiquity, A Review of World Archaeology, 95(382), Cambridge University Press, 2021.

Siim Orasmaa, Kristjan Poska, Kadri Muischnek, Anna Edela

Named Entity Recognition in 19th Century Communal Court Minute Books

Keywords *historical texts, corpus annotation, named entity recognition*

Contribution short paper

Affiliation University of Tartu, Estonia

Abstract The National Archives of Estonia holds a crowdsourcing project to transcribe Estonian 19th century communal court minute books1. This is a large historical resource, which sheds light to everyday lives of the peasantry: which minor offences they were tried for, and how did they solve their civil disputes. The material offers a wide range of opportunities for research, allowing historians to trace family history and study peasant life of that time period, and linguists to study language change and regional variations. The crowdsourcing task also includes manual annotation of the named entities (persons and locations) in the transcribed documents. Named entity (NE) annotation enables better search and browsing of the collection, focusing on names mentioned in documents; and annotated names could also be used in a more advanced network analysis. However, as the main focus of the crowdsourcing project is the transcription, the quality of manual NE annotation is heterogeneous and varies greatly from annotator to annotator. Systematising the NE annotation process with automatic pre-annotation is desirable.

We present the work on developing new annotation guidelines for NE annotation in the minute books, introduce a manually annotated NE corpus and discuss the machine learning experiments on that corpus. The main characteristics of this corpus that make annotation challenging are the uneven distribution of entities (84% of all names are person names, names from other categories are scarce), and ambiguity of location names. Location names can refer both to the location, e.g. a farm, and to the “Authors ” related to location, e.g. people living and working in a farm. So we distinguish between regular locations (such as countries and towns), regular Authors (such as names of the higher court instances) and ambiguous locations-Authors (farms, manors, villages etc entities that were frequently mentioned as communities).

We have manually annotated 1500 minute books. To validate our guidelines, we measured inter-annotator agreement f-score (see Hripcsak & Rothschild, 2005) on 250 documents. Annotators following our guidelines obtained mean agreement f-score 0.95, while agreement with crowdsourcing annotators was only 0.68.

Finally, we present our first machine learning experiments on NE recognition. We have retrained EstNLTK’s named entity recognizer (Laur et al. 2020) on the manually annotated corpus, and analysed the effect of different (linguistic/non-

linguistic) features on the machine learning performance. Our best model reaches f-score 0.90 on a test set. For comparison, EstNLTK's default named entity recognizer (trained on nowadays newspapers) obtains only f-score 0.56 on the test set.

Bibliography Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association, 12(3), pp. 296-298.

Laur, S., Orasmaa, S., Särg, D., & Tammo, P. (2020). EstNLTK 1.6: Remastered Estonian NLP Pipeline. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 7152-7160). https://www.ra.ee/vallakohtud/

Jacob Orrje

A digital history of a scientific academy. Exploring the actors of the Royal Swedish Academy of Sciences 1742–1826

Keywords *Digital history, digitization, historical actors*

Contribution short paper

Affiliation Center for history of science (Royal Swedish Academy of Science), Department for history of science and ideas (Uppsala university)

Abstract What might a digital and actor-centric history of the eighteenth-century Swedish Royal Academy of Sciences look like? In an emerging pilot study, I turn a twentieth-century typewritten edition of the academy's eighteenth-century and early-nineteenth-century protocols (c. 3800 pages) into accurate text files using the open-source software OCR4All. Using the same process, and a custom Python script, I turn an index of mentioned persons into a digital database. By using the database to tag mentioned actors in the text, I survey new ways of writing the academy's history.

From its practical turn in 80s and 90s, historians of science have had a particular focus on actor networks, and on highlighting previously neglected participants in making scientific knowledge. Countless microhistories have rewritten the roles that women, artisans, and subalterns played in the making of modern science. More recently, such detailed studies have been criticized for being too narrow and synchronic by historians demanding a new big picture. Moreover, digital historians have argued that diachronic narratives, which both cover long timelines and maintain the detail of previous microstudies, might be realized by leveraging digital methods.

Using my dataset, I thus explore ways of writing a new history both of the central and more peripheral knowledge actors that have surrounded the academy. What categories of historical actors were mentioned during the

academy's first eighty years? Can we find clusters of actors that tend to be mentioned together? Did the categories of actors mentioned in the protocols change of time, as the academy's role in Swedish society changed and scientific practice was professionalized? Finally, I discuss how this relatively small dataset could be supplemented, so that we may create a longer and more detailed account that extends into the twentieth century.

Petri Paju[1], Hannu Salmi[1], Heli Rantala[1], Patrik Lundell[2], Jani Marjanen[3], Aleksi Vesanto[1]

Textual Migration across the Baltic Sea: Creating a database of text reuse between Finland and Sweden

Keywords *text reuse, historical newspapers, digital collections, transnational history, database construction, Sweden, Finland*

Contribution short paper

Affiliation 1: University of Turku, Finland;
2: Örebro University;
3: University of Helsinki

Abstract The short paper presents a database and an interface on text reuse between newspapers and journals published in the Swedish language in Sweden and Finland in the period 1645–1918. The aim of the database and the accompanying project is to study information flows between the two countries from the period when present-day Finland was part of the Swedish kingdom to the establishment of Finland as a Grand Duchy in the Russian Empire after 1809 and until 1918. Even after the 1809 separation, news and other texts circulated because of the common cultural heritage and the shared language, Swedish. The border was relatively easy to cross, and newspapers were delivered from Sweden to Finland and vice versa. Because of later, separate national histories, however, these press materials have been preserved, processed and siloed in two national libraries. The digitization of the press makes it possible to study overlaps in the text masses and thereby analyse how information was spread across the Baltic Sea. In our project, textual migration was traced with a method based on the software BLAST, which can be applied to text reuse detection.

The results of text reuse detection are presented in the form of a database and its interface, which include all detected text reuse passages and provide clusters of repeatedly reused passages. The database bears the name Text Reuse in the Swedish-language Press, 1645–1918 and it has been accessible online since October 2021.

The materials consisted of more than five million pages of digitised content, 1.79 million pages from Finland and 3.24 million pages from Sweden. The total number of clusters of text reuse found is approximately 17.8 million, out of

which 2.4 million clusters are shared between countries. These clusters and their texts comprise the database which includes texts from over 1100 s of newspapers and journals published on circa 150 locations at some time during the time period of 273 years.

The interface was designed as an easy tool to explore the passages as well as the, sometimes viral, clusters they belong to. It allows for studying both long-term and short-term text reuse, text reuse between different countries, towns and newspaper s, and reuse within the countries. A map function illustrates potential viral chains. Overall, the database and interface provide ample opportunities for studying concrete cases of virality and cultural mediation between Sweden and Finland, but also provides a way of quantitatively assessing the cultural asymmetries present between the two countries in the nineteenth century.

The paper further discusses critical issues that are involved in combining historical newspaper collections from countries with different size and historical trajectories. It further raises the issue of how the different choices made in the digitization process of the two newspaper collections influences the ways in which these data sets can be connected and further processed. The paper concludes in the assessment of both benefits and pitfalls of a database as a historical representation.

Niko Partanen[1], Rogier Blokland[2], Michael Rießler[3], Jack Rueter[1]

Transforming archived resources with language technology: From manuscripts to language documentation

Keywords *documentary linguistics, language technology, text recognition, forced alignment, Zyrian Komi*

Contribution short paper

Affiliation 1: University of Helsinki;
2: Uppsala University;
3: University of Eastern Finland

Abstract Transcriptions in different languages are an ubiquitous data format in linguistics and in many other fields in the humanities. However, the majority of these resources remain both under-used and under-studied, even when they have been published in print. Our paper presents a workflow adapted in the research project Language Documentation Meets Language Technology, which combines text recognition, automatic transliteration and forced alignment into a process which allows us to bring earlier transcribed documents to a structure

that is comparable to contemporary language documentation corpora. This has complex practical and methodological considerations.

Natalia Perkova[1], Kirill Kozhanov[2]

Towards the corpus of Latvian Romani texts: deciphering the manuscripts from Jānis Leimanis' archive

Keywords *Romani studies, corpus linguistics, handwritten text recognition, digitalization, crowdsourcing*

Contribution short paper

Affiliation 1: Uppsala University, Sweden;
2: Södertorn University, Sweden

Abstract This study focuses on the development of a corpus for one of the undestudied Romani dialects, Latvian Romani (Lotfitka Roma). It belongs to the group of Northeastern Romani dialects (Matras 2002; Tenser 2008) and is spoken in Latvia, Estonia and Lithuania. There exist only a few published texts in Latvian Romani available both for the wider public and specialists; the data in the existing dictionary (Mānušs, Neilands & Rudevičs 1997) and the Romani Morphosyntax Database (Matras, White & Elšík 2009) are represented by words and separate phrases or sentences. Our attempt of creating a corpus has been encouraged by recent digitalization initiatives in several countries (Estonia, Latvia, and Finland) which resulted in providing open access to the two important archives with Latvian Romani texts, both compiled before the World War II: the collection by Jānis Leimanis, a prominent Latvian Romani personality of the interwar Latvia, and the collection by Paul Ariste, a brilliant Estonian linguist.

In the 1930s Jānis Leimanis collected Romani folklore for the Latvian Folklore Archive (Latvijas Folkloras Krātuve, LFK), and his archive comprises 75 copybooks (three of them not currently available), with about 500 folklore units of different genres (463 of them accessible) and 1254 manuscript pages. This collection is rather unique, as all Romani texts have translations into Latvian, which makes it possible to compile a parallel corpus with them; overall, 884 pages contain such bilingual texts. The archive has been digitalized by the LFK and uploaded into their crowdsourcing platform garamantas.lv, launched in 2014 (Reinsone 2020) as a separate collection (http://garamantas.lv/lv/collection/886320/Jana-Leimana-ciganu-folkloras-vakums). Since then, some files have been deciphered by volunteers, but by October 2020 it was only about 25% of all files and only about 21% of files with a text in Romani.

In 2021, we started moving towards the more time-efficient pipeline with automatic text recognition for Jānis Leimanis archive. Obviously this collection has advantages in both its reasonably large size and the consistent use of the same handwriting. We use Transkribus (Kahle et al. 2017) for work with the collection and its deciphering. First of all, we copied transcriptions for those pages that had been deciphered by volunteers and were available at garamantas.lv; after adjusting them to the line-by-line layout, these transcriptions became our ground truth data. We trained two HTR models: Model 1 (235 pages, a training set of 25890 words and 4911 lines, 50 epochs) gave CER of 8.44% on the validation set, while Model 2, with only several more pages of ground truth data added improved CER up to 3.95% (a training set of 16952 words and 3136 lines; uses Model 1 as its base model, 100 epochs). The latter model has been used since then to decipher other texts in the collection, notably accelerating work on the corpus.

We are going to continue our work with automatic HTR followed by manual correction and hope to be ready with at least the bilingual part in the nearest future, first of all with its 65 longer narrative texts (fairytales and stories).

Bibliography

Kahle, P., Colutto, S., Hackl, G. and Mühlberger, G. 2017. Transkribus – a Platform for Transcription, Recognition and Retrieval of Document Images. IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE, pp. 19–24.

Ginta Perle-Sile, Sanita Reinsone

Transcription as a Tool for Deep Reading and Teaching of Folklore

Keywords *deep reading, digital tools for humanities, digital humanities, transcription, actualisation of cultural heritage*

Contribution short paper

Affiliation Institute of Literature, Folklore and Art of the University of Latvia, Latvia

Abstract Development and increasing use of modern technologies has given opportunity to use different tools in education process. On the other hand, they have had an impact on the perception of what is read, encouraging scanning and skimming, as opposed to a critical and analytical reading of the text. Deep reading as an activity that promotes immersion and the understanding of long and complex texts becomes particularly important in the light of increasing trend of text reading digitally. Transcription as a tools that stimulates re-reading, interaction with a text as whole, has the potential to fit into different instrumental models in education, i.e. complex of lesson structure and methods, for reading elaboration.

The aim of the paper is to reflect the experience of using transcription as part of an instrumental model to promote deep reading and increase students' knowledge and interest in folkloristics. The paper shares first results of a pilot study carried out in 2020–2021 on transcription as a tool for deep reading in folklore studies. Behind this study is a project aiming to improve folk song database developed by the Archives of Latvian Folklore. It involved students from different study programs of humanities in Latvia. Their task consisted of checking the electronically available data on the comprehensive publication of Latvian folk songs “Latvju Dainas” which forms the basis of the database in question, transcribing (i.e. entering data from) scanned images of original manuscript, as well as adding different attributes for each text unit and comparing published and transcribed texts. A specially designed editing tool was developed for this task.

The feedback provided by the students, analyzed together with individual interviews, provides a valuable insight into the motivation for participation, as well providing information on the disincentives to continue working. The pilot study shows the impact of family and school on the understanding of folk songs as part of a nationally significant cultural heritage. At the same time, it is also evident that taking part in the project impacts the perception of folk songs and, generally,improves comprehension of particular cultural heritage. During the interviews, students expressed their discovery of the folk song as a "living", changing, adaptable tradition. Study tackles also challenges when implementing the task that can be improved in further work resulting in development of new instrumental model for development of deep reading and interaction with cultural heritage.

Nadezhda Povroznik

Museum Digital Identity: Building a New Vision of Museum Functions in Virtual Environments

Keywords *museums' digital identity, museum functions, digital environment, virtual museums, web history*

Contribution short paper

Affiliation Perm State University, Russian Federation

Abstract The paper focuses on the new concept of the museum's digital identity which can be considered as a vision of the museums themselves and the museum functions in the virtual environments. The author outlines the specifics of the museum's digital identity, defines its’ components, and traces the evolution of the museum websites through the recent 25 years. The research is based on the analysis of the snapshots of the museum's websites preserved in the web

archives. The author articulates the museum's digital identity at the collective and individual levels. The individual level of identity is associated with respective museums and their activities in the digital environment, and more attention has been paid to it. At the individual level, the components of museum digital identity are determined, which are associated with an understanding of the value of digital museum content, the importance of investments in the development of virtual museums as an equal part, a virtual extension of a physical museum.

Q-Ö

Heikki Rantala[1], Esko Ikkala[1], Mikko Koho[1,2], Ville Rohiola[4], Eljas Oksanen[2,5], Jouni Tuominen[1,2,3], Eero Hyvönen[1,2]

FindSampo: A Linked Data Based Service for Analyzing and Disseminating Archaeological Finds

Keywords *Digital Citizen Science, Archeology, Semantic Computing, Linked Open Data, Data Analysis*

Contribution poster

Affiliation 1: Semantic Computing Research Group (SeCo), Aalto University, Finland;
2: HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland;
3: HSSH – Helsinki Institute for Social Sciences and Humanities, University of Helsinki, Finland;
4: Finnish Heritage Agency (FHA), Archaeological Collections, Finland;
5: Department of Cultures, University of Helsinki, Finland

Abstract The popularity of recreational metal detecting has grown rapidly in many countries such as in Finland during the last decade, creating a large amount of new archaeological data. This paper demonstrates how archaeological object finds made by the public can be analyzed using the Linked Open Data (LOD) based FindSampo service. A demonstrator based on data of some 3000 archaeological object finds catalogued in the archaeological collection of the Finnish Heritage Agency (FHA) from 2015 to 2020 has been open to public use since May 2021. FindSampo demonstrator consists of a data service (see: https://www.ldf.fi/dataset/findsampo) and a semantic portal (see: https://loytosampo.fi/en) with search functions and analytical tools.
The archaeological data included in the service has been recovered by the public mainly by metal-detecting and reported to the FHA for recording. FindSampo data constitutes an unprecedented reservoir of citizen science-generated archaeological heritage in Finland, equally accessible to researchers and to the broader public interested in heritage. The new archaeological object finds material it contains has proven to hold the potential for substantially adding to our understanding of the Finnish prehistorical and historical periods. The data is stored and published as Linked Data using a simple data model, and it can be queried by anyone with SPARQL query language. The published data follows the EU General Data Protection Regulation (GDPR), and details relating

to the identity of the finder are not included. The key properties of object finds in the data include object type, material, dating, and the location where the object was found. Values of object type, material, and dating are expressed using ontologies based on the MAO/TAO – Ontology for Museum Domain and Applied Arts. Applying the framework to data from other countries would be relatively simple in terms of the data model. However, the terminology used especially for object types varies between countries, and this would create challenges to fully harmonizing, for example, all the Nordic data.

The FindSampo portal is based on the Sampo-UI (https://github.com/SemanticComputing/sampo-ui) framework and FindSampo data service. The portal can be used to filter the data with faceted search and to visualize the data with various tools. Visualization tools include various charts and maps. Combining faceted search with visualizations enables the user to analyze the data quickly, and with very little technical knowledge. FindSampo increases the democratization of archaeological heritage data and opens up new possibilities for involving different audiences in knowledge discovery and creation.

Krista Stinne Greve Rasmussen, Jon Tafdrup, Kim Steen Ravn, Katrine Frøkjær Baunvig

The case for scholarly editions

Keywords *data validation, data reliability, data infrastructure*

Contribution short paper

Affiliation Center for Grundtvig Studies, Aarhus University, Denmark

Abstract The digitization of texts often transforms large collections of texts into digital corpora by means of OCR (Optical Character Recognition). This straightforward procedure is the foundation for valuable mega-scaled projects such as the Google Books and for meso-scaled projects such as the digitization of Danish newspapers published from 1666 and onwards named Mediastream (Mediestream) at The Royal Library in Copenhagen. Having received only cursory proofreading (or perhaps none at all) it is, nevertheless, well-known that material such as this contain variating degrees of 'dirt' in the form of transcription errors.

The big and dirty corpora stand in opposition to smaller, cleansed and philologically digested data-sets that ensure research validity. For a computational cultural historian the former can, because of their large scale, be used to paint a blurry background of semantic trends in a given period, despite the 'dirtiness' of the data. However, the often smaller, but purer and

philologically handled data helps to paint the foreground, which is more accurate and greater in details.

It goes without saying that extensive scholarly editions are increasingly more difficult to fund. Grundtvig's Works (2010-) is the largest scholarly edition ever produced in Denmark. The corpus consist of 37,000 standard pages of 2400 units per page spread across 1073 works. The scholarly digital edition (SDE) has a 20 years' period, and so far, the project is on time. The input is a large 'dirty' data set comprised of digitized and OCR-treated printed texts and the output is not just a clean dataset, but also a scholarly digital edition with introductions, commentaries and several tools such as diverse indices of named identities, such as persons, places, mythological entities and biblical references.

In this short presentation, we want to make a case for digital scholarly editions. We want to argue that a scholarly digital edition can help to put the computational humanities into the center of humanistic scholarship. Large scaled projects such as Google Books, the HathiTrust Digital Library, the Project Gutenberg and the ever-increasing amount of national digital archives based primarily on ORC technology will continue to pave the way for easy access to historical documents. For this they must be applauded and appreciated. But the most likely scenario is that small, curated digital editions, strenuously enriched by philologists, will be the key player in the process of integrating computational humanities into traditional scholarship; first and foremost because this material is clean, reliable, and flexible: Thorough markup leaves it open to comprehensive, fine-grained, hermeneutically complex explorations.

Annika Rockenberger

Digital Humanities in the Nordic and Baltic Countries - A Living Bibliography

Keywords *bibliography, DHNB, history of DH*

Contribution poster

Affiliation University of Oslo Library, Norway

Abstract In late 2019, I started collecting material on DHNB as an organisation, a community of researchers and practitioners, and a central facilitator of digital humanities knowledge. Though DHNB was only six years old, I met many obstacles collecting its conference outputs. It was challenging to gather information about how the Nordic-Baltic, the international DH community, and the wider Humanities and Social Sciences communities received the organisation and its conferences. Webpages for previous conferences had died,

books of abstracts were hard to come by, and blog posts, tweets, newsletters, and other forms of web-based publications about DHNB and its conferences were disappearing quickly.

I decided to initiate a living archive for DHNB and its community. I began collecting all DHNB related materials on the web, starting with the central area of its activity: the DHNB conferences and their papers in BoAs, Conference Proceedings, special issues of journals, and other publication outlets. I decided to use Zotero (https://www.zotero.org/) and added all publications about DHNB wherever I could find them.

However, it quickly became apparent that just saving links to websites wasn't enough since they rotted rapidly. I, therefore, decided to preserve all links - to blogs, newsletters, news items on university websites, online magazines, and other platforms - using the Internet Archive's Wayback Machine (https://archive.org/web/) to save at least one time-stamped version of their content for the future.

I manually combed Twitter for DHN conference-related content and all the official DHN Twitter account tweets. After only a few weeks of research, what grew in the Zotero library became a veritable living bibliography of DHN and its community.

I continued the work in 2020, adding the outputs of DHN2020 pre-and post-conference and preparing the bibliography to become an open community resource when I transferred my private Zotero collection into a public Zotero group.

For the future, this living bibliography has the following goals:
- iteratively update the bibliography with all DHNB abstracts from BoAs, all articles published in the conference proceedings, and special issues of journals
- iteratively update with news items, newsletters, blog posts, and other smaller pieces reporting about DHNB, including conference reports, workshop reports, reviews, opinion pieces
- categorising the DHNB publications using the TaDiRAH taxonomy (https://tadirah.info/) by creating and assigning tags in the Zotero group library
- supplying full-text versions of all abstracts, papers, and posters via the Zotero group library
- preserving all links to websites containing DHNB-related material to the Internet Archive or similar services
- systematically harvesting Twitter for DHNB-related hashtags and archiving tweets as a version-controlled data package on Dataverse.no (https://dataverse.no/)
- systematically harvesting Facebook and Instagram for DHNB-related hashtags and posts
- establishing a workflow for feeding DHNB conference data into the Index of Digital Humanities Conferences (https://dh-abstracts.library.cmu.edu/)

This endeavour cannot stay the side project of one individual. I call the DHNB community and its friends to help collect relevant material in the Zotero group library and an open GitHub repository.
I aim to make the collection of DHNB publications and data openly available for the community to use, re-use, share, and build on.

Anne Grethe Sæbø

Serving as a Method's Partner to the Researcher in Complex Digital Times: A vision for a new era for the University of Oslo Library

Keywords *library digitization hub, digital special collections, digital scholarship center, research method partner*

Contribution poster

Affiliation University of Oslo Library, Norway

Abstract In 2019, to meet the support needs of our researchers who are navigating a digital turn with open science demands and data protection requirements, the University of Oslo Library launched a Digital Scholarship Center (DSC) as a two-year project. The center has become a trusted advisory service for researchers working on their Research Data Management Plans in particular. The center organizes courses, workshops, and events related to open access and FAIR, digital scholarship and digital skill development, including Carpentries and CodeRefinery.

Then in 2021, the University of Oslo Library centralized its teams working with the physical and digital collections and the digital services to better meet the requirements of open access and open research with more complex demands on systems and integrations. The Humanities and Social Sciences Library (HUMSAM Library) in turn reorganized with a new section devoted to research support, including a team specializing in digital research methods and another focused on special collections and digitization. Both these teams collaborate with the DSC and strive to support our researchers with the management, collaboration, publication, and storage of their research data. At the end of its project period, it was determined that the DSC would be integrated into the new centralized division for Collections and Digital Services.

In 2022-2023, we will also open a digitization hub with scanners and OCR & TCR software. We have received funding to find digital solutions for a select pilot group of archival and special collection projects that we will digitize and publish on a digital platform. We are trying out the Swedish platform Alvin for these

pilots, and may commit to a long-term partnership with them for the publication and cataloging of more from our special collections. And the DSC will open a physical space for our researchers where they can access some of our specialized equipment, including scanners and powerful machines (PCs and/or MACs).

While these initiatives are promising for our goal to meet the digital needs of our researchers as their method's partner, we have encountered several challenges that still remain to be resolved that pertain to technical, financial, and administrative interests and circumstances. In my presentation, I will address these as well as the possibilities we see in how to resolve some of them.

Kirsi Sandberg[1], Mykola Andrushchenko[1], Risto Turunen[1], Jani Marjanen[2], Jussi Kurunmäki[1], Jaakko Peltonen[1], Timo Nummenmaa[1], Jyrki Nummenmaa[1]

Analyzing temporalities in parliamentary speech about ideologies using dependency parsed data

Keywords *temporality, parliamentary records, natural language processing, universal dependencies*

Contribution short paper

Affiliation 1: Tampere University, Finland;
2: University of Helsinki, Finland

Abstract The temporal aspects of politics have been discussed extensively by political theorists, but have not been explored using grammatically parsed textual datasets. This paper explores the ways in which future, present and past are projected and referred to in speeches in the Finnish parliament that talk about ideologies. Ideologies are crucial categories of thinking about the political past and future and therefore serve as a case in which temporality is expressed in a variety of ways. We use a dataset drawn from Finnish parliamentary records from 1980 to 2021 and operationalize morpho-syntactic information onclause structures and grammatical tense system to explore the different temporal profiles of ideologies. We show how some isms, like communism and fascism, are much more likely to appear in the context of the past, whereas others, like capitalism and racism, tend to appear in the present tense. We further develop a framework for analyzing temporality based on clause structures and grammatical tense and relate that to how the study of politics has approached time in parliamentary speaking.

Thomas Schmidt, Sabrina Hartl, Konstantin Kulik, Vera Wittmann

Systematic Evaluation of Annotation Tools and Analysis of Annotation Behavior for Three Annotation Tasks

Keywords *Annotation, Usability, User Experience, Tools, Usability Engineering*

Contribution short paper

Affiliation Media Informatics Group, University of Regensburg, Germany

Abstract Annotation is one of the most important activities in Digital Humanities (DH) and is used for texts to enrich them with metadata, linguistic or semantic information. Various annotation tools, desktop and web-based have been developed that can be used by researchers for their projects. However, while there are evaluations for certain tools and certain tasks, systematic evaluations of various annotation tools for multiple tasks are rare. Furthermore, we argue that the usability of tools not just influences the satisfaction of annotators but also the annotation behavior meaning annotation distributions and correctness. Therefore, next to the systematic evaluation of tools across different tasks, we also tested this statement by looking at correlations of usability/user experience metrics with annotation distributions and the correctness of annotations.

We focus on accessible and ready-to-use annotation tools and evaluate the three established text annotation tools: CATMA, eMargin, WebAnno. As tasks we investigate the following three standard tasks in DH: (1) Coreference annotation (annotation of mentions of characters), (2) part-of-speech annotation for nouns, verbs and adjectives and (3) speech type annotation (direct vs indirect speeches). For each task we selected different appropriate texts of the literary domain and created a correct gold annotation. The study was performed via a within-subject design with 12 participants with no prior experience with the annotation tasks and tools. Each participant performed every task with a different tool marking the specific text. The order of task and the distribution of tools was counterbalanced, and each participant was introduced shortly into the task and the tool. After each task, annotators filled out several questionnaires to measure subjective usability and user experience: the System Usability Scale (SUS), the User Experience Questionnaire (UEQ) and the NASA Task Load Index (NASA-TLX). As objective metrics, we also measured the task completion time and the number of correct annotations. To examine the general annotation behavior, we also looked at the overall number of annotations.

The statistical analysis of our results shows that CATMA achieves the best results considering the SUS, UEQ and NASA-TLX score, and the difference compared to the other tools is significant; thus, this tool is perceived as having a better usability and user experience. The result for the task completion time is the same: annotators completed the tasks faster with CATMA, however the

correctness only shows small and descriptive differences and is quite similarly high across tools. We did not find task-specific differences.

Considering the question regarding the relationship between usability/user experience metrics and annotation behavior we were not able to find significant results. We did however find descriptive relationships for the NASA-TLX metric, which correlates positively with the number of correct annotations meaning the more difficult the task was perceived the less mistakes were made. Nevertheless, we could not fully validate our assumption of this relationship. We argue that the tools and the tasks were too similar to show significant relationships between usability and annotation behavior. We intend to discuss our results in more detail in the presentation and report ideas for future work.

Maria Skeppstedt, Rickard Domeij, Gunnar Eriksson, Jenny Öqvist

Digital humanities for the spreadsheet nerd: Presenting the output of a topic modelling tool as tabular data

Keywords *topic modelling, text mining, audio recording descriptions, dialects*

Contribution Poster

Affiliation Institute for Language and Folklore, Sweden

Abstract The text mining algorithm 'topic modelling' can be used to automatically extract frequently re-occurring topics from a large text collection. There are a number of digital humanities tools for visualising the output of the algorithm, e.g., tools which visualise words representing the topics extracted, and which support a manual analysis of text snippets representative to the topics. Given the general popularity of the standard spreadsheet, we suggest an additional way to present the output of a topic modelling-based text analysis: as simple tabular data. We therefore (i) applied the topic modelling tool Topics2Themes to our text collection and used its visual user interface for an initial analysis of the topic modelling output, (ii) devised and implemented a tabular format for exporting the automatic and manual analysis performed using the visual tool, and (iii) imported the tabular data exported from the tool into a spreadsheet program for further exploration.

The work was carried out within the project ’Tilltal’ ("Accessible cultural heritage for speech research"), which is aimed at making archival speech recordings more accessible to research in the humanities and social sciences. As the text collection on which to apply the tool, we used text descriptions of audio recordings, i.e., textual entry points to the recorded speech. With a

future goal to study audio recordings that discuss different aspects of language, we selected recording descriptions that had previously been manually coded as containing content related to dialects and languages. Since each recording description often contains several subjects, we divided the descriptions into subtexts by splitting them on time indications, which typically also indicate a change of subject. This resulted in a total of 2,378 documents, to which we applied the Topics2Themes tool.

The tool automatically detected 16 topics, of which we assessed 8 to contain language related content. For these 8 topics, we used Topics2Themes to manually analyse the 25 most closely associated documents, which resulted in that 21 re-occurring language-related themes were identified. This automatic and manual analysis was then exported from the tool and imported into a spreadsheet program. The tabular format devised allowed for a number of possibilities for sorting and exploring the texts, albeit not as many options as provided by the Topics2Themes interface. But there are also benefits of the tabular format compared to the Topics2Themes interface. The structure of the spreadsheet is more readily comprehensible to researchers not involved in the work of devising its format, and the spreadsheet structure offers a greater flexibility with regards to adding your own data categories and to disseminating the output of the topic modelling-based analysis. We therefore believe the tabular data-export functionality to be a useful addition to the graphical user interface of Topics2Themes. As future work, we plan to provide the user with more flexibility regarding which content to export from the Topics2Themes tool. We also plan to implement functionality for importing content from the spreadsheet-based analysis back into Topics2Themes.

Topics2Themes is available at: github.com/mariask2/topics2themes. Word lists, configuration files and output examples are available at: github.com/mariask2/spreadsheetnerd.

Bo Ærenlund Sørensen[1], Lars Kjær[2]

Looking for dangerous liquids in Chinese literature: a programmatic approach

Keywords *NLP, python, ontology, liquids, personhood*

Contribution short paper

Affiliation 1: University of Copenhagen, Denmark;
2: Royal Danish Library

Abstract Many scholars have noted that liquids and gasses have been vested with particular significance in Chinese social life for centuries. In terms of medical

theory and practice, the damaging and healing powers of liquids and gasses have been a central concern at least since the 2nd century BCE (Kuriyama 2002) and continues to be so today (Hsu 2007; Shapiro 1998). In Chinese religious practice, offerings to ancestors are often burnt or dissolved in water—in other words, turned into liquids and gasses—so that they may traverse the boundary to the domain beyond (Mueggler 2001). Water also features prominently in Chinese imaginings of the underworld that is often referred to as the "Yellow Springs" (Lewis 2006). Obviously, the tradition of fengshui—a term meaning "wind water"—also focuses attention on the flows of liquids and energies (Bruun 2008). Scholars of Chinese literature have noted that, at least in the case of 17th century novels, liquids feature with conspicuous frequency in relation to illicit sexual desires (Epstein 1999).

This paper uses python programming to examine whether a similar erotic investment appear in the form of liquids in contemporary Chinese literature. Working with a midsize corpus of novels and short stories, python is used to locate passages where main protagonists meet for the first time. Python extracts these passages which are then read by a human being who codes the passages by paying attention to the occurrence of liquids. Terms identified to be of central significance are then employ for further computer-aided investigations of the original corpus. In short, do liquids appear with remarkable frequency and in conspicuous roles in the first meetings of characters who will later develop romantic relations? The findings are discussed in relation to comparative literature and theory of mind.

Bibliography

Bruun, Ole. 2008. An Introduction to Feng Shui. Cambridge: Cambridge University Press.

Epstein, Maram. 1999. "Inscribing the Essentials: Culture and the Body in Ming-Qing Fiction." Ming Studies 1999(1): 6–36.

Hsu, Elisabeth. 2007. "The Experience of Wind in Early and Medieval Chinese Medicine." Journal of the Royal Anthropological Institute 13: 117–34.

Kuriyama, Shigehisa. 2002. The Expressiveness of the Body and the Divergence of Greek and Chinese Medicine. New York: Zone Books.

Lewis, Mark Edward. 2006. The Flood Myths of Early China. Albany, NY: SUNY Press.

Mueggler, Erik. 2001. The Age of Wild Ghosts: Memory, Violence, and Place in Southwest China. Univ of California Press.

Shapiro, Hugh. 1998. "The Puzzle of Spermatorrhea in Republican China." positions 6(3): 551–95.

Polina Staroverova[1], Natalia Perkova[2]

Towards improving the OCR quality of 18th century Russian pre-reform books and periodicals

Keywords *optical character recognition, digitalization, digital collections*

Contribution Poster

Affiliation 1: Higher School of Economics, Russia;
2: Uppsala University, Sweden

Abstract The revolutionary reform of the Russian alphabet and script in the beginning of the 18th century, known as the introduction of the so-called civil script or typeface (graždanskij šrift), accelerated the development of book production: new typographies were opened, the first periodicals were published, and the amount of printed production increased enormously. The importance of the Russian printed heritage of the 18th century is immense from the philological perspective, but it also expands to numerous areas of science and humanities. However, the accessibility of these sources leaves much to be desired. Even though digitalization by major Russian libraries has progressed considerably in the past years, most historical books and periodicals are still often not truly searchable and are typically provided either as downloadable PDF files without OCR or put into library-specific viewer mode platforms. Of course, works of most prominent 18th century Contribution have been published extensively and made available digitally with time; however, they are typically rendered in modern orthography, and this mainly concerns poetry and prose. As for periodicals and scientific literature, there is still much work to be done on making them fully accessible for wider audience.

In our project, we aim at the creation of ground truth data representing various typographies (and thereby various fonts used in the 18th century) to improve the OCR quality of of Russian pre-reform scanned books and periodicals. The best known Russian old orthography OCR model is available within ABBYY FineReader and has been used, among others, by a number of national libraries, but this is proprietary software, and it seems to misclassify some specific features of 18th century fonts. On the other hand, open-source libraries Tesseract and Ocropy lack special models for pre-reform Russian. We have started our work in Transkribus by building ground truth data for several books printed by the typography of the Russian Academy of Sciences, to be later expanded by other typographies. This typography was highly infuential in terms of the development of the Russian civic script (Šicgal 1985). The choice of the books was determined by the existence of proofread transcriptions which had been prepared for the purposes of the Russian National Corpus as texts preserved in old orthography. This allowed us to faster and easier match lines to their transcriptions, as it required only minor corrections.

We managed to train the first pilot model on a small dataset of 6459 words and 1154 lines (43 pages from 7 books) with a close to zero CER on validation set. This is, of course, not unexpectable due to the print character of the documents and the small size of the dataset, but we tested the model on some pages printed by other 18th c. typographies, and the quality still remains very satisfying. We also tested the OCR quality of one random page outside the training/validation data, but from the same typography, using our model, Google Cloud Vision and Tesseract, with expectably higher quality obtained by our custom model.

Emil Stjernholm

Distant Reading Televised Information: Exploring the Communication of Swedish Government Agencies, 1978–2020

Keywords *television studies, content analysis, corpus analysis, automatic speech recognition, government information*

Contribution long paper

Authors Stjernholm, Emil
Lund University, Sweden

Abstract Starting in the early 1970s, the program Anslagstavlan aired government information twice a week on Swedish public service television (SVT). There were only two channels at the time, hence this pinboard of government information on taxes, health care and public insurance reached millions of Swedes. Though often part of larger information campaigns, the televised government information stood out, making the program well-known amongst generations of Swedish audiences.

Anslagstavlan is a neglected but important part of Sweden's modern audiovisual heritage. While much research has been devoted to the Swedish public service model, wherein regulated independence from the government has been a cornerstone, little is known about Swedish television's function as a communication tool for government authorities. Whereas digital research methodologies for data-driven text analysis have been developed and established over the last decades, the use of digital tools in the analysis of audiovisual sources has only recently gained increased attention. To analyze the aesthetic, narrative and rhetorical development of Anslagstavlan, the research project Televising Information (Swedish Research Council, 2020–2023) relies on digital video analytics tools as well as speech-to-text-algorithms—foremost in co-operation with developers at the KB-lab research environment at the National Library of Sweden where all TV-programs are preserved.

For this paper, KB-lab's automatic speech recognition method ('Fine-tuning

XLSR-Wav2Vec2') was used to transform a sample of spoken messages in Anslagstavlan between 1978–2000 to text. Following this, the corpus analysis tool AntConc was utilized to explore keywords and collocational patterns in the material. By comparing and contrasting key themes and discourses in different time periods, this paper highlights how Swedish government agencies' communication has developed over time. Notably, even though the National Library of Sweden's audiovisual collection is frequently used in media historical scholarship, no previous research project has utilized digital methods to analyze material from their collection. Therefore, this paper additionally offers a reflection on the theoretical and methodological challenges facing the project at large, and in particular the novel approach to television content analysis using speech-to-text algorithms.

Bibliography Tenser, Anton. 2008. The Northeastern Group of Romani Dialects. Ph.D. dissertation, University of Manchester

Jana Sverdljuk[1], Lars Johnsen[2], Magnus Birkenes[2]

Merging Digital Humanities and Discourse Analysis in the Study of COVID-19 Vaccine Distribution in Norwegian Newspapers

Keywords *collocation, concordance, discourse analysis, newspapers, vaccine distribution*

Contribution short paper

Affiliation 1: University of Agder, Norway;
2: National Library of Norway

Abstract While anti-vaccine, anti-science attitudes are among the top 10 health threats facing the world, according to WHO, in Norway, there is a general high trust towards the Corona vaccine. Most of the Norwegian population has been vaccinated. At the same time, in Norway, extensive debates about the distribution of vaccines contributed to the politicization of this issue. National and regional newspapers were important channels, which mediated the scientific information on distribution forming public opinion on the authorities. By using computer-assisted methods in combination with discourse analysis, the article analyzes linguistic patterns and key topoi which were characteristic for the studied discourse. It shows the preoccupation with the number of doses, which were to be sent to the regions and reveals the topos of "the rule of ordinary people". This topos was interpreted as treating everyone equally, including the population living in the regions and those who belonged to the lower middle class. At the same time, "ordinary people" implied Norwegians as opposed to the rest of the world. Newspapers coverage of vaccine distribution played along with the rhetoric of the Labor and the Center parties, which were the main favorites in the coming Parliamentary elections.

Bibliography https://nbviewer.jupyter.org/github/DH-LAB-NB/DHLAB/blob/master/DHLAB_ved_Nasjonalbiblioteket.ipynb

Birkenes, Magnus Breder, Lars G. Johnsen, Arne M. Lindstad og Johanne Ostad (2015): «From digital library to n-grams: NB N-gram». In Proceedings of the 20th Nordic Conference of Computational Linguistics, 293–295. Linköping: Linköping University Electronic Press.

Minna Tamper[1,2], Jouni Tuominen[1,2,3], Eero Hyvönen[1,2]

Extending the Finnish Linked Data Infrastructure with Natural Language Processing Services in FIN-CLARIAH

Keywords *NLP services, NLP, data tranformation and enriching, digital humanities, Finnish*

Contribution poster

Affiliation 1: Aalto University, Semantic Computing Research Group (SeCo), Finland;
2: University of Helsinki, HELDIG – Helsinki Centre for Digital Humanities, Finland;
3: University of Helsinki, HSSH – Helsinki Institute for Social Sciences and Humanities, Finland

Abstract The EU-DARIAH infrastructure[1] for Digital Humanities (DH) is often focusing on using structured data for quantitative studies, while the EU-CLARIN infrastructure[2] deals primarily with unstructured natural language texts. However, in DH research based on both texts and structured data are needed. Therefore, it often makes sense to develop and use both infrastructures together, as suggested in the Dutch CLARIAH programme[3] and the corresponding FIN-CLARIAH initiative[4] in Finland, a part of the new research infrastructure roadmap of the Academy of Finland. This paper presents ongoing work in FIN-CLARIAH on integrating natural language processing (NLP) tools with the Linked Open Data (LOD) Infrastructure for Digital Humanities in Finland (LODI4DH)[5] (Hyvönen, 2020) using the existing Linked Data Finland platform[6]. We present a set of NLP services, to be opened to public use, that are used on the other hand for knowledge extraction from (Finnish and Swedish) texts (named entity recognition, linking, relation extraction, and data anonymization) for weaving LOD, and on the other hand in linguistic DH data-analyses of the published datasets, such as national biography collections[7], Finnish legislation and case law[8], and the speeches 1907–2021 of the Parliament of Finland[9] available through the Sampo series of LOD services and portals[10].

The NLP services are byproducts of projects that have cleaned, transformed, and enriched LOD datasets. The services have been largely created from needs stemming from input data and project goals. The input in PDF, CSV, HTML, and text form is converted into RDF. The data can simultaneously be cleaned from, e.g., OCR errors or broken HTML tags. Our tools use various knowledge extraction methods for enriching the documents semantically. The results have been applied to create several in-use new applications on the web. Selected tools are available for open use at GitHub.

A portal for NLP services is underway resembling other NLP portals, such as the GATE Cloud (Tablan, 2013), but with a focus on providing tools to process and enrich Finnish texts. The portal provides users with demo applications and APIs, unified output formats (e.g., JSON, RDF), documentation, and software delivery in Docker containers, which lowers the bar for deployment. The tools can be tested with custom input through APIs directly or through online demo applications.

Notes

1 https://www.dariah.eu
2 https://www.clarin.eu
3 https://www.clariah.nl
4 https://www.kielipankki.fi/Authors /fin-clariah/
5 https://seco.cs.aalto.fi/projects/lodi4dh/
6 https://ldf.fi
7 https://seco.cs.aalto.fi/projects/biografiasampo/
8 https://seco.cs.aalto.fi/projects/lakisampo/
9 https://seco.cs.aalto.fi/projects/semparl/
10 https://seco.cs.aalto.fi/applications/sampo/

Bibliography

Hyvönen, E. (2020). Linked Open Data Infrastructure for Digital Humanities in Finland. In Reinsone, S., Skadiņa, I., Baklāne, A., & Daugavietis, J. (Eds.) Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, pp. 254–259, CEUR Workshop Proceedings, vol. 2612. http://ceur-ws.org/Vol-2612/short10.pdf

Tablan, V., Roberts, I., Cunningham, H., & Bontcheva, K. (2013). GATECloud.net: a platform for large-scale, open-source text processing on the cloud. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1983), 20120071. doi:10.1098/rsta.2012.0071

Iiro Lassi Ilmari Tiihonen, Yann Ryan, Lidia Pivovarova, Aatu Liimatta, Tanja Säily, Mikko Tolonen

Distinguishing Discourses: a Data-Driven Analysis of Works and Publishing Networks of the Scottish Enlightenment

Keywords *computational history, eighteenth-century studies, economic discourse, social network analysis*

Contribution short paper

Affiliation University of Helsinki, Finland

Abstract A key feature of the Enlightenment is the development of a discourse on commerce and economy entangled in larger discussions around politics, morality, and progress. However, this discourse was not formalised in the way it may be seen today. Instead, it was an emerging subject incorporating diverse and contested ideas. The objective of this case study, then, is to use various methods to identify the boundaries of these emerging economic, political, and moral discourses in a data driven way, using a unified version of the metadata (e.g. publisher, publication year, format of the print product) of the English Short Catalogue (ESTC) and full texts of the Eighteenth Century Collections Online (ECCO). We approach the task iteratively, first making a separation between broadly defined economic documents and other eighteenth century documents by modeling the features which separate samples from two collections of historical economic texts from the wider ECCO data. Then, based on the previous step, we distinguish works similar to Hume's \textit{Political Discourses} (a text widely seen as the epitome of the Scottish Enlightenment) from other branches of commercial and economic discourse, and analyse this set of works in more detail. We also experiment on how a purely unsupervised approach -- a contextualized topic model using BERT encodings -- groups our set of economic texts.

Previous historical scholarship has taken the perspective that we ought to identify language use from large corpora of text. The aim has been to contextually understand language from the perspective of a particular group of historical actors or, as is the case with conceptual historians, detect contested and changing concepts. Our approach is different in that we closely link language use to the material and historical circumstances of the individual texts within which these uses can be found. Essentially, we combine computation and social network analysis with the study of changing concepts and word uses over time, taking individual editions rather than language abstracted from them as the basic object of study. This approach allows us to identify additional contextually relevant works by both their linguistic features, and the material and network history of their production.

Jointly, the combination of iterative data-driven discourse detection and the

focus on manifested editions allows us not only to extract a significant proportion of the Scottish Enlightenment discourse in a data driven manner, but to link it to the social networks and commercial context in which it was produced and evolved. This approach allows us to evaluate the existence, scope, and social and economic contexts of historical discourses (in particular, economic discourse) in the eighteenth century in a way which is both computationally state-of-the-art and relevant to historical practices and interests. It also demonstrates how data-driven analysis and the traditional hermeneutic approach of the humanities can be combined with an iterative approach to study meanings and their changes over time.

Pihla Toivanen

A workflow for selecting the research material: Using BERT and active learning in catching the phenomenon

Keywords *text classification, active learning*

Contribution short paper

Affiliation University of Helsinki

Abstract Computational methods have been increasingly used in media studies for both collecting and analyzing data. A less studied area is how to select the relevant research material from an enormous amount of data when it's not possible to take a qualitative look at every document. A common strategy for finding relevant documents in media studies is to formulate keyword-based queries for searching documents (e.g. Rogers, 2019). However, the keyword strategy doesn't work properly when the phenomenon is conceptually too wide and abstract to be represented with only a few keywords. Using keywords as the only selection criterion also encourages studying phenomena that are easily definable with those few words: for example, events, persons and places. To address this problem, in this paper, we present a BERT-based workflow for selecting relevant research material from big data when studying a particular debate that can't easily be defined with a keyword-based approach.

The workflow is based on first identifying a manual classification scheme for relevant articles and then building a model to detect the articles automatically. We use machine learning, which allows reasoning the language characteristics of the phenomenon inductively without specifying keywords. As the model, we use BERT (Bidirectional Encoder Representations from Transformers) classifier (Devlin et al., 2018) with an active learning training strategy.

Our workflow includes the following steps:
A media researcher identifies the phenomenon in interest and develops a classification scheme for selecting the example articles
Initial training and validation datasets are classified into categories manually according to the classification scheme
The training dataset is expanded through the active learning technique: the model suggests new articles to be classified manually based on uncertainty sampling
The resulting classifier is used in selecting the relevant research material from the whole dataset
We demonstrate the approach with a research case studying media power in four Finnish newspapers, specifically finding articles concerning alcohol policy. We use Finnish BERT (FinBERT) (Virtanen et al., 2019) to identify documents where the research interest is present.

Bibliography

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.

Rogers, R. (2019). Doing Digital Methods. SAGE

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. ArXiv, abs/1912.07076.

Jouni Tuominen[1,2,3], Mikko Koho[1,2], Ilona Pikkanen[4], Senka Drobac[1], Johanna Enqvist[4,5], Eero Hyvönen[1,2], Matti La Mela[2,6], Petri Leskinen[1], Hanna-Leena Paloposki[4,7], Heikki Rantala[1]

Constellations of Correspondence: a Linked Data Service and Portal for Studying Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland

Keywords *Epistolary culture, letter metadata, Linked Open Data, semantic portal, data analysis and visualisation*

Contribution short paper

Affiliation 1: Semantic Computing Research Group (SeCo), Aalto University, Finland;
2: HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland;
3: HSSH – Helsinki Institute for Social Sciences and Humanities, University of Helsinki, Finland;
4: Finnish Literature Society, Finland;
5: Cultural heritage studies, University of Helsinki, Finland;

6: Uppsala University, Sweden;
7: Finnish National Gallery, Finland

Abstract

This paper presents the vision of aggregating, harmonizing, and publishing letter catalogue metadata from cultural heritage (CH) institutions in Finland as a single reconciled Linked Open Data (LOD) service and a semantic portal providing data analytical tools for researchers. The Constellations of Correspondence (CoCo) project will enable scholars to conduct empirical, bottom-up case studies on epistolary culture and social networks in the Grand Duchy of Finland (1809–1917), ask ambitious research questions in the field of computer science and make currently scattered, heterogeneous epistolary metadata interoperable and available.

The project builds upon the experiences accumulated during the Reassembling the Republic of Letters (RRL) [1], 1500–1800 (2014–2018) COST action and works in close collaboration with the LetterSampo initiative (https://seco.cs.aalto.fi/projects/rrl/) [2,3] that is building a framework for representing, publishing, and using epistolary data as LOD on the Semantic Web for Digital Humanities research. Compared with RRL, CoCo is focused on data that are both temporarily and geographically more restricted but, due to this, much more comprehensive regarding social layers of 19th-century epistolary culture.

The project is currently surveying and collecting data from letter metadata collections scattered in different archives in Finland. Based on the analysis of the existing data and state-of-the-art data models, we will create a harmonizing data model for epistolary data. The collected datasets will be transformed into the data model accompanied with automatic and manually curated disambiguation processes for the reconciliation of the identities of pivotal entities (people, places). The recognized people are linked to established LOD registers and enriched with their metadata, such as occupation, gender, and other social connections. For datasets with full-text contents of letters, we investigate options for enriching the metadata by utilizing natural language processing methods, such as named entity recognition.

Based on the integrated letter metadata, the project develops and utilizes social network analysis, data visualization, and knowledge discovery methods to respond to the empirical research questions presented in the project and to discover interesting patterns and phenomena in the data. The analysis tools are packaged and provided for public and scholarly use as a LOD service, SPARQL endpoint, and semantic portal.

The multidisciplinary CoCo project consortium, funded by the Academy of Finland, consists of historians, heritage specialists, and computer scientists at the Finnish Literature Society, Aalto University, and the University of Helsinki.

Bibliography [1] Hotson, H., & Wallnig, T. (Eds.). (2019). Reassembling the Republic of Letters in the Digital Age: Standards, Systems, Scholarship. Göttingen University Press.

[2] Tuominen, J., Mäkelä, E., Hyvönen, E., Bosse, A., Lewis, M., & Hotson, H. (2018). Reassembling the Republic of Letters – A Linked Data Approach. In E. Mäkelä, M. Tolonen, & J. Tuominen (Eds.), DHN 2018: Digital Humanities in the Nordic Countries 3rd Conference (pp. 76–88). CEUR Workshop Proceedings, vol. 2084. http://www.ceur-ws.org/Vol-2084/paper6.pdf

[3] Hyvönen, E., Leskinen, P., & Tuominen, J. (2021). LetterSampo – Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data of the Republic of Letters. Under review.

Risto Juhani Turunen, Ilari Taskinen, Lauri Uusitalo, Ville Kivimäki

Mining Emotions from the Finnish War Letter Collection, 1939-1944

Keywords *sentiment analysis, history of emotions, digital history, war letters*

Contribution long paper

Affiliation Tampere University, Finland

Abstract Our paper analyzes emotional language used by Finnish soldiers and civilians in their private communication during World War II. The dataset consists of 7,000 handwritten letters that have been transformed into a machine-readable corpus with rich metadata. The dataset offers a unique opportunity to analyze statistically people's emotional responses to the war. We engage in key questions of cultural history of war such as the connection between soldiers' emotional language and violence: did initial optimism shift towards an unemotional way of seeing the world during the war?

While computational approaches to mine emotions have been common in fields like computer science and linguistics, they have not gained wider popularity in historical research. The pioneering attempts have been based on individual emotion words carefully chosen by an historian, or on more generic emotion lexicons that have been easily available. Compared to machine-learning solutions, lexicon-based approaches demand less computational effort and are more transparent to interpret. Our methodology combines the ready-made word list FEIL with contextual knowledge of historians. FEIL gives around 7,000 Finnish words an emotion category and an intensity score. First, the emotion lexicon was filtered based on high intensity. Then, the domain expert manually cleaned words that were not emotionally especially intensive in the context of war letters. The expert also annotated the list of the most frequent words in the war letter collection and handpicked emotionally intensive words

not included in the FEIL. Our final list covered 298 emotion words. We quantified how their use changed over time.

Based on our analysis, soldiers' and civilians' emotionality does not appear to be very different in letters during World War II. Soldiers' use of emotion words saw a decline in the last moments of the war, but overall their letters were almost as emotional as the civilians' letters. We could identify some changes in the individual emotion words used by the soldiers in their letters: especially patriotic words decreased in the course of the war. In addition to empirical findings, our paper sheds light on the problem of universal emotion lexicons in historical research: linguistic, cultural and temporal differences between present-day lexicons and historical datasets can lead into biased interpretations. Thus, our paper contributes not only to the history of emotions but also to emotion mining that is historically sensitive.

Raf Van Rooy[1], Xander Feys[2], Maxime Maleux[2], Andy Peetermans[2]

The rocky road to DaLeT: Pitfalls and successes in developing a database of the Trilingual College of Leuven (1517–1578)

Keywords *database creation and management, digital edition with metadata, student notes, Trilingual College Leuven, 16th-century history*

Contribution long paper

Affiliation 1: University of Oslo, Norway;
2: KU Leuven

Abstract In this paper, we will present our Trilingue e-resource, which we have baptized DaLeT, Database of the Leuven Trilingue, and already online in a beta-version at https://www.dalet.be/ but still under construction. It is the first database devoted to an early modern trilingual institute, the Collegium Trilingue Lovaniense, or the Three-Language College in Leuven, where one could study the three so-called sacred languages Latin, Greek, and Hebrew for free. The Trilingue was also known as the Collegium Buslidianum, or Busleyden College, named after its material founder Jerome of Busleyden (ca. 1470-1517), a prominent diplomat with humanist interests and many important connections throughout Habsburg Europe. The brainchild of Desiderius Erasmus, the institute existed from 1517 until 1797, when it was dissolved in the wake of the French revolution and the annexation of most parts of present-day Belgium by France. The Trilingue was most influential during its acme in the 16th century, especially the years 1517-1578, when it educated numerous prominent scholars, scientists, and politicians. DaLeT focuses on this early period of the college, after which Leuven intellectual life became seriously upset by the Eighty Years' War and other dire circumstances such as the plague.

DaLeT offers first and foremost a new and dynamic way for publishing student notes included in 16th-century prints, applied to three core Trilingue sources, one for each language taught, going beyond the transcription and translations of notes as happens, for instance, in ABO. This part is already available in our beta-version. Currently we are also working on integrating a section on books and people at the Trilingue during its acme, planned for release in late 2021 or early 2022. In a final step, presumably late 2022, we hope to construct a teacher's corner (in Dutch) for high school teachers of Classics in the Low Countries who want to initiate their students into daily life at the Trilingue, in and outside the classroom, and give them a taste of the very beginning of a multilingual classical tradition in this area.

In our paper, we will present the rocky road to DaLeT, the product of a cross-disciplinary collaboration between literary scholars, linguists, historians, and digital humanities specialists. We will discuss the pitfalls we faced – including the limitations of HTR in our case – and the successes we achieved – first and foremost the visualization of marginal handwritten notes in relation to the printed text on which they bear. We conclude by looking at future opportunities for digital pedagogy through DaLeT, an aspect of our database on which we have only started to brainstorm, and where feedback from an international community of digital humanities scholars is most welcome.

Joshua Wilbur

The usefulness of "small data" in capturing syntactic change in an under-documented endangered language

Keywords *language technology, syntax, endangered languages, Pite Saami*

Contribution short paper

Affiliatoin Tartu Ülikool, Estonia

Abstract This paper presents the results of a pilot study exploring the usefulness of "small data" for studying syntactic change over time in a small, highly endangered language. Most larger, well known, national and/or regional languages provide linguists plenty of "big data" to work with; such datasets allow even fine nuances in the patterns of linguistic structures to be identified while referencing a variety of factors. In contrast to this, most small, local languages or language variants are typically under-documented or have only a minimal set of available texts, especially when the language is endangered. This is particularly true if the language only has a limited history of being written. Pite Saami, as a critically endangered Uralic language (spoken in northern Sweden by currently approximately 35 speakers), clearly belongs to the latter

group. However, Pite Saami provides us with a somewhat unique opportunity because there are relatively old heritage texts as well as contemporary texts, all of which are essentially transcripts of spoken-mode language use; these form the two small subcorpora for this study. The older texts were collected in 1893 (by Ignácz Halász) and 1921 (by Eliel Lagercrantz, but published in 1963 and 1957); the collection of contemporary texts was collected over the last decade as part of the Pite Saami Documentation Project. Thus there is approximately a century of difference between the texts in these two subcorpora.

Thanks to language technology tools (specifically Finite State Transducer and Constraint Grammar), these texts have been automatically tokenized, lemmatized, and tagged for part of speech and morphological categories (cf., e.g., Gerstenberger et al. 2017), and stored in XML format (in the form of ELAN transcription files). In my pilot study, I investigate how to use computational methods to extract syntactic patterns, specifically clause-level constituent ordering, from this specific data set (comprising around 10,000 tokens from about 1000 clauses). Furthermore, I control for the age of the respective subcorpora, in order to determine if any significant differences between constituent ordering in the heritage texts vs. the contemporary texts exist. Aside from the potentially interesting linguistic results, this pilot study is explores to what extent computational results obtained from ”small data”, which is often the only data we have in some areas of the Humanities, can be meaningful.

Bibliography

Gerstenberger, Ciprian, Niko Partanen, Michael Rießler, & Joshua Wilbur (2017). “Instant annotations. Applying NLP methods to the annotation of spoken language documentation corpora”. In: International Workshop on Computational Linguistics for Uralic languages. St. Petersburg: Association for Computational Linguistics. 25–36.