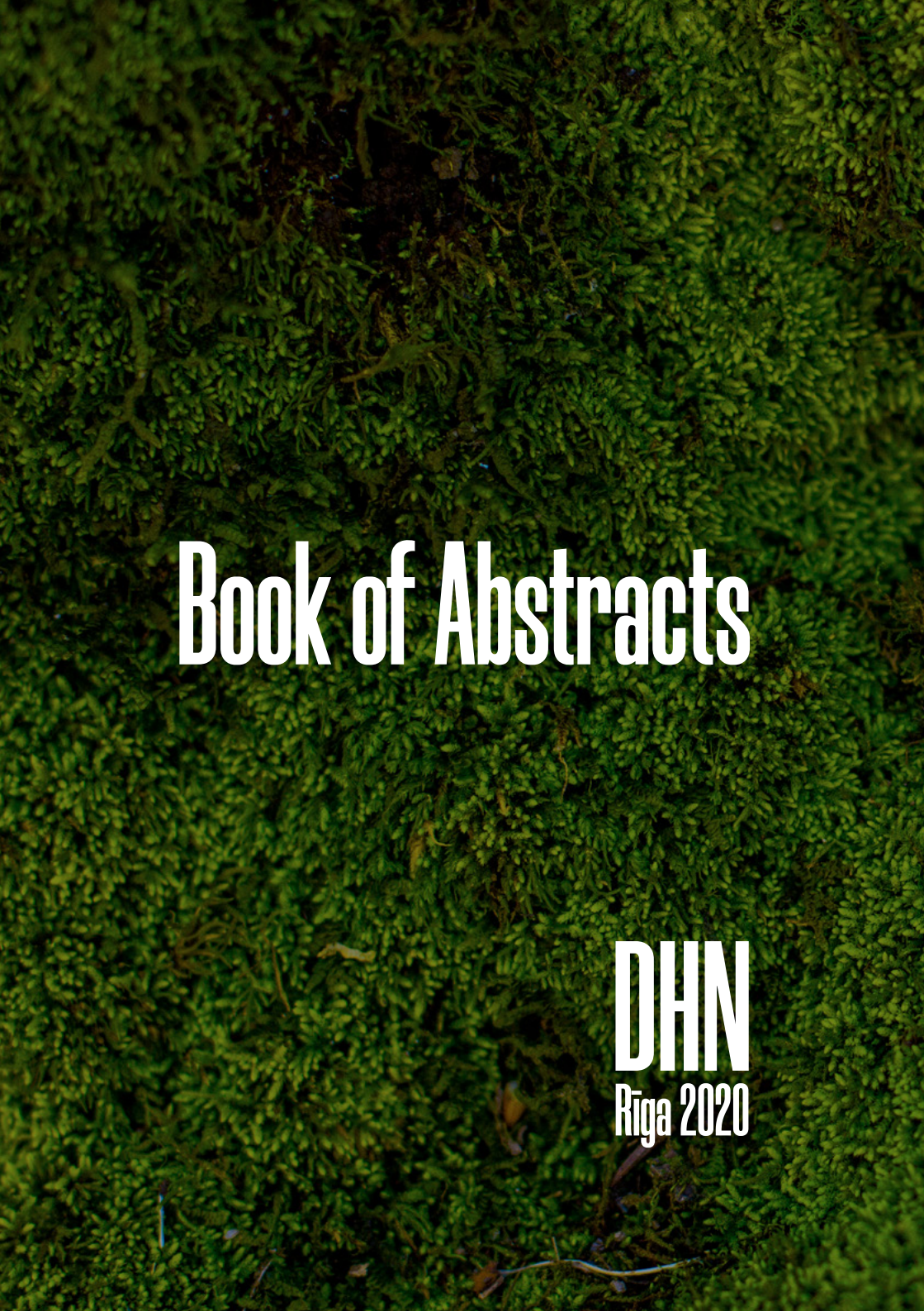


Book of Abstracts

DHN

Rīga 2020

ISBN 978-9984-850-83-2
DOI 10.5281/zenodo.4107117

Book of Abstracts of the Digital Humanities in the Nordic Countries 5th conference. Riga, 20–23 October 2020

Conference website:
http://dig-hum-nord.eu/conferences/dhn2020

Editors:
Sanita Reinsone, Anda Baklāne, Jānis Daugavietis

Editorial assistants:
Justīne Jaudzema, Ilze Ļaksa-Timinska

Cover: Anete Krūmiņa

Publisher:
Institute of Literature, Folklore and Art (University of Latvia)
lulfmi.lv

Riga, 2020

TABLE OF CONTENTS

ORGANIZERS 11

PROGRAMME COMMITTEE..... 12

SUPPORTERS 13

DHN2020 ABSTRACTS 14

Open a GLAM Lab 15
Aisha Al Abdulla, Sarah Ames, Paula Bray, Gustavo Candela, Sally Chambers, Caleb Derven, Milena Dobreva, Katrine Gasser, Stefan Karner, Kristy Kokegei, Ditte Laursen, Mahendra Mahey, Abigail Potter, Armin Straube, Sophie-Carolin Wagner, Lotte Wilms

New Empirical Resources for the Study of Contemporary Preaching – Presenting a Danish Sermon Corpus through Two Cases on Cultural Conformity and Representation of Christian Concepts 18
Anne Agersnap, Kirstine Helboe Johansen, Uffe Schjødt, Kristoffer Laigaard Nielbo, Ross Deans Kristensen-McLachlan

A Proposed Workflow for Future Monograph Digitization Projects 23
Niklas Kristian Alén

Short Texts in the Corpus of Early Written Latvian (www.korpuss.lv/senie) 27
Everita Andronova

Time-Layered Cultural Map of Australia 32
Paul Arthur, Erik Champion, Hugh Craig, Ning Gu, Mark Harvey, Victoria Haskins, Andrew May, Bill Pascoe, Alana Piper, Lyndall Ryan, Rosalind Smith, Deb Verhoeven

Emotional Imprints: Letter-Spacing in N.F.S. Grundtvig's Writings 33
Katrine F. Baunvig, Oliver S. Jarvis, Kristoffer Laigaard Nielbo

Linked Data for Digital Humanities Scholars and Researchers: "Rainis and Aspazija" Collection 34
Uldis Bojārs, Anita Rašmane, Anita Goldberga

Linked Open Data Vocabularies and Identifiers for Medieval Studies *38*
Toby Burrows, Antoine Brix, Doug Emery, Mitch Fraas, Eero Hyvönen, Esko Ikkala, Mikko Koho, David Lewis, Synnøve Myking, Kevin Page, Lynn Ransom, Emma Thomson, Jouni Tuominen, Hanno Wijsman, Pip Willcox

"Memes" as Activism in the Context of the US and Mexico Border *39*
Martin Camps

No Longer Obsolete: Mapping Digital Literacy Skills for Museum Professionals in Sweden and Lithuania *42*
Nadzeya Charapan

Comparing Word Frequencies and Lexical Diversity with the ZipfExplorer Tool *46*
Steven Coats

Digital Maps for Linguistic Diversity *47*
Coppélie Cocq

An Artificial Intelligence Approach to Segmenting Medieval Manuscripts with Complex Layouts *50*
Lisandra S. Costiner, Lizeth Gonzalez Carabarin

Óravíddir: Interactive Exhibition about the Icelandic Language ... *54*
Trausti Dagsson, Jón Hilmar Jónsson, Eva María Jónsdóttir

Classification of Medieval Documents: Determining the Issuer, Place of Issue, and Decade for Old Swedish Charters *55*
Mats Dahllöf

Supervised OCR Post-Correction of Historical Swedish Texts: What Role Does the OCR System Play? *57*
Dana Dannells, Simon Persson

Museums, Technology and Social Interaction in "Anyone Can Innovate!" *58*
Gabriella Di Feola, Erik Einebrant, Fredrik Trella

Responsible Artificial Intelligence *59*
Virginia Dignum

Taking the Livonians into the Digital Space *60*
Valts Ernštreits, Gunta Kļava

Inheriting Digital Projects: How to Keep Ibsen Alive Online *65*

Nina Marie Evensen

Human-Centered Humanities: Using Stimulus Material for Requirements Elicitation in the Design Process of a Digital Archive *66*
Tamás Fergencs, Dominika Illés, Olga Pilawka, Florian Meier

Birth Certificate Enslavement – A Conspiracy from the Archives to the Internet *67*
Rikard Lars Friberg von Sydow

Automatic Morphological Annotation of Ego-Documents: Evaluating Automatically Disambiguated Annotation of Estonian Semper-Barbarus Correspondence Corpus *71*
Olga Gerassimenko, Kadri Vider, Neeme Kahusk, Marin Laak, Kaarel Veskis

From Cow Sheds to Computer Screens: Some Thoughts on the Uses of Digital Humanities for the Study of Folkloristics in Iceland and the Other Nordic Countries *75*
Terry Gunnell

Supporting Research Use of WEB Archives: A 'Labs' Approach *77*
Olga Holownia, Sally Chambers

Evaluating a DH Tool. The First 18 Months of the Gale Digital Scholar Lab and the Future of Academic/Corporate Partnerships *81*
Christopher Michael Houghton

"Sampo" Model and Semantic Portals for Digital Humanities on the Semantic Web *84*
Eero Hyvönen

Linked Open Data Infrastructure for Digital Humanities in Finland *86*
Eero Hyvönen

Studying Transnational Digital Spaces: Methodological Vistas and Challenges *87*
Anastasia A. Ivanova

A Workflow for Integrating Close Reading and Automated Text Annotation *93*
Maciej Janicki, Eetu Mäkelä, Anu Koivunen, Antti Kanner, Auli Harju, Julius Hokkanen, Olli Seuri

Online Participatory Memory Work: Understanding the Potential Roles of Online Mnemonic Communities in Building the Collections of Public Memory Institutions *99*
Ina-Maria Jansson, Olle Sköld

Studying Semantic Domains in Akkadian Texts *105*
Heidi Jauhiainen, Krister Lindén, Saana Svärd, Tero Alstola, Aleksi Sahala

Legacy Data in a Digital Age *111*
Ellert Thor Johannsson, Simonetta Battista, Tarrin Wills

Text Mining Themes of the Urban Night in Historical Literary Corpora *116*
Hanne Emilia Juntunen

Emotion Preservation in Translation: Evaluating Datasets for Annotation Projection *121*
Kaisla Kajava, Emily Öhman, Hui Piao, Jörg Tiedemann

Tracing Complexity in Food Blogging Entries *122*
Maija Kāle, Ebenezer Agbozo

Modal Grammar and Metaphoricity as Vehicles of Affectivity in Political Newspaper Journalism *124*
Antti Kanner, Anu Koivunen, Eetu Mäkelä

Becoming a State Language: Finnish Public Debate and Modal Grammar 1820–1917 ..... *129*
Antti Kanner, Hege Roivainen, Tuuli Tahko, Jani Marjanen

Targeted, Neural Re-OCR of Norwegian Fraktur *135*
Andre Kåsen, Lars G. Johnsen

Digging Deeper into the Finnish Parliamentary Protocols – Using a Lexical Semantic Tagger for Studying Meaning Change of Everyman's Rights (Allemansrätten) *141*
Kimmo Kettunen, Matti La Mela

Can Umlauts Ruin Your Research in Digitized Newspaper Collections? A NewsEye Case Study on 'the Dark Sides of War' (1914–1918) *142*
Barbara Klaus

In Quest of Transition Books *143*
Denis Kotkov, Kati Launis, Mats Neovius

Digital Mapping: Research Method and Data Representation Tool in garamantas.lv *144*
Sandis Laime

Towards Better Structured Online Data with the Project "News, Opinions or Something Else? Modeling Text Varieties in the Multilingual Internet" *147*
Veronika Laippala, Saara Hellström, Sampo Pyysalo, Liina Repo, Samuel Rönnqvist, Anna Salmela, Valtteri Skantsi

LIBDAT: Towards a More Advanced Loaning and Reading Culture and Its Information Service *152*
Kati Johanna Launis, Erkki Sevänen

Linked Open Data Service about Historical Finnish Academic People in 1640–1899 ..... *156*
Petri Leskinen, Eero Hyvönen

The 10M Balanced Corpus of Modern Latvian (LVK2018) *157*
Kristīne Levāne-Petrova

Wrangling with Non-Standard Data *165*
Eetu Mäkelä, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi, Terttu Nevalainen

Object Recognition in Illustrated Children Books: Challenges of Applying Computer Vision Systems *166*
Thomas Mandl, Im Chanjong, Helm Wiebke, Schmideler Sebastian

Evaluating, Monitoring and Regulating the Identification of Offensive Content *174*
Thomas Mandl, Prasenjit Majumder, Sandip Modha, Mohana Dave

Keeping It Simple: Word Trend Analysis for the Intellectual History of International Relations *179*
Benjamin G. Martin

What Is Russian Elegy? Computational Study of a Nineteenth-Century Poetic Genre *180*
Antonina Martynenko

Starting Points in French Discourse Analysis' Lexicometry to Study Political Tweets *184*
Marge Käsper, Liina Maurer

Exploring the Potential of Bootstrap Consensus Networks for Large-Scale Authorship Attribution in Luxdorph's Freedom of the Press Writings *187*
Florian Meier, Birger Larsen, Frederik Stjernfelt

Implications of Multifractal Theory for Fictional Narratives – a Dynamic Perspective on Sentiment-Based Story Arcs Exemplified by Ishiguro's Never Let Me Go *188*
Kristoffer Nielbo, Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao

Evolving Political Keywords 1945–1989: Clustering Word Distributions in 3100 Swedish Governmental Official Reports .. *190*
Fredrik Norén, Roger Mähler

Detecting Social Structures Using Library Loan Data *194*
Olli Nurmi, Kati Launis, Erkki Sevänen

Arctic Visible: Mapping the Visual Representations of Indigenous Peoples in the Nineteenth-Century Western Arctic *198*
Eavan Fiona O'Dochartaigh

Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task *201*
Emily Öhman

Handwritten Text Recognition and Linguistic Research *202*
Erik M. Petzell

Digital History of Virtual Museums: The Transition from Analog to Internet Environment *206*
Nadezhda Povroznik

Computer-Based Identification of Metric Verse Structures in Literary Prose of Portuguese Language *207*
Joao Queiroz, Ricardo Carvalho, Angelo Loula

Building a Linked Open Data Portal of War Victims in Finland 1914–1922 ..... *211*
Heikki Rantala, Ilkka Jokipii, Mikko Koho, Esko Ikkala, Jouni Tuominen, Eero Hyvönen

Personal Names as Mirrors of the Past in Medieval Northwestern Russia *212*
Jaakko Raunamaa, Antti Kanner

A Vaccine Against Fake News *218*
Jon Roozenbeek

The Grammar of Politics. Modelling Technocratic Speech and Argumentation in Parliamentary Debate 1918–2017 ..... *219*
Ruben Ros

Creating an Annotated Corpus for Aspect-Based Sentiment Analysis in Swedish *224*
Jacobo Rouces, Lars Borin, Nina Tahmasebi

Name the Name – Named Entity Recognition in OCRed 19th and Early 20th Century Finnish Newspaper and Journal Collection Data *225*
Teemu Ruokolainen, Kimmo Kettunen

Crowdsourcing Metadata for Audiovisual Cultural Heritage: Finnish Full-Length Films, 1946–1985 ..... *227*
Hannu Salmi, Kimmo Laine, Tommi Römpötti, Noora Kallioniemi, Elina Karvo

Using Word Statistics in Studying Variation of Folksongs *228*
Mari Sarv

Towards an Analysis of Gender in Video Game Culture: Exploring Gender-Specific Vocabulary in Video Game Magazines *230*
Thomas Schmidt, Isabella Engl, Juliane Herzog, Lisa Judisch

Live Sentiment Annotation of Movies via Arduino and a Slider . *231*
Thomas Schmidt, David Halbhuber

Distant Reading of Religious Online Communities: A Case Study for Three Religious Forums on Reddit *232*
Thomas Schmidt, Florian Kaindl, Christian Wolff

CLARIN in Latvia: From the Preparatory Phase to the Construction Phase and Operation *233*
Inguna Skadiņa, Ilze Auziņa, Normunds Grūzītis, Arturs Znotiņš

Adapting a Topic Modelling Tool to the Task of Finding Recurring Themes in Folk Legends *234*
Maria Skeppstedt, Rickard Domeij, Fredrik Skott

Digital Emotions: Hybrid Structure of Emotional Impacts *235*
Jurģis Šķilters

Discourse on Safety / Security in the Parliamentary Corpus of Latvian Saeima *236*
Ilva Skulte, Normunds Kozlovs

The New Possibilities for Philological Research in the Digital Archive: The Case of "The Voices of Spring" by Maironis *237*
Magdalena Slavinska

Digital Analysis and Machine View on Latvian National Catalogue of Museum Collections *241*
Maija Spuriņa

Impact of Technologies on Political Behaviour: What Does It Mean to Be "Good Digital Citizen" *242*
Ieva Strode

Analyzing Candidate Speaking Time in Estonian Parliament Election Debates *244*
Siim Talts, Tanel Alumäe

Gulag Literature: Looking Through the Glass of Digital Humanities *245*
Kseniia Alexandrovna Tereshchenko

Estonian Language Community ca. 1900: Learning from Linked Metadata *249*
Peeter Tinits

Hearth Tax Digital: New Narratives on Restoration England *254*
Andrew Wareham, Jakob Sonnberger, Theresa Dellinger, Georg Vogeler

Foreignizing the Other: National Identity and the Concept of Aristocrat in Dutch Historical Newspapers *258*
Leon Wessels

From Research to Revitalization: Fighting Language Endangerment with Digital Humanities *262*
Joshua Wilburs

Integrating TEI/XML Text with Semantic Lexicographic Data ... *269*
Tarrin Wills, Ellert Thór Jóhannsson, Simonetta Battista

Beginning Latvian and Lithuanian as University Level Distance Learning Courses – Experiences and Reflections from the Past Two Years of Teaching *275*
Lilita Zalkalns

Limits of Authenticity of Digitized Objects *277*
Alžbeta Zavřelová, Petr Žabička

3D and AI Technologies for the Development of Automated Monitoring of Urban Cultural Heritage *282*
Tadas Ziziunas, Darius Amilevicius

Disappearing Discourses: Avoiding Anachronisms and Teleology with Data-Driven Methods in Studying Digital Newspaper Collections *287*
Elaine Zosa, Simon Hengchen, Jani Marjanen, Lidia Pivovarova, Mikko Tolonen

Collecting and Storing the Historical Statistics Data on Baltic Countries in 1897–1939 ..... *292*
Giedrius Žvaliauskas

TOPIC INDEX 296

AUTHOR INDEX 299

ORGANIZERS

lnb.lv

lulfmi.lv

Artificial Intelligence Laboratory, Institute of Mathematics and Computer Sciences, University of Latvia

ailab.lv

PROGRAMME COMMITTEE

Sanita Reinsone (Latvia), chair, sanita.reinsone@lulfmi.lv
Ilze Auziņa (Latvia)
Anda Baklāne (Latvia)
Jānis Daugavietis (Latvia)
Eva Eglāja-Kristsone (Latvia)
Koraljka Golub (Sweden)
Normunds Grūzītis (Latvia)
Olga Holownia (Iceland)
Jānis Kreicbergs (Latvia)
Sandis Laime (Latvia)
Liisi Laineste (Estonia)
Veronika Laippala (Finland)
Ditte Laursen (Denmark)
Bente Maegaard (Denmark)
Annika Rockenberger (Norway)
Inguna Skadiņa (Latvia)
Olle Sköld (Sweden)
Mikko Tolonen (Finland)
Jurgita Vaičenonienė (Lithuania)

SUPPORTERS

Nordic Council of Ministers

Ministry of Education and Science Republic of Latvia

VALSTS KULTŪRKAPITĀLA FONDS

Budget sub-programme No. 05.04.00

VPP Valsts pētījumu programma

Digital Resources for Humanities: Integration and Development
No. VPP-IZM-DH-2020/1-0001

ARCTIC PAPER

GALE A Cengage Company

DHN2020 ABSTRACTS

ID: 179
Poster
Topics: library & information science, cultural heritage collections, data mining / text mining, digital resources – publication and discovery, GLAM: galleries / libraries / archives / museums, interdisciplinary collaboration, computational science, citizen humanities, citizen science
Keywords: Labs

Open a GLAM Lab

Aisha Al Abdulla[3], Sarah Ames[12], Paula Bray[10], Gustavo Candela[5], Sally Chambers[11], Caleb Derven[4], Milena Dobreva[9], Katrine Gasser[1], Stefan Karner[13], Kristy Kokegei[6], Ditte Laursen[1], Mahendra Mahey[8], Abigail Potter[2], Armin Straube[9], Sophie-Carolin Wagner[13], Lotte Wilms[7]

[1]Royal Danish Library, Denmark; [2]Library of Congress Digital Innovation Lab, US; [3]Qatar University Library, Qatar; [4]University of Limerick, Ireland; [5]Biblioteca Virtual Miguel de Cervantes, University of Alicante, Spain; [6]History Trust of South Australia, Australia; [7]KB Research Lab, The Netherlands; [8]British Library Labs, UL; [9]UCL Qatar, Qatar; [10]State Library of NSW, UK; [11]Ghent Centre for Digital Humanities, Ghent University, Belgium; [12]National Library of Scotland, UK; [13]ONB Labs, Austrian National Library, Austria

In the age of digital production and transformation, Labs are one of the most significant and disruptive influences on organisations such as Galleries, Libraries, Archives and Museums (GLAMs). All over the world, cultural heritage institutions are witnessing the value and dynamism Labs bring to their collections, making them more accessible, used, shared and enjoyed by their users, embracing innovation, development, experimentation, new ideas through disruptive thinking, and generating opportunities. Labs are living, progressive and transformational. They push boundaries, open up new perspectives, create content and encourage engagement with communities.

This poster will present a new book on GLAM Labs. The book is a collective outcome with contributions from 16 people from 14 cultural heritage organisations and universities around the world. The themes reflected in this book, such as being open to experimentation, risk-taking, iteration and innovation, also capture the methodology of the book, which was written in a collective process during five days.

The book describes what an Innovation Lab is in the GLAM context, what an Innovation Lab is for, and, how to make one happen. The book addresses characteristics, aims and objectives, processes and prospects, tools and services, as well as legal, financial and operational issues. Significantly, the book addresses how libraries, archives, museums, heritage institutions and users can operate and benefit from Innovation Labs.

More specifically, the following themes are covered in the book:

Introducing GLAM Labs

A Galleries, Libraries, Archives and Museums (GLAM) Lab is a place for experimenting with digital collections and data. It is where researchers, artists, entrepreneurs, educators and the interested public can collaborate with an engaged group of partners to create new collections, tools, and services that will help transform the future ways in which knowledge and culture are disseminated. The exchange and experimentation in a Lab are open, iterative and shared widely.

Building a GLAM Lab

Building a GLAM Lab involves defining its core values to guide future work, fostering a culture that is open, transparent, generous, collaborative, creative, inclusive, bold, ethical, accessible and encourages a mindset of exploration. The Lab should be grounded in user-centred and participatory design processes and its staff should be able to clearly communicate what the Lab is about. It's important to think big but start small and establish quick wins to get up and running. This chapter describes why and how to open a GLAM Lab and encourages participation in a movement that can transform organisations and the communities they partner with.

GLAM Lab Teams

There are recommendations for the qualities and skills to look for in Labs teams, how to go about finding allies within and outside the institution, and ideas on how to create a nurturing environment for teams to thrive in. Labs teams have no optimal size or composition, and its team members can come from all walks of life. Teams need a healthy culture to ensure a well-

functioning Lab which might be augmented intermittently by fellows, interns or researchers-in-residence. For a Lab to have lasting impact it must be integrated into the parent organisation and have the support of staff at all levels.

User Communities

GLAM Labs will need to engage and connect with potential users and partners. This means rethinking these relationships to help establish clear and targeted messages for specific communities. In turn, this enables Labs to adjust their tools, services and collections to establish deeper partnerships based on co-creation, and open and equal dialogue.

Rethinking Collections and Data

This chapter discusses the digital collections which are an integral part of Labs. It provides insights on how to share the collections as data, and how to identify, assess, describe, access, and reuse the collections. In addition, there is information about messy and curated data, digitisation, metadata, rights and preservation.

Transformation

Experimentation is the core of the Lab's process. Insights about how to transform tools to operational services are demonstrated. It shows that experimentation can prepare the organisational culture and services for transformation. There is an examination of funding and the advantages and disadvantages of various models through discussion of the different mechanisms and options that an organisation can apply to Lab set-ups.

Funding and Sustainability

This chapter provides insights on how to plan for a Lab's sustainability as well as a step-by-step guide for when an organisation is retiring or decommissioning a Lab.

Curious? Come and see our poster or get involved!

ID: 145
Short paper presentations
Topics: cultural studies, gender studies, theology, corpus linguistics, cultural heritage collections, digitisation – theory and practice, discourse analysis, religious studies
Keywords: sermons, large scale text studies, gender, discourse, theological concepts, semantic network analysis

New Empirical Resources for the Study of Contemporary Preaching ï Presenting a Danish Sermon Corpus through Two Cases on Cultural Conformity and Representation of Christian Concepts

Anne Agersnap, Kirstine Helboe Johansen, Uffe Schjødt, Kristoffer Laigaard Nielbo, Ross Deans Kristensen-McLachlan

Aarhus University, Denmark

Pastors within the Evangelical Lutheran Church of Denmark (ELCD) continually produce an enormous religio-cultural text material, when they write their weekly sermons. These sermons are individually prepared; but they are also the product of a communal practice, since pastors prepare them synchronically within the same national context. This text material thus contains valuable and comprehensive knowledge regarding pastors' representation, interpretation and engagement with respectively biblical, historical and contemporary narratives.

The growing field of Digital Humanities has paved the way for new methods to study sermons as collective text productions. Through computational tools, it has provided pertinent approaches to detect text components containing cultural information and their connections in large corpora.

We have therefore constructed a corpus of 11955 contemporary Danish sermons written from 2011–2016 by pastors in the ELCD. With this paper, we present this newly established corpus and introduce two analytical cases; they illustrate how the corpus can be utilized to explore core questions regarding the pastoral preaching to congregations. Case 1 presents a study on gender constructions and

emphasizes cultural conformity as a pertinent aspect of preaching. In case 2, we explore how pastors through key theological concepts activate the Christian symbolic system in contemporary sermons.

Sermons: Genre and Research Approaches

A sermon represents a unique oral event carried out by a specific pastor of a specific parish at a specific time. Once the pastor steps down from the pulpit, the content of that individual sermon is no longer directly accessible, neither to the congregants nor to the researcher. This transient nature of sermons complicates a systematic study of general themes and discourses conveyed by pastors of the ELCD.

Recent decades of sermon studies have accommodated research designs to study this performativity of preaching. Based on primarily observational studies and interviews, much research has focused on congregants' and pastors' experience of the transmitted sermon content. This has significantly improved our understanding of preaching as a form of ritual communication between pastors and congregants. Meanwhile, though sermons are intended to be performed in church, they are typically highly scripted, as pastors carefully prepare them in the week before the service – most often in the form of written manuscripts. These documents contain valuable knowledge; not only about dialogues between pastor and congregants, but especially about pastors' dialogues with various cultural sources – such as biblical texts, news, literature, television etc. – that confront pastors during the preparation of their sermons. This requires a focus on the transmitted content itself; on sermons as texts containing pivotal knowledge about pastors' engagement with religious, historical and contemporary culture.

To explore the content of sermons, previous qualitative studies have focused on few written sermons by specific pastors. Yet, the ELCD produces between 1500–2000 sermons on a weekly basis, so analyses of few individual contributions are unlikely to reveal any insights on the nature of ELCD sermons on a broader spectrum. Indeed, how Danish pastors collectively

represent major cultural themes, events and discourses in sermons is largely unknown.

This calls for a need to integrate digital resources in the study of contemporary sermons. In this endeavour, especially corpus linguistic practices have provided pertinent solutions in regards to archiving and analysing our corpus of 11.955 sermon manuscripts; all written for church services in the ELCD by pastors beforehand. From metadata annotations, we can recapture the external contexts of the sermons with focus on pastor, location and time and thus explore whether these factors explain content. In the study of sermons, time must be approached from two perspectives. On the one hand, sermons adhere to a linear timeline, as they are prepared anew every week and thus prone to absorb themes and events from pastors' immediate surroundings. But they also adhere to a cyclical timeline, since pastors are expected to interpret biblical passages that are officially prescribed by the ELCD and reoccurring on yearly basis. Both dimensions are reflected in the metadata of the corpus through information on specific date as well as name of holiday.

However, to understand what influences representations of themes and discourses in sermons, we need to understand first how these are textually constructed. Linguistic annotations and semantic network analysis can provide access to such content and their semantic structures. In the following cases, we will focus on these approaches.

Case 1: Gender Discourse

In our first case, we demonstrate an approach to uncover discursive themes from linguistic information in order to understand how pastors represent discourses on gender in sermons.

In 2014 a counting disclosed that the gender balance among Danish pastors had shifted with an overweight of 55,9 % female pastors. Nevertheless, explicit discussions on theology and gender in public debates are not prominent among Danish pastors compared to other – especially Nordic – Lutheran countries, where feminist theologies are thriving. Thus, gaining

insights into gender discourses in a Danish theological context requires uncovering them implicitly. The sermon corpus presents pertinent data for this endeavour, as it is a rich text collection, in which pastors willingly or unwillingly construct gender since distinguishing binarily between male and female gender is an innate feature of Danish language. From a POS-tagged version of the sub-corpus, we extracted collocations of gendered pronouns in subject or object position (he, she, him, and her) and associated verb within a sentence. This process revealed that pastors mention male characters approximately six times as frequently as female characters. To measure the associative relationship between verb and gendered pronoun, we calculated a pointwise mutual score on every verb in relation to each of the four pronouns. Based on inductive codings of the verbs in relation to pronouns, we find that males tend to have more distinct, but diverse roles (either authoritative or victimized) compared to female agents, who carry more diffuse and varied roles. This indicate that female agents are less remarkable or central characters in the corpus.

Case 2: Love and Sin

The sermon corpus further constitutes a unique source for uncovering how the symbolic system of Christianity unfolds in contemporary preaching practices. In our second case, we explore this aspect of preaching, as we demonstrate how semantic network analysis can provide knowledge about pastors' utilization of Christian concepts and their conceptualization in the corpus. We focus specifically on the concepts "love" and "sin" and their respective associative structure. Both concepts are central Christian concepts, but their incorporation in contemporary preaching seems to deviate, as love is represented far more often than sin in the sermons; love is the 98th most frequent word in the corpus, whereas sin is number 492. From the semantic network analysis, we find two rather separate networks around each concept. The words around "sin" illustrate negative actions that can be understood as either preconditions for sin such as "trespassing", or consequences of sin, such as "atone" or "punishment". "Love" is characterized entirely by positive

connotations, such as “devotion”, “unlimited”, “generosity”, and “mercy” describing love as a state of carefree idyllic being. Though the semantic network distinguishes the two concepts “love” and “sin” rather clearly, “forgiveness”, a term primarily associated with “love”, become a mediating concept between love and sin. This indicates a narrative structure of a movement from sinfulness as a temporary state of being to love as an eternal one. Based on the raw word count as well as the associative structure of both concepts, our findings thus suggest that love is a considerably more pertinent concept compared to sin in contemporary Danish preaching.

Perspectives

From the two analytical cases, we wish to emphasise how digital methods provide new and pertinent resources for the study of contemporary sermons. The methodological endeavours for this corpus have been to find approaches to gain intimate knowledge about aspects concerning pastors’ dialogues with Christianity and culture. In this effort, corpus linguistics facilitate a systematic approach to capture textual components of which cultural discourse is build. Further, semantic network analysis enables uncovering conceptualizations in a genre, where activating and interpreting specific themes and concepts are of the essence. Though both methods imply removing terms from their immediate contexts within the corpus, they allow for detailed knowledge about specific text components; both approaches entail re-reading our distant readings closely, as we study how the extracted textual components are embedded semantically in the corpus. Through these rather qualitative readings of collocations and networks, we find potential for establishing “thick descriptions” of repetitive structures within comprehensive data. When working on a large corpus of full texts never read before, we have found it important to find various methods that can help us getting to know this particular text collection. For this purpose, the shift between distant and close readings seems promising.

ID: 113
Poster
Topics: historical studies, library & information science, copyright / licensing / Open Access, cultural heritage collections, data mining / text mining, digitisation – theory and practice, digital resources – publication and discovery, GLAM: galleries / libraries / archives / museums, project design / organization / management, web research, archiving
Keywords: digitization, monographs, licensing, publishing, open access

A Proposed Workflow for Future Monograph Digitization Projects

Niklas Kristian Alén

Suomalaisen Kirjallisuuden Seura / Finnish Literature Society, Finland

As the cost of digitization has come down and the open access movement has gained worldwide momentum, many learned societies have started looking into digitizing their own publications. In this poster session we'll be taking a closer look at the Finnish Literature Society's (Suomalaisen Kirjallisuuden Seura, SKS) and the Finnish Historical Society's (Suomen Historiallinen Seura, SHS) joint-project to digitize all of the SHS's publications during the period 1866–2000. The poster will first present the realized workflow, and then propose an improved one for future projects.

During the last decade there has been a substantial amount of scholarship to a varying degree of granularity, ranging from practical guides to white papers and scholarly research, on the topic of digitization. The Finnish Literature Society anchored its workflow on current best practices. Based on our realized workflow, we found that the most time-consuming part of the process was securing authors' consent to publish their work online. We would therefore strongly suggest that this phase be prioritised in any future digitization projects.

The Finnish Literature Society is a learned society founded in 1831. To consists of a library, an archive, a publishing house, an expert organization for the export of Finnish literature (FILI) and a research department.

In 2016 the Finnish Literature Society received a grant from the Finnish Association for Scholarly Publishing for the digitization and publication of the Finnish Historical Society's published

works ranging from 1866 to 2000. The main objective of the project was to digitize and publish in open access form all monographs and edited volumes published in 8 different SHS scholarly series. The digitization project was carried out by the publishing house and the research department.

The digitized material constitutes a major historical and cultural corpus that incorporates the main research output of historical research in Finnish. As an established open access publisher the SKS also knew that by offering this valuable resource online, free of charge and under a Creative Commons license, it would greatly facilitate Finnish historical research and also improve its dissemination, visibility and its potential impact. Successful open access publishing is, however, far more complicated than simply uploading digitized material to the internet. To be successful the material has to be indexed and distributed through the optimal channels. Search engines must be able to index the whole texts (SEO, Search Engine Optimization), the books need to have adequate metadata and they need to have persistent identifiers of one kind or another. Next we'll be describing the projects workflow and after this we will be offering a revision of the workflow.

The 1st phase in the digitization project was to chart the number of books published and the corresponding number of books available in libraries and archives. The necessary information was collated from library catalogues, SHS online book lists and old publication databases. All of this data was then merged into a relational database that gave us a general overview of the situation. Most of the books could be found in the SHS publication archive. There were, however, gaps in the records and these where filled with book loans from different libraries.

The 2nd phase was to decide on the optimal file format. Here it was crucial to settle on an adequately futureproof approach. The chosen file format had to comply with national long-term digital preservation standards (KDK-PAS), the image quality and resolution should also be adequate to ensure that at least in the short term there would be no need to rescan the material. Besides these archival requirements, the file format also had to

be user friendly. It had to be a well-established and widely used file format that scholars were used to. It was also important that the file format could aid in the discoverability and better usability of the book by integrating an OCR-text layer with the image layer. For these reasons the PDF file format was chosen. It was also decided that two PDF versions would be produced at the same time: the first would be a KDK-PAS compliant high resolution colour scan PDF with an OCR text-layer (PDF/A-1b), and the second would be a web optimized PDF also equipped with an OCR text-layer. The web optimized PDF file had to be a good compromise between file size and resolution. Here it was settled on a MRC (mixed raster content) PDF file. This file format segments each image into different layers, and then applies an optimal compression to each layer . The addition of an OCR-layer also made it possible to produce a corpus-like XML-file. If required this file could be used as research data in linguistic research.

The 3rd phase was to choose a digital repository that would support persistent identifiers, high quality metadata and have a good SEO. After careful consideration SKS chose to use a DSpace instance provided by the Finnish National Library. This instance allows for the use of URN-PID's administered and maintained by the Finnish National Library. It also supports the Dublin Core vocabulary which met our metadata requirements. The DSpace instance also allows for its material to be indexed in the national Finnish portal for libraries, archives and museums Finna and BASE, which dramatically improves the dissemination and visibility of the monographs.

The 4th phase was to produce the required metadata for the books. As most of the books were catalogued in the National Bibliography of Finland (Fennica) which has an open API, it only made sense to harvest these records. Here the SKS's library could lend its expertise, in describing which MARC fields were critical and to which Dublin Core elements they could be mapped. After this a programme was written that harvested the records in MARCXML and mapped them to a Dublin Core dialect that DSpace understands. After examining the records it was, however, decided that supplemental cataloguing was

required. To facilitate this, an online metadata editor application was developed. This editor was used by a third party to check and supplement the metadata. The application allowed multiple specialists to edit the metadata simultaneously. After editing, the application allowed the export of all records in a DSpace compliant format. The supplemented metadata was then uploaded alongside the web optimized PDF file to SKS's DSpace instance.

The 5th phase, which was going on in conjunction with steps 3-4, consisted of acquiring authors' consent to license their works under a Creative Commons –license as well as the right to upload their monographs to the internet. This phase of the project proved to be much more time consuming and complex than what we had anticipated. After quickly making contact with well-known and well-established authors as well as identifying all authors whose copyright had expired, the project began to face mounting difficulties. Many authors had passed away and contacting all of their heirs proved to be too time consuming, others had retired after only authoring a couple of articles during their career, others could not simply be found (their names are too common, they've moved or changed names etc.), and last but not least there were a sizable number of foreign authors.

Our experience shows clearly, and we would strongly suggest, that any future digitization efforts should focus on phase 5 first. This is the most time-consuming phase as technology can't really facilitate it in any way. A thorough investigation in this phase may even dictate whether or not the digitization effort is feasible. From our experience all other issues can be solved by innovative use of technology.

ID: 186
Short paper presentations
Topics: cultural studies, linguistics, corpus linguistics, cultural heritage collections
Keywords: corpus, early written Latvian

Short Texts in the Corpus of Early Written Latvian (www.korpuss.lv/senie)*

Everita Andronova

The Institute of Mathematics and Computer Science, University of Latvia, Latvia

Early written Latvian texts are important sources not only for humanities, but also in culture and social studies. Unfortunately, being scattered in different libraries and archives (in different countries), they have not been much investigated; they are very much treated isolated and in many cases are used for quite narrow purposes. There was a serious lack of general overviews introducing the sources and studies on them, and more important, even now there are still a few interdisciplinary studies carried out. Fortunately, the last two decades have seen a growth in popularization and dissemination of the early written sources. The 21st c. brought new chances for lesser-used and lesser-studied languages, namely, the era of digitalization has resulted in the development of different general and special corpora.

The diachronic Corpus of early written Latvian was launched in 2003 and is intended to cover the history of written Latvian of the 16th–18th cc. (Andronova 2007). The aim of the corpus is to facilitate studies of early Latvian in general and to serve as the basis for the Historical dictionary of the Latvian language (this is a good example of successful co-operation between linguists and software engineers in creating a new kind of dictionary in Latvian lexicography; 1200 pilot entries are now available on the web: www.korpuss.lv/lvvv).

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

The development of the corpus has gone through several phases. Early written Latvian texts have been acquired thanks to close co-operation with Latvian and Lithuanian libraries, as well as with researchers across Europe interested in the history of early Latvian texts. Undergraduate students at the University of Latvia and St. Petersburg State University (Russia) have also been involved in the process of transliterating some texts during the compilation of the corpus. This has served to raise the interest of the history of the Latvian language, and subsequently some bachelors' theses have been defended on the basis of these texts.

The first digitized text copies were handed over to the National Library of Latvia in 2002. Some new sources have been discovered since then: thus e.g. a unique copy of Agenda Parva (1622), earlier reported unknown, has recently been published on the website of the Warmia-Mazury Digital Library (http://wmbc.olsztyn.pl/dlibra/doccontent?id=926). We are presently processing Latvian fragments in this Agenda that will be added to the corpus.

One of the challenges in this work is the crucial need of comparison between different editions of the same source, as well as an analysis showing the tradition of circulation of different parts of religious texts from one source to another.

One of the advantages of this corpus is that it provides the exact location of a word-form (usually the abbreviation of the source, page and line number of the text or the Bible Book, chapter and verse). This makes it easy to accurately cite the corpus data. There is a possibility to look at facsimiles of the sources as well, which gives an extra added-value to this resource.

All sources in the corpus are included in toto, no samples are chosen. Quite a wide range of short texts has either been added to the corpus recently or is presently in the process of being included; these texts can be divided into 3 groups:

1) individual short texts, e.g., occasional poetry, oath texts;

2) Latvian texts found in sources written in other foreign languages, e.g., the prayer Pater Noster published in the 16th

c.; sentences in Latvian in several editions of ‘Stratagema oeconomicum oder Akker-Student’ written in German by S. Gubert in the 17th c. or Latvian text in Agenda Parva (1622 and later editions);

3) shorter texts in Latvian appended to some individual Latvian sources.

The description of these three groups and the methodology of their inclusion in the corpus is the topic of the present study.

1. Individual short texts

These include both poetry and certain legal texts (different oaths, laws of war court). The bulk of the sources in this group is occasional poetry, written in the 17th and 18th c.

The beginnings of Latvian occasional poetry have recently been the object of in-depth studies. A broad inspection of the 16th and 17th c. poetry in the cultural context has been carried out by Māra Grudule (2017). The book shows the long way of evolution of this type of texts: they were profoundly influenced by German culture but later little by little turned into Latvian poetry. Three early dedication poems were already added to the corpus in 2016. In 2019 around 70 poems from 15 sources have been collected in different libraries and are now in the process of being included in the corpus. One of these new poems is a unicum kept at the Russian National Library – ‘Mūsu visu upurs tai priecas dienā’ (1791). These new poems are of wide thematic range, covering different occasions – birthday congratulations, wedding songs, popular New Year’s wishes, which can be printed on cards or written in letters, funeral songs and others.

These songs may be interesting not only for literature and linguistic studies, but also in order to examine the culture, history and ethnography in Livonia at that time. One can examine New Years dedication poems in ‘Jaunā Gada vēlēšanas pēc ikkatra gribēšanas’ (1781) and ‘Jaunā Gada vēlēšanas’ (1793) not only for literary analysis, but also to understand the soul, psychology and manners of people. Thus, we would like to encourage not only linguistic, but all other

kinds of studies by means of the corpus. These texts will be included in the Corpus as individual sources.

2. Latvian inscriptions in texts written in other languages

This group covers single words, phrases, sentences and longer passages in Latvian in books printed in other languages. Latvian proper names – personal names and places names – have been found in several sources dated to the 15th century (e.g. chronicles). The lists of craftsmen guilds from the 16th c. should be examined and excerpted for the purposes of the corpus). The history of written Latvian rises with the period of Reformation and the claim of Martin Luther to use native language. There are already a number of prayers Pater Noster from the 16th c. in the corpus, before including them a linguistic analysis was performed in order to define which prayer to include (see Vanags 2014).

At the moment 2 new sources are being processed for inclusion in the corpus:

(1) Agenda Parva (1622) with its texts written in Polish, German, Estonian and Latvian. For the needs of the corpus only the Latvian sentences are excerpted and processed, and a Latvian word-list will be created on the basis of this material.

(2) The popular 17th c. book by S. Gubert, 'Stratagema oeconomicum oder Akker-Student' (1st ed. 1645 and later editions in the 17th c.), is a good example of so-called Hausväterliteratur and is a valuable source for ethnographical studies among others (e.g. the description of instruments and agriculture cultures known in Livonia at that time; 'Bauer=Prognosticon' for weather forecast is often mentioned, later included in the volumes of Latvian beliefs compiled by P. Šmits (1940–1941). In this book we can find Latvian phrases and hymnals at the end (last edition printed in 1757 excludes hymnals). Single words and phrases are encountered within the German sentences, commonly introduced by the phrase 'die Bauern nennen', e.g. names of insects (circiņš 'criket'), names of plants (vavieriņi 'marsh tea'), phrases like dvēsel laiks liter. 'time of souls' meaning 'time span between Michael's Day

(29th of September) and Martin's Day (10th of November)). In this case the whole sentence will be copied and marked as German, but only the Latvian phrase will be included in the word list. There are some hymnals added at the end of the book both in German and Latvian (most probably the songs were translated by S. Gubert himself, the last edition printed in 1757 lacks songs). All the songs will be included in the corpus in order to facilitate the analysis of the source text in German and its translation into Latvian.

3. Texts in Latvian added (later) to some individual Latvian sources

At the moment we have only one such source – a letter written by the peasant Anšs to the priest Loder dated June 1771 and added to the transcript of the 'Lettisches und Teutsches Wörterbuch' by Ch. Fürecker. This letter has already been included in the Corpus (http://senie.korpuss.lv/static/V1771_SZA.html) as a separate item.

The development of the Corpus of early written Latvian texts 'SENIE' is an on-going activity within other research projects; in 2018–2020 it is funded by the State Research program 'The Latvian Language' (No. VPP-IZM-2018/2-0002).

References

Andronova, E. The Corpus of Early Written Latvian: current state and future tasks. In: Proceedings of Corpus Linguistics, 2007, Birmingham, UK. Available at: http://ucrel.lancs.ac.uk/publications/CL2007/paper/245_Paper.pdf

Grudule Māra. Latviešu dzejas sākotne 16. un 17. gadsimtā kultūrvēsturiskos kontekstos. Rīga (2017).

Vanags Pēteris. Latviešu valodas vēsturiskās vārdnīcas projekts. In: Valodas prakse: Vērojumi un ieteikumi. Rīga (2014), pp. 97–109.

ID: 103
Short paper presentations
Topics: communication studies, cultural studies, design, geography, historical studies, cultural heritage collections, digitisation – theory and practice, digital resources – publication and discovery, geospatial analysis – interfaces & technology, infrastructure, information architectures, information retrieval, interface & user experience design, software design and development, standards and interoperability, visualisation
Keywords: deep mapping, distributed network, spatiotemporal data, statistical analysis, indigenous history

Time-Layered Cultural Map of Australia*

Paul Arthur[1], Erik Champion[2], Hugh Craig[3], Ning Gu[4], Mark Harvey[3], Victoria Haskins[3], Andrew May[5], Bill Pascoe[3], Alana Piper[6], Lyndall Ryan[3], Rosalind Smith[3], Deb Verhoeven[7]

[1]Edith Cowan University, Australia; [2]Curtin University, Australia; [3]University of Newcastle, Australia; [4]University of South Australia, Australia; [5]The University of Melbourne, Australia; [6]University of Technology Sydney, Australia; [7]University of Alberta, Edmonton, Canada

This paper reports on an Australian project that is developing an online system to deliver researcher-driven national-scale infrastructure for the humanities, focused on mapping, time series, and data integration. Australian scholars and scholars of Australia worldwide are well served with digital resources and tools to deepen the understanding of Australia and its historical and cultural heritage. There are, however, significant barriers to use. The Time-Layered Cultural Map of Australia (TLCMap) will provide an umbrella infrastructure related to time and space, helping to activate and draw together existing high-quality resources. TLCMap expands the use of Australian cultural and historical data for research through sharply defined and powerful discovery mechanisms.

See https://tlcmap.newcastle.edu.au/.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 206
Short paper presentations
Topics: literary studies, information retrieval, scholarly editing, religious studies
Keywords: N.F.S. Grundtvig, 19th Century, letter-spacing, typography, emotional history

Emotional Imprints: Letter-Spacing in N.F.S. Grundtvig's Writings*

Katrine F. Baunvig[1], Oliver S. Jarvis[2], Kristoffer Laigaard Nielbo[2]

[1]Aarhus University, The Grundtvig Study Centre, Denmark; [2]Aarhus University, Centre for Humanities Computing, Denmark

Undertaking a distant reading of letter-spacings in the digitized and annotated N.F.S. Grundtvig data, this paper targets a trait of an overall romanticist emotionalizing trend in a corpus of 19th century literature: It proposes to analyze the letter-spacings as a deposition of heightened attention to subjective emotional experience in printed matter and typesetting in the writings of the Danish poet, priest and politician N.F.S. Grundtvig (1783–1872), who is widely regarded as the central figure in the 19th century Danish religious development and nation building process. As such this paper sketches the temporal and semantic contexts of the letter-spacings.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 136
Short paper presentations
Topics: cultural studies, cultural heritage collections, GLAM: galleries / libraries / archives / museums, linked data / semantic web / ontologies, linking and annotation
Keywords: linked data, annotation, named entities, digital humanities

Linked Data for Digital Humanities Scholars and Researchers: ñRainis and Aspazijaò Collection

Uldis Bojārs, Anita Rašmane, Anita Goldberga

National Library of Latvia, Latvia

This demonstration paper will focus on the Linked Digital Collection “Rainis and Aspazija” https://runa.lnb.lv/en/ that showcases the use of Linked Data in Digital Humanities. This collection offers interlinked digital objects and data from several memory institutions and private repositories related to two Latvian poets of the period of National Awakening. The talk will also introduce the semantic annotation tool developed for cultural heritage needs that was used to create this collection, and it`s potential in Digital Humanities research and teaching.

In 2016, the National Library of Latvia (NLL) together with the National Archives of Latvia, the Institute of Literature, Folklore and Art of the University of Latvia, the Association of Memorial Museums, and the Literature and Music Museum published RunA – the first digital cross-sectoral cultural heritage pilot-collection in Linked Data form in Latvia. RunA highlights the NLL`s efforts in developing new knowledge base for memory institutions and researchers. During 2018–2019, a special semantic annotation tool and a separate entity datastore was developed by the NLL to enhance RunA textual documents analysis. Although there already exist tools handling annotations of entities and links to external sources, they do not exactly provide for specific purposes of historical cultural heritage document research, like correspondence from the late 19th century, archival documents etc.

The RunA annotation tool includes support for three core types of annotations – simple annotations that may link to named

entities, structural annotations that mark up portions of the document that have a special meaning within the context of the document (e.g. – direct citation of another published material) and composite annotations for more complex use cases (e.g., for representing an event described in a document with mentions of place, time and participants, all marked and identified in their own annotations).

The tool allows users to import text documents, create manual annotations and entity pages. The tool also includes a semi-automated named entity recognition technique where entity mentions in unstructured content are identified and linked to the existing entity pages in the annotation tool.

The process of semantic annotation of cultural heritage documents using special NLL`s tool includes all classical stages of annotation: text analysis to identify concepts such as people, things, places, events, etc.; concept extraction, classification of identified entities; manual relationship extraction between known and newly recognized entities; linking entities to internal and external controlled vocabularies; storing entity information with links in the datastore.

Information about the entities referenced from annotations is maintained in a dedicated entity datastore that supports links between entities and can point to additional information about these entities (e.g., to Linked Open Data resources such as VIAF, Wikipedia, etc.). The datastore provides for storing, sharing, and reusing data, extracted from individual annotations and those added by researchers. This allows experts to build a knowledge base about the entities referenced from annotations while annotating documents. This entity information could evolve as the annotation task progresses. It is possible to enhance the completeness of data on entities later. For example, they may create an entry for an entity that needs further research (with comments about what is known about the entity and what is not) which can be extended with additional information (for example, identifiers for the entity in other authoritative data sources) when it becomes available. Machine-readable information about all entities is published

according to Linked Data principles (in Turtle RDF and RDF/XML format).

Expanded annotated materials could be the research subject of students, who, whilst doing research, could become RunA's annotation tool testers. After providing some guidance the NLL plans to involve students in the annotation process of the correspondence of Rainis and Aspazija. Students, educators and early-career researchers will have a chance to learn about the possible methods of using digitized material from cultural heritage collections. The knowledge base generated by the NLL could be integrated into the education process of the study courses at the Faculty of Humanities of the University of Latvia through the use of RunA.

This presentation will give concrete example and recommendations for exploiting the RunA as a resource and the potential of textual documents annotation tool in Digital Humanities research and teaching.

Acknowledgements

This research was supported by the Latvian Council of Science Project Nr. lzp-2019/1- 0365 "Latvian Memory Institution Data in the Digital Space: Connecting Cultural Heritage" and RunAdevelopment was partlyfinanced by the European Regional Development Fund's (ERDF) project Digitization of culture heritage content (1st stage) in 2018–2019.

References

1. Bojārs, U., Rašmane, A., Žogla, A., Bāliņa, S., Salna, E. Semantic Annotation Tool for Cultural Heritage Content. Baltic Journal of Modern Computing, Vol. 6, No. 4 (2018).

2. Goldberga, Kreislere, Rašmane, Stūrmane, Salna. (2018). Identification of entities in the Linked Data collection "Rainis and Aspazija" (RunA). Italian Journal of Library, Archives and Information Science (JLIS.it). V.9, No.1. https://www.jlis.it/article/view/83-106, DOI: http://dx.doi.org/10.4403/jlis.it-12444

3. Bojars, U., Rasmane, A., Zogla, A. The Requirements for Semantic Annotation of Cultural Heritage Content. Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria,

October 22, 2017. CEUR Workshop Proceedings, vol. 2014. URL: http://ceur-ws.org/Vol-2014/

4. Goldberga, A., Rašmane, A. (2017) The Linked Data collection “Rainis and Aspazija” (RunA) and the potential of IFLA FRBR LRM key entities for annotating textual documents. IFLA Library. http://library.ifla.org/1762/

5. Bojars, U. Case Study: Towards a Linked Digital Collection of Latvian Cultural Heritage. Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe 2016), pp.21–26.

ID: 166
Short paper presentations
Topics: medieval studies, linked data / semantic web / ontologies
Keywords: linked open data, medieval studies, data reconciliation, vocabularies

Linked Open Data Vocabularies and Identifiers for Medieval Studies*

Toby Burrows[1], Antoine Brix[2], Doug Emery[3], Mitch Fraas[3], Eero Hyvönen[4], Esko Ikkala[4], Mikko Koho[4], David Lewis[1], Synnøve Myking[2], Kevin Page[1], Lynn Ransom[3], Emma Thomson[3], Jouni Tuominen[4], Hanno Wijsman[2], Pip Willcox[5]

[1]University of Oxford, UK; [2]Institut de recherche et d'histoire des textes, France; [3]University of Pennyslvania, US; [4]Aalto University, Finland; [5]The National Archives, UK

This paper examines the use of Linked Open Data in the research field of medieval studies. We report on a survey of common identifiers and vocabularies used across digitized medieval resources, with a focus on four internationally significant collections in the field. This survey has been undertaken within the “Mapping Manuscript Migrations” (MMM) project since 2017, aimed at aggregating and linking disparate datasets relating to the history of medieval manuscripts. This has included reconciliation and matching of data for five main classes of entities: Persons, Places, Organizations, Works, and Manuscripts. For each of these classes, we review the identifiers used in MMM’s source datasets, and note the way in which they tend to rely on generic vocabularies rather than specialist medieval ones. As well as discussing some of the major issues and difficulties involved in conceptualizing each of these types of entity in a medieval context, we suggest some possible directions for building a more specialized Linked Open Data environment for medieval studies in the future.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 231
Long paper presentations
Topics: social media
Keywords: social media

ñMemesò as Activism in the Context of the US and Mexico Border

Martin Camps

University of the Pacific, US

Memes function as “digital graffiti” in the streets of social media, a cultural electronic product that satirizes current popular events, and can be used to criticize those in power. The success of a meme is measured by its “virality” and the mutations that are reproduced like a germ or a part of the genetic trend of subculture societies. I am interested in analyzing these eckphrastic texts in the context of the construction of the wall between the US and Mexico. I examine popular memes in Mexico and the US from both sides of the border. I believe these “political haikus” work as an escape valve for the tensions generated in the culture wars that consume American politics. The border is an “open wound” (Mexican writer Carlos Fuentes dixit) that was opened after the War of 1847 and resulted in Mexico losing half of its territory. Currently, the wall functions as a political membrane barring the “expelled citizens” of the Global South from the economic benefits of the North. Memes help to expunge the gravity of a two-thousand-mile concrete wall in a region that shares cultural traits, languages, and natural environment, a region that cannot be domesticated with symbolic monuments to hatred. Memes are rhetorical devices that convey the absurdity of a situation, as in a recent popular meme that shows a colorful piñata on the edge of the border, a meme that infantilizes the State-funded project of a fence. The meme’s iconoclastography sets in motion a discussion of the real issues at hand–global economic disparities and the human planetary right to migrate.

The term meme was coined by Richard Dawkins, a British evolutionary biologist, in 1976 in his book The Selfish Gene as

a unit of cultural transmission. He wrote: “We need a name for the new replicator, a noun which conveys the idea of a unit of cultural transmission, or a unit of imitation. ‘Mimeme’ comes from a suitable Greek root, but I want a monosyllable that sounds a bit like ‘gene’. I hope my classicist friends forgive me if I abbreviate mimeme to meme.” (The Selfish Gene 192). There are many popular memes that relate to different cultural trends, such as “Leave Britney Alone,” “Gangnam Style,” “Situation Room,” “Advise Dogs,” “LolCats,” “Success Kid”, but in this presentation, I will concentrate on the genre of “border memes”. I offer a cross- cultural study of border memes, a cultural software, produced in Mexico and the United States about the mutual issue of the border wall that was raised during the 2016 American presidential campaign and continues until now.

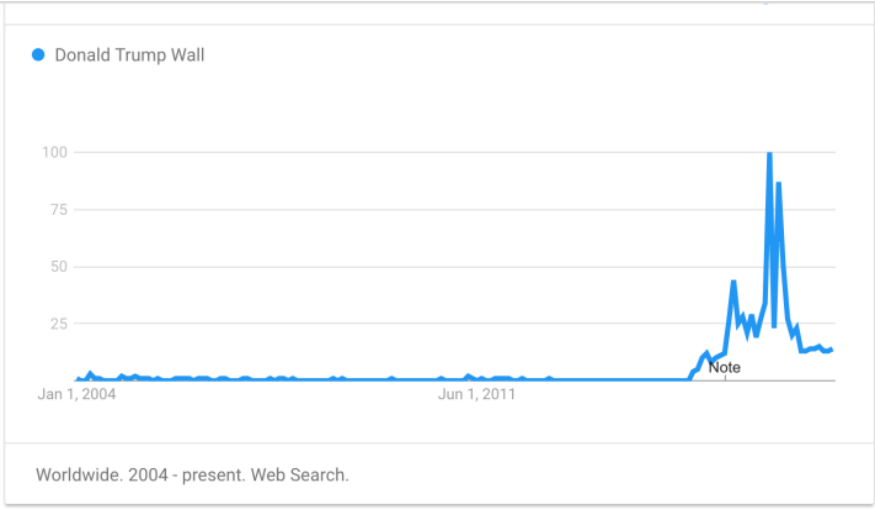


Fig. 1. Increase of border memes after the 2016 Presidential Election USA.


SOME MEME-IGRANTS CROSSING
THE BORDER INTO IMGFLIP.


Weeeeeee!
Back over the border you go!


OH, YOU WERE JUST EXPLORING?
NOT IN MY COUNTRY!


The caravan has disappeared overnight, this was found at the border.....

ID: 203
Short paper presentations
Topics: art history, cultural heritage collections, digitisation – theory and practice, GLAM: galleries / libraries / archives / museums, infrastructure, teaching / pedagogy / curriculum design, user studies / user needs, museology
Keywords: digital literacy, museum, digitization, knowledge transfer, Sweden, Lithuania

No Longer Obsolete: Mapping Digital Literacy Skills for Museum Professionals in Sweden and Lithuania

Nadzeya Charapan

Uppsala University, Sweden; Vilnius University, Lithuania

Contemporary museums as open systems are constantly transforming in response to economic, technological, social and cultural trends. The past decade has witnessed an increasing demand for information about the digitization of, access to, and preservation of museum collections to produce digital cultural heritage and new affordances for visitor-museum encounters. The post-digital turn normalizes the application of the ICTs as a basic attribute of the museum practice for preservation, collection, display and communication functions (Parry, 2010; 2013). Thus, the practitioners must be equipped with the transferable competencies to be able to successfully perform their duties and facilitate the successful digital transformation of the cultural institutions (Borowiecki & Navarrete, 2017).

The previous research into the digital competencies demonstrates the paucity in its understanding and conceptualization of digital literacy (Marty, 2006; Tallon, 2017). For example, Jisc (2014) defines it as “capabilities which fit an individual for living, learning and working in a digital society. Digital literacy looks beyond functional IT skills to describe a richer set of digital behaviors, practices, and identities”. This definition provides a general view of the concept and requires further elaboration and adaptation to the specificity of the museum sector. Moreover, due to the constant and speedy

change, the creative industries sector experiences a permanent gap in transferrable skills (Creative and Cultural Skills, 2011; Howard, 2013). Against the background of these trends, there is a need for further investigation into digital literacy and approaches to assessment and evaluation.

The existing European (eCult Skills 2013–2015, Mu.SA project 2016–2019) and British national research projects (One by One: Building Digital Literacies 2017–2020) serve as important facilitators in addressing the existing research and practice gaps in the digital literacies and advancement of the museum sector, however, the empirically-driven conclusions are partly applicable to the Baltic and Nordic context.

The goal of this paper is to provide a nuanced understanding of how the digital skills and literacies are understood, operationalized, and supplied in the Swedish and Lithuanian museological contexts.

A conceptual model of the museum digital skills ecosystem, suggested by Parry, R., Eikhof, D. R., Barnes, S. A., & Kispeter, E. (2018) is adopted as a theoretical framework to scrutinize the landscape of the digital literacy skills in two case studies.

The paper addresses the following interrelated blocks of research questions:

1. How do national cultural policies and legislation regulate the digitalization of museums and the provision of digital literacy skills in Lithuania and Sweden?

2. How do museum practitioners understand and deploy digital literacy skills in their daily professional practices?

3. What measures are required to bridge the gap (if any) and reach the balance in demand and supply of the skills?

To depict the national peculiarities, the study will use the data from a) desk-study about the evidence on the national museum regulations and digitization in Lithuania and Sweden, and 2) qualitative research methods, based on the in-depth interviews with the museum practitioners to gain a nuanced understanding of how digital skills are developed and deployed in different structural units.

The comparative thematic analysis of Kulturarvspolitik (2017) and Museilag (2017), in Sweden; and New National Museum Decree (2018) in Lithuania will create the legislative framework for the analysis of the existing regulations and infrastructures. Furthermore, the empirical data will be obtained from the museum professionals of two national art museums: the Nationalmuseum (Stockholm), incorporating Digital Laboratory; and Lithuanian Art Museum (LAM), incorporating Lithuanian Museums' Centre for Information, Digitisation. The choice of the museums is determined by the following factors: similarity of the institutional context – art museums; the status – both museum are national cultural institutions; and they both serve as national digital hubs, incorporating the Digital Laboratory (Nationalmuseum), and Lithuanian Museums' Centre for Information, Digitisation (Lithuanian Art Museum).

The empirical data will benchmark the national peculiarities of the digital skills ecosystems and digitization processes in Lithuania and Sweden. The Baltic-Nordic comparative perspective will generate a consolidated view on the digitization of the museum sectors, discussing the existing threats and opportunities for digitalization, as well as supply and demand of the digital competencies.

As an outcome, a set of recommendations for the prospective collaboration and knowledge transfer will be developed. These guidelines will provide a glimpse into nationally-tailored and regional specificity of digital skills ecosystems that will address the existing gap.

References

Borowiecki, K. J., & T. Navarrete (2017). Digitization of Heritage Collections as Indicator of Innovation. Economics of Innovation and New Technology, 26, 3, 227–246.

Creative and Cultural Skills (2011). Sector Skills Assessment for the Creative Industries of the UK. London: Creative and Cultural Skills. Available from: https://creativeskillset.org/assets/0000/6023/Sector_Skills_Assessment_for_the_Creative_Industries _-_Skillset_and_CCSkills_2011.pdf

eCult Skills [Desk and Field Research: Guidelines and Templates] V.1.0. Available from: http://files.groupspaces.com/eCult/files/1152507/RQMMdZeHqGSV1EEiHKk5

/R2a+%26+R3a+Methodology+for+identification+of+K%2C+S%2C+C+needed+in+the+e-cult+sector+%26+Trainings+availalbe+in+the+EU.pdf

Jisc (2014). Developing Digital Literacies (online guide). Bristol: Jisc. Available from: https://www.jisc.ac.uk/guides/developing-digital-literacies

Howard, K. (2013). GLAM (Re-)Convergence and the Education of Information Professionals. Paper presented at A GLAMorous Future? Reflecting on Integrative Practice Between Galleries, Libraries, Archives, and Museums. Victoria University, Wellington, New Zealand.

Lithuanian Art Museum. Available from: https://www.ldm.lt/en/

Lithuanian Museums' Centre for Information, Digitisation. Available from: https://www.limis.lt/en/projektas

Marty, P. F. (2006). Finding the skills for tomorrow: Information literacy and museum information professionals. Museum Management and Curatorship, 21, 4, 317–335.

Mu.SA: Museum Sector Alliance (2019). Available from: http://www.project-musa.eu/about/

Nationalmuseum, Available from http://collection.nationalmuseum.se/

Parry, R. (ed.) (2010). Museums in a Digital Age. Abingdon and New York: Routledge.

Parry, R. (2013). The End of the Beginning: Normativity in the postdigital museum. Museum Worlds, 1,24–39. Pedro, A. R. (2010). Portuguese Museums and Web 2.0. [Os museus portugueses e a Web 2.0]. Ciencia da Informacao, 39, 2, 92–100.

Parry, R., Eikhof, D. R., Barnes, S. A., & Kispeter, E. (2018). Mapping the Museum Digital Skills Ecosystem-Phase One Report.

Tallon, L. (2017). Digital is More Than a Department, it is a Collective Responsibility. The Met. Published 24 October 2017. Available from: https://www.metmuseum.org/blogs/now-at-themet/2017/digital-future-at-the-met

ID: 163
Short paper presentations
Topics: linguistics, corpus linguistics, teaching / pedagogy / curriculum design, visualisation, computational science
Keywords: word frequencies, visualization, lexical diversity, Zipf

Comparing Word Frequencies and Lexical Diversity with the ZipfExplorer Tool*

Steven Coats

University of Oulu, Finland

The ZipfExplorer is a tool for the interactive comparison and visualization of shared word type frequencies for two texts or corpora. The tool can be used to give insight into similarities and differences in textual and discourse content in terms of individual keywords or groups of keywords, and also calculates several measures of lexical diversity for the shared types of the selected texts. A selection of texts and corpora can be analyzed, and users can upload their own files for interactive comparison.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 204
Poster
Topics: cultural studies, diversity and multilingual / multicultural approaches, visualisation, ethnography
Keywords: multilingualism, language diversity, maps

Digital Maps for Linguistic Diversity

Coppélie Cocq

University of Helsinki, Finland

Language maps have a central role in educational books, atlases, etc, illustrating pedagogical efforts toward a presentation of linguistic data. The characteristics of languages are however problematic to render on a map: flows and movements as well as the lack of clear borders, for instance, demand contextualization and clarifications that can hardly been rendered. Language maps have therefore been criticized for being “generalized snapshots in time of a variable that is in constant change” (Luebbering et al 2013a: 383; 2013b).

This poster takes its point of departure in a research project investigating linguistic landscapes, that is landscapes constructed by the combination of “road signs, advertising billboards, street names, place names, commercial shop signs, and public signs on government buildings” in a given “territory, region, or urban agglomeration” (Landry & Bourhis, 1997:25). By studying which languages materialize in our surroundings, we can reach an understanding about which languages that are used and represented in a society, and thereby about which languages that are allowed to be seen and which are not, providing information about language discourses, policy and power relations.

Digital mapping for visualization can offer solutions for meeting the challenge of successfully representing linguistic diversity – and this is what this poster proposes to discuss. One of digital humanities most valuable contributions is within the area of visualization. It is not only a mode to convey scientific results in graspable packages – it is also a way to raise new questions, make visible new patterns and causal relations between variables. Digital maps, more specifically deep maps (Least Heat- Moon 1991; Bodenham & Corrigan 2015) can combine

complex layers based on various data linked to objects and enable the user to interactively compare these different layers. Geographic Information Systems (GIS) allow for a greater flexibility in the use of data in term of accessibility, analysis and display, which in one of the reasons why digital cartography is today an increasing area in visualization studies (Foka, Buckland, Cocq, Gelfgren, forthcoming; Luebbering et al 2013a; 2013b).

Language maps have been criticized for being oversimplified (Mackey 1998), for failing to represent today's diverse linguistic environment and for embedding issues of power and perception, for instance in cartographic decisions (Luebbering, Kolivras and Prisley 2013b), with implications for the representation of various groups of language speakers. Here, we will seek to discuss digital forms of visualization that are non-authoritative and allow to render the flows and dynamism of languages and language use.

Such form of visualization shall not only serve as methodological tools, but also as a means for communication knowledge about the presence of languages and their speakers. Thereby, with this proposal we wish to contribute to an increased awareness about linguistic diversity and multilingualism – a first step for developing means for inclusion, for understanding place-making processes and apply this knowledge to the creation of public spaces that are inclusive and provide a better understanding and the prevention of segregation.

References

Bodenham, D. Corrigan J. & Harris T. (2015). Deep Maps and Spatial Narratives. Indiana University Press.

Foka, Anna, Cocq Coppélie, Buckland Phillip I. & Gelfgren Stefan Mapping Socio-ecological Landscapes: Geovisualization as Method. In: Routledge Research Methods Handbook: Digital Humanities, eds. Stuart Dunn and Kristen Schuster (Routledge, forthcoming).

Landry, R. & Bourhis, R.Y. (1997). Linguistic landscape and ethnolinguistic vitality: An empirical study. Journal of language and Social Psychology, 16, 23–49.

Least Heat-Moon, William PrairyErth: A Deep Map. (1991) Boston: Houghton Mifflin Company.

Luebbering, Candice R., Korine N. Kolivras & Stephen P. Prisley (2013a) The lay of the language: surveying the cartographic characteristics of language maps, Cartography and Geographic Information Science.

Luebbering, Candice R., Korine N. Kolivras & Stephen P. Prisley (2013b) Visualizing Linguistic Diversity Through Cartography and GIS, The Professional Geographer, 65:4, 580–593.

Mackey, W. F. 1988. Geolinguistics: Its scope and principles. In Language in geographic context, ed. C. H. Williams, 20–46. Philadelphia: Multilingual Matters.

ID: 187
Short paper presentations
Topics: art history, historical studies, medieval studies, image processing, artificial intelligence
Keywords: image/document segmentation, document and image processing, artificial intelligence, machine learning, medieval manuscripts

An Artificial Intelligence Approach to Segmenting Medieval Manuscripts with Complex Layouts

Lisandra S. Costiner[1], Lizeth Gonzalez Carabarin[2]

[1]Merton College, University of Oxford, UK; [2]Eindhoven University of Technology, The Netherlands

Digitization initiatives undertaken by libraries, museums and collections around the globe are rapidly increasing the number of manuscript images online. Given the large volume of such data, it is important to devise new ways to automatically process and extract relevant information from these images, saving valuable human time invested in manual transcription and image extraction.

Digitized documents pose a number of challenges for the extraction of relevant information, the key ones being the location of areas of text and illustration. Medieval manuscripts are especially challenging for automatic segmentation. Each surviving book was hand produced so its page layout, script used, and illustrations widely vary. Furthermore, medieval decorations do not typically conform to uniform rectangular registers – they can be unframed, be placed throughout the text at irregular intervals and extend into page margins. Given this, such documents pose particular difficulties for traditional methods of segmentation designed for printed text, requiring instead the development of customized algorithms.

Although many techniques have been developed for image segmentation (Eskenazi et al, 2017), there is a need for a generic tool that is flexible in dealing with a range of documents, low on processing power, and white-box, allowing every step to be queried. This paper proposes such a technique for the automatic identification and extraction of

images (illuminations or miniatures), and of lines of text from Western medieval manuscripts.

Algorithms for the extraction of images and texts in layout analysis (segmentation) can be generally divided into three classes. Most of the approaches employed in document segmentation are adapted to specific types of records (Shafait et al, 2008), so there is a need for a global or generic approach that will be able to adapt to different types of documents. Older approaches rely on rule-based algorithms which have reduced versatility, generality, robustness and accuracy when segmenting hand-written documents (Shafait et al, 2008). Recent developments have tended to focus on the use of neural networks (Eskenazi et al, 2017) (Gao et al, 2017) (Ares Oliviera et al, 2018). Although effective, neural networks (NNs) require manually-annotated data for training, expending large amounts of human time; they are computationally heavy, and are black boxes, meaning that their inner workings are not understood. New approaches with increased versatility, stability, generality, ability to perform multi- scale analysis, and to handle color remain a desiderata (Eskenazi et al, 2017).

The current approach proposes to address these needs. It is based on k-means algorithm with a very limited number of features. Although k-means has been applied for document segmentation previously, the number of features used in these approaches was large, increasing the computational cost. The current methodology relies on only three features.

Although for the segmentation of historical documents with challenging layouts, a number of annotated datasets have been created (Gruning et al, 2018; Simistira et al, 2016), no such dataset exists for illuminated medieval manuscripts. For the current study a dataset was created containing manuscripts with a range of layouts, decorations, and containing a variety of texts (devotional and medical), produced in different regions in different time periods. The images, freely available (Digital Bodleian) derive from the following manuscripts in Oxford's Bodleian Library: MS Canon. Misc 476, MS Add. A. 185, MS Ashmole 1462, MS Auct. 2.2, MS Buchanan e 7.

Fig. 1. Oxford, Bod. Library, MS Canon. Misc. 476. ©Bodleia

Fig. 2. Oxford, Bod. Library, MS Ashmole 1462. ©Bodleian Library

As a pre-processing step, the image is converted into gray format, a uniform filter is then applied using a kernel size of 13 in order to obtain a smoother format. After pre-processing, three features are proposed for clustering.

Once all features are computed and standardized, k-means algorithm is performed over 5 clusters. Additionally, after computing k-means for 120 images belonging to 4 different manuscripts, the centroids of each cluster are calculated and plotted.

This approach uses clustering and filtering techniques for segmenting challenging illuminated medieval manuscripts. Traditional approaches to text segmentation assume that text regions are enclosed in rectangular shapes, which is not true for many illuminated medieval books. Although k-means and filtering have been previously used for this task, the uniqueness of this approach is its reliance on only three features. The strength of the method further lies in its transparency at every step of the process, low-memory use, potential to produce highly refined results, and versatility. This stands as an alternative to programs such as neural networks which are black-boxes, do not allow for querying of their decision-making process, are computationally intensive, and demand manually-annotated training sets. This approach, therefore, provides not only a solution for the segmentation of challenging images with mixed textual and visual content, but more importantly leads towards algorithms with improved robustness, stability and versatility.

References

Ares Oliveira, S., Seguin, B. & Kaplan F. (2018). 'dhSegment: A generic deep-learning approach for document segmentation'. 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR): n. pag. Crossref. Web.

Digital Bodleian. https://digital.bodleian.ox.ac.uk/.

Eskenazi, S., Gomez-Krämer, P., & Ogier, J. (2017). 'A comprehensive survey of mostly textual document segmentation algorithms since 2008'. Pattern Recognition, 64, pp. 1–14.

Grüning, T., Labahn R., Diem M., Kleber F., and Fiel S. (2018). 'Read-bad: a new dataset and evaluation scheme for baseline detection in archival documents'. 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), pp. 351–356.

Shafait, F., Keysers, D., & Breuel T. (2008). 'Performance evaluation and benchmarking of six-page segmentation algorithms'. Pattern Analysis and Machine Intelligence, IEEE Transactions, 30, 6, pp. 941–

954.

Simistira, F., Seuret, M., Eichenberger, N., Garz, A., Liwicki, M., & Ingold, R. (2016). 'Diva-hisdb: a precisely annotated large dataset of challenging medieval manuscripts'. Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference, pp. 471–476.

ID: 111
Short paper presentations
Topics: linguistics, 3D modeling / virtual and augmented reality, audio / video / multimedia, networks / relationships / graphs, software design and development, visualisation, computational science
Keywords: network, lexicography, exhibition, word net, visualization

Óravíddir: Interactive Exhibition about the Icelandic Language*

Trausti Dagsson, Jón Hilmar Jónsson, Eva María Jónsdóttir
The Árni Magnússon Institute for Icelandic Studies, Iceland

This paper describes an interactive exhibition about the vocabulary of the Icelandic language. The exhibition is called Óravíddir – Orðaforðinn í nýju ljósi (e. Vastness – The Vocabulary in a New Light) and was opened at the Culture House in Reykjavík, a part of The National Museum of Iceland in May 2019. The exhibition used data from the word database Íslenskt orðanet (The Icelandic Word Web) and illustrates semantic relations between words in a three-dimensional visualization. The paper introduces Íslenskt orðanet followed by a description on how the data was used to create the network graph visualization. Then we discuss the setup of the exhibition and finally we conclude by reflecting on future possibilities and further development.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 116
Long paper presentations
Topics: historical studies, linguistics, medieval studies, philology, corpus linguistics, cultural heritage collections, data mining / text mining, information retrieval, natural language processing, stylistics and stylometry / authorship attribution
Keywords: text classification, medieval charters, author attribution, dating, Old Swedish

Classification of Medieval Documents: Determining the Issuer, Place of Issue, and Decade for Old Swedish Charters*

Mats Dahllöf

Uppsala University, Sweden

The present study is a comparative exploration of different classification tasks for Swedish medieval charters (transcriptions from the SDHK collection) and different classifier setups. In particular, we explore the identification of the issuer, place of issue, and decade of production. The experiments used features based on lowercased words and character 3- and 4-grams. We evaluated the performance of two learning algorithms: linear discriminant analysis and decision trees. For evaluation, five-fold cross-validation was performed. We report accuracy and macro-averaged F1 score. The validation made use of six labeled subsets of SDHK combining the three tasks with Old Swedish and Latin. Issuer identification for the Latin dataset (595 charters from 12 issuers) reached the highest scores, above 0.9, for the decision tree classifier using word features. The best corresponding accuracy for Old Swedish was 0.81. Place and decade identification produced lower performance scores for both languages. Which classifier design is the best one seems to depend on peculiarities of the dataset and the classification task. The present study does however support the idea that text classification is useful also for medieval documents characterized by extreme spelling variation.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 119
Long paper presentations
Topics: historical studies, cultural heritage collections, digital resources – publication and discovery, natural language processing
Keywords: historical newspaper corpus, OCR, machine learning, post processing

Supervised OCR Post-Correction of Historical Swedish Texts: What Role Does the OCR System Play?*

Dana Dannells[1], Simon Persson[2]

[1]University of Gothenburg, Sweden; [2]Chalmers University of Technology, Sweden

Current approaches for post-correction of OCR errors offer solutions that are tailored to a specific OCR system. This can be problematic if the post-correction method was trained on a specific OCR system but have to be applied on the result of another system. Whereas OCR post-correction of historical text has received much attention lately, the question of what role does the OCR system play for the post-correction method has not been addressed. In this study we explore a dataset of 400 documents of historical Swedish text which has been OCR processed by three state-of-the-art OCR systems: Abbyy Finereader, Tesseract and Ocropus. We examine the OCR results of each system and present a supervised machine learning post-correction method that tries to approach the challenges exhibited by each system. We study the performance of our method by using three evaluation tools: PrimA, Språkbanken evaluation tool and Frontiers Toolkit. Based on the evaluation analysis we discuss the impact each of the OCR systems has on the results of the post-correction method. We report on quantitative and qualitative results showing varying degrees of OCR post-processing complexity that are important to consider when developing an OCR post-correction method.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 173
Short paper presentations
Topics: design, 3D modeling / virtual and augmented reality, project design / organization / management, user studies / user needs, museology
Keywords: museum, social interaction, participatory design, VR, virtual reality

Museums, Technology and Social Interaction in ñAnyone Can Innovate!ò

Gabriella Di Feola, Erik Einebrant, Fredrik Trella

Research Institutes of Sweden, Sweden

The purpose of this paper is to describe insights gained from a collaboration project between RISE, an experimental research institute, and Borås Museum, a local cultural heritage institution, around the topic of how technology can be used in museums to encourage social interaction between visitors and between visitors and the museum staff. This is investigated through a case study of the project “Anyone Can Innovate!”, which was a multi-participatory VR-installation, using a perspective of participatory design. The study was conducted through observations by the developers, formal user testing with externally recruited testers, and by an interview with the responsible project leader and curator from Borås Museum. The VR-installation was tested in two iterations with different levels of embedded guidance, and included different roles for the participants, as an attempt to boost collaboration and interaction. One conclusion of the study is that the use of technology in a museum doesn’t per se mean that it will be participatory, and that it does not necessarily exclude the role of a human guide. In the discussion part, examples are given on how technology can be used as a tool to use participatory design.

ID: 250
Keynote speaker

Responsible Artificial Intelligence

Virginia Dignum

Umeå University, Sweden

The last few years have seen a huge growth in the capabilities and applications of Artificial Intelligence (AI). Hardly a day goes by without news about technological advances and the societal impact of the use of AI. Not only are there large expectations of AI's potential to help to solve many current problems and to support the well-being of all, but also concerns are growing about the impact of AI on society and human wellbeing. Currently, many principles and guidelines have been proposed for trustworthy, ethical or responsible AI.

In this talk, I argue that ensuring responsible AI is more than designing systems whose behavior is aligned with ethical principles and societal values. It is above all, about the way we design them, why we design them, and who is involved in designing them. This requires novel theories and methods that ensure that we put in place the social and technical constructs that ensure that the development and use of AI systems is done responsibly and that their behavior can be trusted.

ID: 117
Long paper presentations
Topics: cultural studies, geography, linguistics, bibliographic studies, corpus linguistics, cultural heritage collections, data mining / text mining, interdisciplinary collaboration, open science, sustainability and preservation, ethnography, computational science
Keywords: Livonian, corpora, place names, geospatial information, mapping, low-resource languages, database, endangered languages

Taking the Livonians into the Digital Space

Valts Ernštreits, Gunta Kļava

University of Latvia Livonian Institute, Latvia

Currently there are approximately 30 individuals who are able to communicate in Livonian (Livones) and approximately 250 individuals gave their ethnicity as “Livonian” on the last Latvian national census (Census 2011); however, the actual number of Livonians is considerably greater. Livonian heritage has had a significant, though understudied, role in the formation of modern Latvia, the Latvian language, and the Latvian nation, and has also been important across the broader Northern European region.

The Livonian community has been able to preserve its identity and also its language up to the present day despite its small size (19th century – 2500, mid-20th century – 1500) and the complicated history of the Livonian speech area. This history included the loss of the last compact Livonian-inhabited territory following the creation of a border zone along the Baltic Sea in the area encompassing the Livonian villages at the start of the Soviet occupation after World War II (Druviete, Kļava 2018, 132). As a result, the Livonians were scattered across Latvia and the world, and this fact continues to pose an added challenge for them. The same is true for various archives relating to Livonian (language, folklore, folk cultural objects, etc.), which for historical reasons were collected and therefore are currently stored at various institutions located in different countries (Ernštreits, 2012).

Though the number of Livonians and Livonian speakers is extremely small, the study of Livonian and related topics requires the same opportunities and tools as those for any

other language. As a result, taking into account all of the aforementioned facts, the main issue faced by the Livonian community, Livonian researchers, and society at large is ease of access to sources and archives relating to Livonian heritage. The rapid digitalization of society during the last decades has created new opportunities for many smaller communities, including the Livonians, for solving these problems and designing new tools ensuring that these materials can be easily accessed and used.

Work on a Livonian language database cluster began in 2016 and currently contains three Livonian language databases – a lexical database, morphological database, and corpus – and consists of interconnected data archives. These databases have already significantly simplified the process for learning about, researching, and studying Livonian, and have created a foundation for future solutions directed towards simplifying access to Livonian-related materials. Though initially this system was created as a Livonian language data archive and a tool for language research, standardization, and acquisition, its principles also can be adjusted to suit other types of studies by supplementing it with other digital archives (containing images, audio recordings, video, 3D scans, data from other databases) as well as other information. A Livonian digital text archive, which is currently being built, will be one of these archives.

In 2019, the UL Livonian Institute submitted a project for expanding the existing Livonian language database cluster. This project takes the next logical step and will establish a mapped open-access Livonian place name database linked to geospatial information. This will open up radically new possibilities for access to various Livonian sources archives, ensure their availability for research as well as data interconnectedness in the future. The need for this project was initially due to practical considerations relating to collecting Livonian place names, so that bilingual road signs could be introduced in the territories historically inhabited by the Livonians.

There is no single Livonian place name archive. Livonian place names primarily are found in lexicographic publications or studies, albeit in a more scattered manner and fewer in number (for example, in the Livonian-Estonian-Latvian Dictionary (LELD), Livonian-German Dictionary (LW), the Livonian-Latvian-Esperanto Dictionary (ČDG), the Salaca Livonian Dictionary (SLW), and other lexicographic sources, in Kersti Boiko's dissertation (Boiko 1993)). Currently, recording a large number of place names through field work is also not possible, as there remain only very few Livonian speakers and due to the historical situation of the Livonians, their connection with the territories inhabited by their ancestors is often indirect and their knowledge of Livonian place names is meagre. In large part, the Livonian language corpus is used instead of informants, and while indexing its contents, it is possible to not only identify Livonian place names, but also gain information regarding their localization.

At the same time there is a source, which is not a compilation of place names itself, but can be used effectively for collecting Livonian place names. This source consists of various collections – lexical card files, object card files and descriptions, folklore collections, etc. The metadata in these collections contain references to the time and method as well as the place where each item was collected and this information is usually shown in Livonian. It is also significant that places mentioned in this metadata are of specific importance to Livonian culture and are located within the historical Livonian territory to be documented and mapped for this project. These places are often not found in any available cartographic product (for example, homesteads which have disappeared or have been moved, drained rivers or lakes, etc.).

Therefore, the mapped place name database developed using metadata from various collections will serve as a starting point for the creation of a one-click database cluster. This will make it possible to link place names, geospatial data, and information data from other fields (e.g., information on consultants, dialect materials, objects, folklore, oral history

collections, etc.), thereby opening up new possibilities for the multifaceted documentation and research of Livonian heritage in the future. Furthermore, this will make it possible to use data for many different purposes beyond research. These could include, for example, digital exhibits or using the cartographic products resulting from this project for education exploring regional heritage, cultural tourism, development of municipal strategies, entrepreneurship, and many other areas.

The products and discoveries resulting from this resource will also be useful to other smaller communities, and the synergy and coordination among various archives can create a rich, high-quality resource suitable for multi-faceted studies in many fields or for interdisciplinary research in general. It will also ensure effective use of data and research results for the preservation, maintenance, and development of any low-resource language or cultural community with limited data, personnel, financing, or other resources.

References

Boiko, Kersti (1993). Baltijas jūras somu ģeogrāfiskie apelatīvi un to relikti Latvijas vietvārdos (Balto-Finnic geographic words and their relicts in toponymy of Latvia). Disertācija filoloģijas doktora grāda iegūšanai. Latvijas Universitāte.

Census 2011 = Centrālā statistikas pārvalde. Tautas skaitīšana (Census), 2011. Available online at https://www.csb.gov.lv/lv/statistika/statistikas-temas/iedzivotaji/tautas-skaitisana/tabulas/tsg11- 06/latvijas-pastavigie-iedzivotaji-pec. Accessed on 15.06.2019.

ČDG = Čače, Ints, Damberg, Pētõr, Grīva, Hilda (1964). Esperantisto en Latvio ce livoj = Esperantist Letmāl līvlist jūsõ. Manuscript. Rīga.

Druviete, Ina, Kļava, Gunta (2018). The role of Livonian in Latvia from a sociolinguistic perspective. Eesti ja soome-ugri keeleteaduse ajakiri = Journal of Estonian and Finno-Ugric Linguistics, Vol. 9, N 2, Livonian studies III, pp. 129–146. 10.12697/jeful.2018.9.2.06

Ernštreits, Valts (2012). Lībiešu valodas situācijas attīstība Latvijā. In I. Druviete (ed.) Valodas situācija Latvijā: 2004–2010. Rīga: Latviešu valodas aģentūra, pp. 142−166.

LELD = Līvõkīel-ēstikīel-leļkīel sõnārõntõz. Liivi-eesti-läti sõnaraamat. Lībiešu-igauņu-latviešu vārdnīca (2012). Tartu, Rīga: Tartu Ülikool, Latviešu valodas aģentūra.

Livones = Lībiešu valoda (Livonian language). Available online at http://www.livones.net/lv/valoda/?libiesu-valoda. Accessed on 15.06.2019.

LW = Joh. Andreas Sjögren's Livisch–deutsches und deutsch–livisches Wörterbuch (1861). St. Petersburg: Kaiserlischen Akademie ded Wissenschaften.

SLW = Winkler, Eberhard, Pajusalu, Karl (2009). Salis-livisches Wörterbuch. Linguistica Uralica. Supplementary Series. Volume 3, Tallinn.Livones = Lībiešu valoda (Livonian language). Available online at http://www.livones.net/lv/valoda/?libiesu-valoda. Accessed on 15.06.2019.

LW = Joh. Andreas Sjögren's Livisch–deutsches und deutsch–livisches Wörterbuch (1861). St. Petersburg: Kaiserlischen Akademie ded Wissenschaften.

SLW = Winkler, Eberhard, Pajusalu, Karl (2009). Salis-livisches Wörterbuch. Linguistica Uralica. Supplementary Series. Volume 3, Tallinn.

ID: 224
Short paper presentations
Topics: library & information science, literary studies, copyright / licensing / Open Access, data modeling / knowledge representation, digitisation – theory and practice, digital resources – publication and discovery, diversity and multilingual / multicultural approaches, infrastructure, information architectures, interface & user experience design, linking and annotation, project design / organization / management, scholarly editing, standards and interoperability, sustainability and preservation, research data archiving
Keywords: digital resources, sustainability, interoperability, FAIR

Inheriting Digital Projects: How to Keep Ibsen Alive Online*

Nina Marie Evensen

University of Oslo, Norway

This paper addresses the challenge of managing digital projects on a long-term scale. In most digital projects there is no strategic plan for the afterlife and maintenance of the project results, leaving them to an uncertain fate. This can be illustrated by the inherited digital resources hosted by the Centre for Ibsen Studies at the University of Oslo, and the challenges they represent when it comes to functionality and maintenance. Due to the rapidly increasing number of digital projects, many institutions will be asking the same questions as we do: How do we keep digital resources alive and up to date in a continuously changing digital reality?

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 184
Short paper presentations
Topics: library & information science, literary studies, interface & user experience design
Keywords: human-centered design, stimulus material, requirements elicitation

Human-Centered Humanities: Using Stimulus Material for Requirements Elicitation in the Design Process of a Digital Archive*

Tamás Fergencs, Dominika Illés, Olga Pilawka, Florian Meier
Aalborg University Copenhagen, Denmark

This study proposes the use of so-called stimulus material during interviews for requirements elicitation as part of the design process of a digital archive. Designing complex systems like digital archives is not straightforward as users of these systems have specific needs and tasks that designers need to be aware of before the implementation phase can begin. Stimulus material can support the requirements elicitation to collect domain and content-specific user tasks and needs, which might get overlooked otherwise. We supplemented semi-structured interviews with observational sessions in which print-outs of historical pamphlets and office supplies were handed to participants to give them the opportunity for in-depth study of the material. We found that the use of stimulus material helps participants to focus on the task at hand and articulate their actions and workflow steps more easily. Via thematic analysis the participants statements were turned into a coding schema that serves as requirements specification for an initial prototype.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 100
Poster
Topics: library & information science, information retrieval, social media, data journalism, digital activism
Keywords: conspiracy theory, birth certificate, archives, administration

Birth Certificate Enslavement ï A Conspiracy from the Archives to the Internet

Rikard Lars Friberg von Sydow
Södertörn University, Sweden

Someone that surfs the Internet today has a chance to meet many conspiracy theories online. If the surfer understands English and visits sites where this is the main language of communication, there are a great possibility that a lot of the conspiracy theories s/he meets originate in the USA and in a North American political context. This should make us eager to investigate these conspiracies, their eventual interoperability to other political and administrative context, and the context in which they are created.

The “Birth Certificate Enslavement” – conspiracy is such a conspiracy, mostly set in an North American context. To Archival scientists, as myself, it is interesting because it both involve what is usually called a vital record – the birth certificates – and the Internet – which is the main area where the conspiracy is spread (USA.gov 2019). Thus connecting two different spheres of information – the administrative records and the new electronic media. Connecting these two spheres is interesting from a couple of positions. One position is the administrative agencies and their staff. How are they viewed by the proponents of the conspiracy theory? As enemies or as useful idiots to an evil mastermind? This is of interest because it might give us insight into possible threats towards the agency and its staff. Conspiracy theories have caused violence before: the gunman who attacked the Comet Ping Pong Restaurant because it was targeted as a participator in the Pizzagate child sex ring being one example (Haag et al 2017). Another interesting position to view conspiracies from is the globality of the internet and the often very specific contextual nature of administrative procedures. Administrative procedures and the

document they create: a birth certificate for instance, are often deeply connected to national law and may vary a lot between different states and regions. What do this component add to the conspiracy theories. Are they pointed towards a smaller group of people because of the different administrative circumstances these people live under, or are these differences ignored by the proponents of conspiracy theories? Are they even aware of these administrative differences

The conspiracy of birth certificate enslavement is connected to, among others, a small but rather violent group: the Sovereign Citizen-movement (SPLC 2019), and has through this connection been observed by organizations that monitor violent extremists, like The Southern Poverty Law Center (SPLC). SPLC has described the conspiracy regarding birth certificates as this. According to the believers, the creating of a birth certificate starts the life of an evil administrative doppelgänger to the newly born. A “corporate shell”. This corporate shell is then sold by the Federal Reserve to foreign inverstors as a form of financial security. One of the administrative proof of this, according to the believers of the conspiracy theory, is that the birth certificate is written in capital letters and that bond papers and watermarks are used. This part of the conspiracy, the “proof in the design of the administrative document” is connected to an interpretation of Admiralty law that the sympathisers believe are valid regarding birth certificate (SPLC 2010)

The research will be accomplished in the following way. I will analyze four Youtube videos that aim to explain the Birth Certificate Enslavement-belief from a sympathisers perspective. The videos will be chosen through popularity and the four most viewed videos regarding the subject on Youtube.com will be analyzed. The videos clips will be analysed with the help of three research questions (Research question 1–3). I believe that four videos are an appropriate amount of videos and that it would create the possibility to find differences between various proponents of the conspiracy. At least as introductory research adequate for a poster presentation. Choosing the most popular videos is a way of

reflecting what would be seen by someone casually surfing the web, trying to find explanations of how the world works according to different world views.

The research questions are formulated as a heading with different subquestions.

R1) References done in the video clip to public administration and archives.

How are public administration and archives viewed in the message of the video clips? How are the employees of the agencies responsible for creating, storing and administrating birth certificates viewed. Are they in any way seen as recipients of the message in the video clips? Are there any appeals to confront the agencies or their employees.

R2) References to other conspiracy theories.

Are there any references to other conspiracy theories? Do they differ from video to video depending on the creator? One of the differences between various conspiracy theorists are that they put different “masterminds” behind the execution of the conspiracy. Aliens, Freemasons, Jews, Reptiles, the Catholic church, the Red shoe men and the Bilderberg group et cetera.

R3) References to other political and administrative contexts than those connected to the United States of America.

Are there any references to other political and administrative contexts than those of the United States of America? Do the proponents of the conspiracy theory believe that the theory interoperates with other political and administrative contexts or are they unique to an United States context?

Possible this research will give us a better insight into how the proponents conspiracy theories argue for their world view in connection to actually occurring administrative procedures. How they connect this particular theory of birth certificate enslavement to other conspiracy theories, arguing for specific masterminds that are working behind the scenes. And hopefully also how they contextualise this theory in reference to it being spread on the global internet. Do they stick to explaining it from an American perspective, or is the theory

decontextualised to interoperate with other political and administrative contexts? All questions which answers would be valuable to our understanding of fake news and conspiracies on the internet in general.

References

Haag, Matthew and Salam, Maya (2017) “Gunman in ‘Pizzagate’ Shooting Is Sentenced to 4 Years in Prison” New York Times. Viewed 2019-09-05

SPLC (2010) “The Sovereigns: A dictionary of the peculiar”, viewed 2019-09-05

SPLC (2019) “Sovereign Citizens Movement”, viewed 2019-09-05.

USA.gov - “Replace your vital records”, viewed 2019-09-01

ID: 222
Short paper presentations
Topics: linguistics, literary studies, corpus linguistics, cultural heritage collections, encoding – theory and practice, interdisciplinary collaboration
Keywords: morphological tagging, evaluation, ego-documents, literary research

Automatic Morphological Annotation of Ego-Documents: Evaluating Automatically Disambiguated Annotation of Estonian Semper-Barbarus Correspondence Corpus

Olga Gerassimenko[1], Kadri Vider[1], Neeme Kahusk[1], Marin Laak[2], Kaarel Veskis[2]

[1]University of Tartu, Center of Estonian Language Resources, Estonia; [2]Estonian Literary Museum, Estonia

The digitization of the cultural heritage is massive in Estonia: the national programme of mass digitisation started in 2018, and the creation of digital heritage resources is made a priority for Estonian memory institutions (Viires, Laak 2018). Yet, the majority of the digitised literary data is captured and used in the raw format: at best, the digitised source is transformed to the plain text that is searchable for strings. The digital resources are mostly used by the humanity scholars in the same way as the published texts: digital texts are read at length and analysed qualitatively. The quantitative methods, even as simple as word frequency analysis, are not possible for unannotated texts. The morphological annotation and disambiguation is an undisputed necessity for the digitized data, especially considering the rich morphology of Estonian and the great amount of homoforms. Big amounts of data need to be parsed and disambiguated automatically that implies some error rate but still makes corpus search, data analysis and data mining much more efficient.

There are many challenges for morphological tagging of older cultural heritage sources (especially non-edited ego-documents such as letters and diaries). The authomatical morphological parser and disambiguator of Estonian ESTMORF has been created for contemporary Estonian and trained on the texts of second half of the 20th century, proving

to be 99% efficient on the contemporary published texts (Kaalep, Vaino 2001). The efficiency of parcer has been tested on less normative texts types such as chatroom texts (Kaalep, Muischnek 2011), but ego-documents offer some specific complications: sentences are lengthy and syntactically complicated, and yet letters and diary entries may include ad hoc abbreviations, unmarked switching to other languages, specific orthography and punctuation. Is the automatic morphological annotation of such texts reliable enough for a decent corpus research and for comparison of the target sources with other corpora?

We are exploring it on the data of the Correspondence Corpus of Estonian avant-garde poets and writers Johannes Semper and Johannes Barbarus in 1911–1940 (Laak et al. 2019). This is a unique and multidimensional collection of private letters, the hand-written originals of which are held at the Estonian Cultural History Archives of the Estonian Literary Museum in Tartu. The correspondence consists of 670 letters with more than 1,100 pages and more than 310 890 tokens (249 970 words). The range of subjects touched upon in the letters is extremely wide: Semper and Barbarus as friends and colleagues discuss all events in the Estonian cultural life, organize the publication of their books and discuss the problems of their contemporary literary and political life and even economics in Estonia and in other countries. The letters were already transformed to typewritten and then to electronic format; morphological categories had to be automatically annotated and disambiguated in them and metadata had to be described manually to transform the letters to a machine-readable format. Corpus is openly accessible through KORP query system and is currently being used by the literary scholars for textual search.

In order to evaluate the quality of the morphological analysis of the Semper-Barbarus Correspondence Corpus, we are manually checking certain excerpts of the output and computating the error rate in general and the error rate for Estonian text only (there are lengthy foreign-language excerpts in the Semper-Barbarus correspondence). We are

going to calculate and compare the error rate to the error rate of the texts of the same time period (1920–1930) from Estonian Literary Criticism Corpus containing published publicistic texts and to compare it to the previous work of Liba and Veskis (2008) on Estonian automatic tagger evaluation.

The results of the study are going to be used to propose the systematic modifications of the morfological parser by manually adding words to the parser lexicon. The reliably annotated corpus can be used for quantitative research of phenomena mentioned in texts: for instance, we can evaluate the relative frequency of the words related to politics in the various decades of correspondence. Having a reliable automatic morphological annotation, we can annotate texts syntactically and semantically, and, in perspective, apply sentiment analysis to see whether the affective polarity of texts changes with time.

Acknowledgements

Research supported by the institutional research grant “Formal and Informal Networks of Literature, Based on Sources of Cultural History” (IRG22-2, Estonian Ministry of Education and Research), related to the Centre of Excellence in Estonian Studies (CEES) and by the programme ASTRA (2014–2020.4.01.16–0026) via the European Regional Development Fund (TK145).

Development of KORP and adding corpora in Estonian is supported by the ERDF project “Federated Content Search for the Center of Estonian Language Resources” (2014–2020.4.01.16–0134) under the activity “Support for Research Infrastructures of National Importance, Roadmap”.

References

Barbarus-Semper Correspondence Corpus, https://doi.org/10.15155/9-00-0000-0000-0000-00190L, last accessed 2019-09-14.

Estonian Literary Criticism Corpus, https://doi.org/10.15155/9-00-0000-0000-0000-00193L, last accessed 2019-09-14.

Kaalep, Heiki-Jaan and Tarmo Vaino. 2001. Complete morphological analysis in the linguist’s toolbox. In Proceedings of Congressus Nonus Internationalis Fenno-Ugristarum, Pars V. 9–16.

Kaalep, H.-J.; Muischnek, K. (2011). Morphological analysis of a non-standard language variety. Proceedings of the 18th Nordic Conference of Computational Linguistics: NODALIDA 18, Riia, Läti, 11-13. mai 2011. Ed. Bolette Sandford Pedersen, Gunta Nešpore, Inguna Skadina. Riia, Läti, 130−137. (NEALT Proceedings Series; 11).

Laak, Marin; Veskis, Kaarel; Gerassimenko, Olga; Kahusk, Neeme; Vider, Kadri (2019). Literary Studies Meet Corpus Linguistics: Estonian Pilot Project of Private Letters in KORP. DHN 2019 Digital Humanities in the Nordic Countries, 2364: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference Copenhagen, Denmark, March 5-8, 2019.. Ed. Costanza Navarretta, Manex Agirrezabal, Bente Maegaard. Copenhagen, Denmark: University of Copenhagen, Faculty of Humanities, 283−294.

Viires, P., Laak, M.: Digital humanities meet literary studies: Chal-

lenges facing estonian scholarship. In: Mkel, E., Tolonen, M., Tuomi-

nen, J. (eds.) DHN Helsinki 2018. Book of Abstracts (2018), https://www.helsinki.fi/sites/default/files/atoms/files/dhn2018-book-of-

abstracts.pdf, last accessed 2019-09-14.

Veskis, K., Liba, E.: Automatic tagger evaluation. NLP course assignment report

(2008), https://entu.keeleressursid.ee/public-document/entity-7052, last accessed 2019-09-14.

ID: 252
Keynote speaker

From Cow Sheds to Computer Screens: Some Thoughts on the Uses of Digital Humanities for the Study of Folkloristics in Iceland and the Other Nordic Countries

Terry Gunnell

University of Iceland

In this lecture, I will be comparing the state of the art in Nordic Folkloristics as it was when I entered the field in the nineties to the present state of affairs, underlining among other things the role of digitisation in not only saving but opening up the archives and understanding their contents. The focus will be on several projects, at the heart of which will be the Icelandic folk legend database, Sagnagrunnur, which has been introduced at earlier conferences, and which is now both integrating and being integrated with several other local and international projects. Far from being an expert on computers myself, I will underline the reasons behind why in the early nineties we decided to make a database of all Icelandic folk legends in print as a means of not only making the printed collections more accessible to researchers, but also of opening up new ways of interpreting the material contained within them, not least reconnecting them with the countryside from which they sprang, and the people who told them. (A key problem in Folkloristic research in the past was related to the way physical card indexes in the folklore archives often restricted searches rather than opening them up.) In relation to this, I will discuss how and why in the Nordic countries we have now returned to mapping (something many folklorists had become wary about by the turn of the century), and how and why we expanded this particular project to bring in manuscript materials and a wide range of other contextual materials, ever expanding the borders to allow us to connect and communicate with other Nordic databases such as those at present being developed by Timothy Tangherlini and Fredrik Skott in the US and Sweden, and Theo Meder in Holland. Some discussion will

also be made of our work in Iceland on making our survey of national folk belief more interactive, and of two other interdisciplinary databases we have created in recent years relating to the “creation” of “national culture” (in close relation to the Joep Leerssen’s wide-ranging on-line Encyclopedia of National Romanticism [E. R. N. I. E.] in Amsterdam), databases which are now being connected in various ways to Sagnagrunnur and other Nordic projects, something that has led to a wider international project charting the development of what we call “The Grimm Ripples”. Here we have a prime example of how increased interconnectivity effectively opens up new means of understanding the local (which is very much at the heart of Folkloristics). I will end by discussing briefly how we see things developing in the future (including dreams of connection to the sound archives), and those problems that might hinder such development.

ID: 159
Long paper presentations
Topics: library & information science, cultural heritage collections, data mining / text mining, digital resources – publication and discovery, GLAM: galleries / libraries / archives / museums, user studies / user needs, big data, web research, archiving
Keywords: web archiving, digital humanities, library labs, collections as data, research use of web archives

Supporting Research Use of WEB Archives: A óLabsô Approach

Olga Holownia[1], Sally Chambers[2]

[1]IIPC, British Library, UK; [2]Ghent Centre for Digital Humanities, Ghent University, Belgium

The use of the archived web as an object of research remains at the fringes of (digital) humanities research (Winters, 2017). While a number of surveys and studies have identified common challenges and researchers' requirements (See e.g. Costa & Silva, 2010; Costea, 2018; Riley & Crookston, 2015; Stirling, Chevallier, & Illien, 2012), the conclusion saying that "there is still a gap between the potential community of researchers who have good reason to engage with creating, using, analysing and sharing web archives, and the actual (generally still small) community of researchers currently doing so" (Dougherty et al., 2010, p. 5) largely holds true. In our paper we argue that Library Labs – a growing network of experimental environments which provide data-level access to digitised and born-digital collections – can help bridge that gap.

Research Use of Web Archives

Although many researchers in the humanities and social sciences still need to begin to explore the web archives, some projects have already investigated their potential. Mapping the Danish Web (Brügger & Laursen, 2018; Brügger, Laursen, & Nielsen, 2019), Big UK Domain Data for the Arts and Humanities (BUDDAH) project (Hockx-Yu, 2011; Winters, 2015)[2], text-mining projects such as Néonaute (Cartier, Stirling, & Aubry, 2018) and Semantic Change Detection

[2] Further information about the BUDDAH project is available at https://buddah.projects.history.ac.uk

(McGillivray & Basile, 2018), the research being undertaken by members of the RESAW network (Research Infrastructure for the Study of Archived Web Materials)[2][3] and PROMISE (PReserving Online Multiple Information: towards a Belgian StratEgy)[4] (Geeraert, Michel, & Vlassenroot, 2018; Vlassenroot et al., 2019) being particular examples. The Internet Archive Research Services have provided important use cases that expand beyond national domains while the Archives Unleashed Project[5] has focused on developing a toolkit, a cloud service to work with WARC files and a community around their regular datathons.

Access and Labs as ñIncubators for Researchò

As a result of legal restrictions, many web archives still remain solely accessible through dedicated computers inside (national) libraries. Additionally, managing archived web-resources as large, complex and messy datasets, requires a relatively advanced level of digital literacy, not always at the fingertips of all humanities researchers. In this paper, we will consider whether the concept of 'library labs', as pioneered by organisations such as the British Library, and more recently, exemplified through the international Building Library Labs network[6] (Chambers et al., 2019) could be a) an ideal incubator for both increasing access to archived-web resources, such as within national library buildings themselves and b) whether the inclusion of web-archives as one of the many available resources alongside e.g. digitised newspapers, etc. could increase their take-up and usage in the humanities and social sciences research community. We will also examine case studies from national and university libraries that have experimented with offering datasets from their web archives as part of labs or research services (e.g. Library of Congress,

[3] Further information about the RESAW project is available at www.resaw.eu
[4] Further information about the PROMISE project is available at https://promise.hypotheses.org
[5] Further information about the Archives Unleashed project is available at https://archivesunleashed.org
[6] Further information about the Building Library Labs Network is available at: https://blogs.bl.uk/digital-scholarship/2018/09/building-library-labs-around-the-world.html

Royal Danish Library, Austrian National Library and British Library). Furthermore, the recently established Research Working Group of the International Internet Preservation Consortium (IIPC), which a) seeks to promote the use of web archives and IIPC collections among researchers, b) share information about web archiving research projects at IIPC member organisations, including workflows and lessons learnt, and c) facilitate ways for dissemination and discussion of use cases, which could be an ideal framework for fostering research-use of archived web material,[7] will be introduced.

References

Brügger, N., & Laursen, D. (2018). Historical Studies of National Web Domains. In N. Brügger & I. Milligan (Eds.), The SAGE Handbook of Web History (1. ed., pp. 413–427). London: SAGE Publications.

Brügger, N., Laursen, D., & Nielsen, J. (2019). Establishing a corpus of the archived web: the case of the Danish web from 2005 to 2015. In N. Brügger & D. Laursen (Eds.), The historical web and Digital Humanities: The case of national web domains (pp. 124–142). Abingdon: Routledge.

Cartier, E., Stirling, P., & Aubry, S. (2018). Néonaute: mining web archives for linguistic analysis. Paper presented at the IIPC Web Archiving Conference, Wellington.

Chambers, S., Mahey, M., Gasser, K., Dobreva-McPherson, M., Kokegei, K., Potter, A, Ferriter, M. and Osman, R. (2019). Growing an international Cultural Heritage Labs community. Retrieved from http://doi.org/10.5281/zenodo.3271382

Costa, M., & Silva, M. J. (2010). Understanding the Information Needs of Web Archive Users. Retrieved from http://xldb.di.fc.ul.pt/xldb/publications/costa2010understandingneeds_document.pdf

Costea, M.-D. (2018). Report on the Scholarly Use of Web Archives. Retrieved from http://netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf

Dougherty, M., Meyer, E. T., McCarthy Madsen, C., van den Heuvel, C., Thomas, A., & Wyatt, S. (2010). Researcher Engagement with Web Archives: State of the Art. Retrieved from https://ssrn.com/abstract=1714997

[7] Further information is available at http://netpreserve.org/about-us/working-groups/research-working-group/

Geeraert, F., Michel, A. & Vlassenroot, E. (2018). Critical reflections on unlocking web archives for humanities research. Paper presented at the 5th DH Benelux Conference.

Hockx-Yu, H. (2011). Up close and personal - Researchers and the UK Web Archive Project. Paper presented at the IIPC Web Archiving Conference, The Hague. https://web.archive.org/web/20120501064731/http:/netpreserve.org/events/Hague/Presentations/Out%20of%20the%20Box/Researchers_HockxYu.pdf

McGillivray, B., & Basile, P. (2018). Exploiting the Web for Semantic Change Detection. Paper presented at the 21st International Conference, DS 2018, Limassol, Cyprus.

Riley, H., & Crookston, M. (2015). Awareness and Use of the New Zealand Web Archive: A Survey of New Zealand Academics. Retrieved from https://natlib.govt.nz/files/webarchive/nzwebarchive-awarenessanduse.pdf

Stirling, P., Chevallier, P., & Illien, G. (2012). Web Archives for Researchers: Representations, Expectations and Potential Uses. D-Lib Magazine, 18(3/4). doi:10.1045/march2012-stirling

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. International Journal of Digital Humanities, 1(1), 85–111. doi:10.1007/s42803-019-00007-7

Winters, J. (2015). Big UK Domain Data for the Arts and Humanities. Paper presented at the IIPC Web Archiving Conference, Stanford. https://web.archive.org/web/20170315123348/http:/netpreserve.org/sites/default/files/attachments/2015_IIPC-GA_Slides_07_Winters.ppt

Winters, J. (2017). Coda: Web archives for humanities research: some reflections. In N. Brügger & R. Schroeder (Eds.), The Web as History: Using Web Archives to Understand the Past and Present (pp. 238–248). UCL Press: London.

ID: 211
Long paper presentations
Topics: anthropology, classical studies, cultural studies, gender studies, historical studies, law, library & information science, linguistics, philosophy, data mining / text mining, digitisation – theory and practice, digital resources – publication and discovery, history and theory of digital humanities, interdisciplinary collaboration, interface & user experience design, natural language processing, software design and development, teaching / pedagogy / curriculum design, user studies / user needs, political science, religious studies, computational science, big data
Keywords: partnerships, collaboration, teaching, research, archives

Evaluating a DH Tool. The First 18 Months of the Gale Digital Scholar Lab and the Future of Academic/Corporate Partnerships

Christopher Michael Houghton

Gale, A Cengage Company, USA

This paper will discuss lessons learned in the first 18 months of release of Gale Digital Scholar Lab; a ground-breaking tool designed to make digital scholarship methods more accessible and vastly reduce the time needed to run digital humanities projects. By taking a global view of the users of the Lab, this paper will illustrate regional trends in use, as well as highlighting the key lessons from researchers and academics around the world.

Since 2011, Gale have been working with academics globally to provide access to the OCR and metadata of its world-famous digital archives, including ECCO (Eighteenth Century Collections Online) and the Times Digital Archive.

In 2014, following the decision to make this data available more formally on drives, Gale has kept in touch with many of the researchers in receipt of this data to understand their projects and ideally, the challenges they face in using this data in digital humanities projects.

As a result of this research, Gale identified three common challenges faced by researchers around the world when conducting digital humanities projects. Firstly, the time taken to bring together a significant corpus, clean it and prepare it for analysis often stretch to many months and proved to be

prohibitive for many researchers. Secondly, hosting data was an expensive and labour-intensive process, requiring significant institutional infrastructure that proved to be an obstacle for many. Finally, learning the coding languages necessary to create analytical tools was a challenge for many, especially when considered in the framework of the undergraduate classroom.

Subsequently, Gale began building a tool to meet and ideally, mitigate these challenges. Creating a tool that would be as useful to researchers in Beijing as those in Birmingham proved to be a significant undertaking, and took over four years of development, including one year of active development, at a cost of $2 Million.

In September 2018, we released Gale Digital Scholar Lab, a cloud-hosted text and data mining environment, bringing up to 166 Million pages (to date) of Gale's leading digital archives together with powerful text mining and natural language processing tools. With an aim of drastically reducing the time needed to construct a research corpus, clean large sets of data, customise and run analyses and teach sophisticated digital scholarship methods, Gale Digital Scholar Lab proved to be an extremely popular product.

The launch of the Lab proved a significant evolution in Gale's relationship with academia, as we found ourselves more frequently partnering with academics on projects related to digital humanities. One area of common focus involved collaborating on pedagogies and working together to construct curricula to widen the teaching of digital humanities, with a specific focus on the undergraduate classroom. Increasingly, institutions around the world looked to Gale to assist them in using the Gale Digital Scholar Lab to teach digital methods to humanities students. Not wishing to insert ourselves unnecessarily into the academic process, this proved a great opportunity to collaborate with leading institutions on methods of using the Lab, as part of a suite of tools and techniques, to spread digital humanities methods throughout the HSS department. To this end, Gale began employing academics to collaborate with institutions on creating curricula and teaching.

Alongside this, there proved to be significant and frequent opportunities to partner with academics to create open tools that could be adapted for inclusion into the Lab. This not only allowed us to support valuable research, but also to ensure that tools were created that allowed all users of Gale digital archives to make discoveries and explore them in new and potentially interesting ways.

Working in these new, collaborative ways with academics proved to be both stimulating and challenging for those of us at Gale. It has been particularly noteworthy that the rise in Gale data being used in digital humanities has caused us to ask questions of OCR, metadata, structure, provenance and framing of archives. There is no question that digital humanities has asked challenged the way in which we present archival material and has changed the way we think about putting archives together and presenting them for research.

This paper will break down the first 18 months' usage of the Lab globally, highlighting regional trends and tendencies. Allied to this, the paper will discuss the most common requests for future development and explain Gale's ongoing commitment to evolving the Lab to meet the needs of the global DH community by presenting the development roadmap. By discussing the various partnerships and collaborations, we will show Gale's commitment to growing and amplifying digital humanities research and supporting the values of openness, breaking down barriers and furthering the cause of humanities and social science research.

ID: 123
Long paper presentations
Topics: cultural heritage collections, standards and interoperability, computational science, artificial intelligence, big data
Keywords: semantic portal, linked data, knowledge discovery, artificial intelligence

ñSampoò Model and Semantic Portals for Digital Humanities on the Semantic Web*

Eero Hyvönen

University of Helsinki (HELDIG) and Aalto University, Finland

This paper presents the vision and longstanding work in Finland on creating a national Cultural Heritage ontology infrastructure and semantic portals based on Linked Data on the SemanticWeb. In particular, the “Sampo” series of semantic portals is considered, including CultureSampo (2009), TravelSampo (2011), BookSampo (2011), WarSampo (2015), BiographySampo (2018), Name-Sampo (2019), WarWictimSampo (2019), FindSampo (2019), and LawSampo (2020).

They all share the “Sampo model” for publishing Cultural Heritage content the Semantic Web that involves three components: 1) A model for harmonizing, aggregating, and publishing heterogeneous, distributed contents based on a shared ontology infrastructure. 2) An approach to interface design, where the data can be accessed independently from multiple application perspectives, while the data resides in a single SPARQL endpoint. 3) A two-step model for accessing and analyzing the data where the focus of interest is first filtered out using faceted semantic search, and then visualized or analyzed by ready-to-use Digital Humanities tools of the portal.

This model has been proven useful in practise: Sampo portals have attracted lots users from tens of thousands to millions depending the Sampo. It is argued that the next step ahead could be portals for serendipitous knowledge discovery where

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

the tools, based on Artificial Intelligence techniques, are able to find automatically serendipitous, “interesting” phenomena and research questions in the data, and even solve problems with explanations.

ID: 124
Short paper presentations
Topics: digital resources – publication and discovery, infrastructure, linked data / semantic web / ontologies, open data, open science, computational science, big data, research data archiving
Keywords: linked data, data services, ontology services, semantic web

Linked Open Data Infrastructure for Digital Humanities in Finland*

Eero Hyvönen

University of Helsinki (HELDIG) and Aalto University, Finland

This paper presents and overviews “Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH)”, a joint initiative of Aalto University, Department of Computer Science, and University of Helsinki (UH), HELDIG Centre for Digital Humanities, for creating a centralized national data infrastructure and Linked Data services for open science. The services enable publication and utilization of datasets for data-intensive Digital Humanities (DH) research in structured, standardized formats via open interfaces. LODI4DH is based on a large national collaboration network and software created during a long line of national projects in DH between UH and Aalto since 2002 that created several in-use infrastructure prototypes, such as the ONKI and Finto ontology service now at the National Library of Finland, the Linked Data Finland platform LDF.fi, and the “Sampo” series of semantic portals testing and demonstrating the usability of the approach. Thus far, these systems have had millions of end-users on the Web suggesting a high potential of utilizing the technology and Linked Data infrastructure.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 156
Short paper presentations
Topics: anthropology, communication studies, cultural studies, sociology, social media, ethnography, artificial intelligence, big data, web research, archiving
Keywords: transnational migration, transnationalism online, transhumanism, social network analysis, artificial intelligence, networks of expertise

Studying Transnational Digital Spaces: Methodological Vistas and Challenges

Anastasia A. Ivanova

Saint Petersburg University, Saint Petersburg, Russian Federation

The discussion of transnationalism and transnational migration as well as approaches to studying newly emerged types of migration activities have already become quite conventional and common in social sciences and migration studies in particular. Yet with the development of information and communication technologies (ICT) as well as with the rising penetration of the Internet among different social strata social scientists not only have encountered the promising proliferation of digital data archives (e.g., the studies of the “social life of methods”, see [3]) but also can observe newly emerging transnational digital spaces. Today the network space not only upholds but also transforms the essence of transnationalism leading to a variety of online communicative channels, transnational digital networks, Internet-based transnational identities, communities, diasporas and sororities of compatriots. In this respect, the world wide web (WWW) maintains existing transnational ties, and also actively participates in their formation by involving additional actors, increasing complexity of social ties and causing changes in identities. Thus, a new phenomenon is emerging that can be reasonably called “extended digital transnationalism” or simply “transnationalism online” [4]. Besides, the WWW is a space where transnational processes can be visualized and documented in their most visible form by means of contemporary methods like SNA, Social Network Analysis, that allow to literally “map” transnational digital networks.

This paper presents and compares two on-going studies of transnational digital spaces that apply partly similar, partly different research methods to analysis of different empirical phenomena: transnational migration and transhumanist network. We would outline research questions and research tools of each of the studies and then presents reflections of vistas and challenges for doing social science research online.

Studying networked links allows to reveal transnational processes as they occur online, and also to shed light on the question of how digital interactions redefine transnational migration. In scholarly literature transnationalism is seen predominantly through optimistic lenses; we argue, however, that it can also have a 'dark side'. One of the 'dark' features of transnational migrants' digital spaces is that transnationalism can potentially lead to extremism and radicalization across social networks. Today, Online space is one of the most important sources of information about extremism and one of the most important tools to keep in touch and coordinate actions in extremist organizations. In this regard, the study of assess to Internet, use of gadgets, features of the network behavior of various groups, socio-demographic profile of users, etc. helps answering the question of how ideologies spread. By creating a corpus of websites related to migration and migration processes with the use of SNA (with the major interest in the measures of centrality and the number of landings from one website to another), one can reveal the infrastructure of transnational migrants spaces wherein ideological radicalization takes place.

The analysis of the network trajectories of those who become extremists has its own value, which is not limited to supplementation of traditional studies of extremism. This research perspective takes on particular importance when analyzing transnational online communities, since the online milieu produces a special type of transnationalism – "here, there and somewhere else" / transnationalism online. Three conceptual statements could be made here. First, online communication based on remote interactions produces less social and emotional involvement than face-to-face

communication. Therefore, for ideological radicalization online, it is necessary to have strong, intentionally created emotional triggers and / or constantly repeated interactions. Moreover, an increase in ideological radicalization is more likely if the topics of online interactions are also discussed in face-to-face interactions. Second, in online interactions, artificial intelligence (AI) agents often act as a part of their environment and as participants. Sorting and search AI algorithms channel users' network trajectories. These algorithms could be used to "push" to consume content that promotes radicalization. Third, being online is an experience of passages from one page to another. Therefore, typical and deviating network trajectories are of special interest for studying ideological radicalization.

While in case of 'transnationalism online' we are interested in net mechanisms of involvement in extremism, and thus in the infrastructure of transnational online spaces, in case of transhumanism we are interested in the infrastructure of expert statements 'travel' [2] and particular rhetorical means of persuasion aimed at enlargening transhumanist network. Although current social science literature mostly considers transnationalism in terms of migration and related issues, we suggest that in order to grasp newly emerging social configurations this concept could be extended to an inquiry into social epistemology and social processes of knowledge production, legitimation, and dissemination. Due to transformation of these processes, new domains of knowledge and new epistemic agents emerge, and, the notions of an 'expert' and 'expertise' are becoming blurred demanding reconceptualization.

One of the 'hot' issues that exemplifies these tendencies is scientific immortalism – the edifice based on the endeavor to achieve biological immortality basing on several findings and developments from natural, technical and interdisciplinary sciences, which is increasingly gaining popularity in both Europe and the USA. Whilst this movement can be regarded as 'accelerator' for particular scientific fields, it has also become the large investment pool. Moreover, being a part of a wider intellectual movement (transhumanism or H+) it is aimed

at radical transformation of boundaries between humans and non-humans. They seek political and social support for these purposes, and one of the core premises of its gaining is particular techniques and means of persuasion. In terms of actor-network theory, immortalism is a growing heterogeneous network of knowledge, and persuasion is one of the most important instruments of the network enlargement.

The particular features of newly emerging networks of expertise is that 1) they are more heterogeneous than a community of experts themselves; 2) operating in a digital space is one of the main conditions of possibility for such networks to uphold and enlarge themselves. At the same time such networks are essentially (or in some cases potentially) transnational since a particular set of expert statements crucial for such networks are circulating above boundaries of nation states.

In this case in order to grasp newly emerging configurations of expertise as well as the conditions of possibility for their deployment one might shift the focus from the question of jurisdiction over knowledge production and particular social types of intervening actors to the movement by which knowledge and expertise are mobilized to inform a value-laden intervention in the public sphere [1]. We argue that the crucial component of the conditions of possibility of such a movement is the particular processes of adoption and dissemination of expert statements, which have become possible by the transformation of the statements' "traveling infrastructure".

Here a researcher might rather focus on understanding the particular techniques and means of persuasion within digital area aimed at enlarging the immortalism network. The major part of attempts at enlarging and strengthening of this network can be seen in digital space. Digital technology provides the basis for transformation of interaction and persuasion in particular. Second, sociological literature on expert persuasion (especially in the actor-network fashion, which follows Latour's arguments) basically concerns investigation of how scientific statements should pass in order to transform from a hypothesis to a matter of fact; however, the basic feature of immortalism

is that their edifice cannot be based on any evidence (strictly speaking, everyone has died so far). Thus, the ways of operating with epistemic statements is radically different in this case: immortalists are attempting at problematization and contestation of unproblematic fact-like statements (e.g., “biological death is inevitable”). Third, their rhetorical strategies are performative in terms of both redefinition and blurring the boundaries between human and non-human actors, as their means of persuasion involve appealing to certain types of cyborgs.

What could comparison of the two cases of studying transnational digital spaces tell us about doing research online? On the one hand, conceptually these two cases differ in terms of understanding of what a network and its infrastructure are. On the other hand, in terms of methods they have more in common on the stage of content analysis and discourse analysis. The fundamental task while studying particular means of persuasion within transhumanist network is to determine the core set of expert statements and the work involved in its production. In case of transnationalism online, in order to reveal extremist and terrorist potential of the nodes in the network, the second step after having outlined the network and its dynamics is to conduct content and discourse analysis. Here a statement can also be a unit of analysis, so that further classification can be developed based on differences in the sets of statements and their modalities. These two cases demonstrate how digital space becomes the major condition for new social configurations to emerge and develop – and a very important source of data for social sciences. It also poses new challenges in data collection and analysis. How to formalize processes of data collection – for texts, images, sentiments? What is the balance between anonymity, privacy and scholarly curiosity in WWW? How to grasp diverse contexts of online interactions? How to classify differences in language use in Internet? How to trace labors of AI algorithms and their influences on users? And who are the users?

Therefore, digital space does not prolong and extrapolate a pre-online logic of previously existing social phenomena but

essentially transforms them. Hence mere extrapolating of pre-online methodology and conceptualizations on studying substantially digital phenomena seems to cease being insightful and fruitful. The analysis of patterns and trajectories in the world wide web constitute a world of social relationships and interactions, links and transitions, that has its own causal power and thus deserves careful study.

Acknowledgement

The research is supported by Russian Science Foundation, project No. 18-78-10049.

References

1. Eyal, G. For a Sociology of Expertise: The Social Origins of the Autism Epidemic. Ameri- can Journal of Sociology, 118(4), 863–907 (2013).

2. Morgan, M. S. Travelling Facts // How Well do Facts Travel: The Dissemination of Relia- ble Knowledge / P. Howlett and M.S. Morgan (eds.). Cambridge, UK: Cambridge University Press (2011).

3. Savage, M. The 'Social Life of Methods': Critical Introduction. Theory, Culture & Society, 30 (4), 3–21 (2013).

4. Starikov, V. S., Ivanova, A. A., Nee, M. L. Transnationalism online: exploring migration processes with large data sets. Monitoring of Public Opinion: Economic and Social Changes, No 5, 213–232 (2018)

ID: 244
Poster
Topics: communication studies, interdisciplinary collaboration, linking and annotation, natural language processing, project design / organization / management, visualisation, political science, computational science
Keywords: workflow, close reading, automated annotation, news media, language technology

A Workflow for Integrating Close Reading and Automated Text Annotation

Maciej Janicki[1], Eetu Mäkelä[1], Anu Koivunen[2], Antti Kanner[1], Auli Harju[2], Julius Hokkanen[2], Olli Seuri[3]

[1]University of Helsinki, Department of Digital Humanities, Finland; [2]University of Tampere, Faculty of Social Sciences, Finland; [3]University of Tampere, Faculty of Information Technology and Communication Sciences, Finland

Motivation

Digital Humanities projects often involve application of language technology or machine learning methods in order to identify phenomena of interest in large collections of text. However, in order to maintain credibility for humanities and social sciences, the results gained this way need to be interpretable and investigable and cannot be detached from the more traditional methodologies, which rely on close reading and real text comprehension by domain experts. The bridging of those two approaches with suitable tools and data formats, in a way that allows a flow of information in both directions, often presents a practical challenge.

In this poster, we present an approach to digital humanities research that allows combining computational analysis with the knowledge of domain experts in all steps of the process, from the development of computational indicators to final analysis.

Our approach rests on three pillars. The first of these is an interface for close reading, but crucially one which is able to highlight to the user all results from automated computational annotation. Beyond pure close reading, through this interface, the user is thus also able to evaluate the quality of computational analysis. Further, the interface supports manual

annotation of the material, facilitating correction and teaching of machine-learned approaches.

The second of our pillars is an interface for statistical analysis, where the phenomena of interest can be analyzed en masse. However crucially, this interface is also linked to the close-reading one to further let the users delve into interesting outliers. Through this, they are not only able to derive hypotheses and explanations of the phenomena, but can also identify cases where outliers are more due to errors and omissions in our computational pipeline.

Finally, our third pillar is an agile pipeline to move data between these interfaces and our computational environment. In application, this third pillar is crucial, as it allows us to iteratively experiment with different computational indicators to capture the objects of our interest, with the results quickly making their way to experts for evaluation and explorative analysis. Through this analysis and evaluation, we then equally quickly get back information on not just the technical accuracy of our approach, but also if it captures the question of interest. Further, beside direct training data, we also get suggestions on new phenomena of interest to try to capture.

By maintaining from the start interfaces that allow both computer scientists and social scientists to not only view, but highlight to each other all aspects of the data, we also further a shared understanding between the participants. For example, social scientists are easily able to highlight to the computer scientists new phenomena of interest in the data derived from their close reading, while the computer scientists can easily show what they are currently automatically able to bring forth from the data. Through this, everyone is kept on the same page, misunderstandings are avoided, and the most fruitful avenues for development can be negotiated in a shared space where everyone contributes equally.

Combined with the capability for agile development and experimentation, this provides a versatile template for an iterative and discursive approach to digital humanities research, which moves toward questions of interest both fast,

as well as with high capability to truly capture the phenomena from all viewpoints of interest.

In this poster, we present insights into the interaction between close reading and computational methods gained from the work in our current project: Flows of Power: media as site and agent of politics. The project is a collaboration between journalism scholars, linguists and computer scientists aimed at the analysis of the political reporting in Finnish news media over the last three decades. We study both the linguistic means that media use to achieve certain goals (like appearing objective and credible, or appealing to the reader's emotions), as well as the structure of the public debate reflected there (what actors get a chance to speak and how they are presented).

Software and Data Formats

As many research questions in our project concern linguistic phenomena, a Natural Language Processing pipeline is highly useful. We employ the Turku neural parser pipeline [2], which provides dependency parsing, along with lower levels of annotation (tokenization, sentence splitting, lemmatization and tagging). Further, we apply the rule-based FINER tool [3] for named entity recognition.

Our primary toolbox for statistical analysis is R. This motivates using the 'tidy data' CSV format [4] as our main data format. In order to keep the number and order of columns constant and predictable, only the results of the dependency parsing pipeline are stored together with the text, in a one-token-per-line format very similar to CONLL-U.[1] All additional annotation layers, beginning with named entity recognition, are relegated to separate CSV files, where tuples like (documentId, sentenceId, spanStartId, spanEndId, value) are stored. Such tabular data are easy to manipulate within R.

For visualization, close reading and manual annotation, we decided to employ WebAnno [1].[2] While this tool was originally intended for the creation of datasets for language technology

1 https://universaldependencies.org/format.html
2 https://webanno.github.io/webanno/

tasks, its functionality is designed to be very general, which enabled its use in a wide variety of projects involving text annotation.[3] In addition to the usual linguistic layers of annotation, like lemma or head, it allows the creation of custom layers and feature sets. WebAnno has a simple but powerful visualization facility: annotations are shown as highlighted text spans, feature values as colorful bubbles over the text, and the various annotation layers can be shown or hidden at user's demand. This kind of visualization does not disturb close reading. It allows to concentrate on the features that are currently of interest, while retaining the possibility to look into the whole range of available annotations.

An important advantage is WebAnno's low barrier of entry. It is a Web application, meant to be deployed on a server and used through a Web browser. This kind of usage requires neither technical skills nor any installation on the users' machines. It provides user account and project management. The application can be also run locally, in form of a JAR file, which is useful for trial and demonstration purposes.

WebAnno supports several data formats for import and export. All of them assume one document per file. Among others, different variants of the CONLL format are supported. WebAnno-TSV is an own tab-separated text format, which, as opposed to CONLL, includes the custom annotation layers. Because it is a text format and is well documented, we are able to implement a fully automatic bidirectional conversion between our corpus-wide, per-annotation CSV files and per-document WebAnno-TSV files.

Thus, using WebAnno as an interface to interact with the domain experts who perform close reading and manual annotation, we are able to exchange our results quickly and with a high degree of automatization.

Case Study: Affective and Metaphorical Expressions in Political News

3 see: https://webanno.github.io/webanno/use-case-gallery/

We applied the methodology outlined above in a recently conducted case study. The subject of the study was the use of affective and metaphorical language in a media debate about a controversial labour market policy reform, called 'competitiveness pact' which was debated in Finland in 2015–16.

The linguistic phenomenon in question is complex and not readily defined. It is also highly subject-dependent: 'the ball is in play' is metaphoric when referred to politics, but not when referred to sports. There is no straightforward method or tool for automatic recognition of such phrases. Therefore, we started the study with a close reading phase, in which the media scholars identified and marked the phrases they recognized as affective or metaphorical in the material covering the competitiveness pact. The marked passages were subsequently manually post-processed to extract single words with 'metaphoric' or 'affective' charge. The list of words obtained this way was further expanded with their synonyms, obtained via word embeddings. Using this list, we were able to mark the potential metaphoric expressions in the unread text as well.

The final step, which is still in progress, is to validate the automatic annotation via close reading of another set of articles. WebAnno's functionality of highlighting and manual correction of annotations greatly facilitates such work.

References

1. Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), pages 76–84, Osaka, Japan, 2016.

2. Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 133–142, Brussels, Belgium, October 2018. Association for Computational Linguistics.

3. Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. A Finnish news corpus for named entity recognition. Lang Resources & Evaluation, August 2019.

4. Hadley Wickham. Tidy data. Journal of Statistical Software, 59(10), August 2014.

ID: 168
Short paper presentations
Topics: cultural studies, library & information science, crowdsourcing, cultural heritage collections, GLAM: galleries / libraries / archives / museums, infrastructure
Keywords: memory institutions, information infrastructures, participatory memory work, online communities

Online Participatory Memory Work: Understanding the Potential Roles of Online Mnemonic Communities in Building the Collections of Public Memory Institutions

Ina-Maria Jansson, Olle Sköld

Department of ALM, Uppsala University, Sweden

Introduction

A key task for humanist scholarship is to continuously interrogate the workings of human culture, community, and the many iterations and permutations in and by which they exist. This task has been taken on by digital humanists as well, not seldom with sights set on better understanding the impact of information technology and networked communication on present-day human affairs (e.g., Kirschenbaum, 2008). This paper shares this ethos, and argues for the importance of understanding digital information infrastructures and platforms when seeking to include collective memories of diverse groups into the collections of public memory institutions (e.g., archives, libraries, museums). The paper presumes that participatory memory work of online communities are not isolated processes but elements of digital ecologies. A purposeful documentation and preservation of those ecologies are essential for contextualization and understanding of the outcome of community memory work.

Technical changes in how people communicate create ripple effects that extend through a wide range of human endeavors and processes, including the shaping, communication, and reformulation of collective memory (Hoskins, 2009). Collective memories are thus dependent on, and carried on by technological and structural frameworks (for example common

entities like standards) that reconstructs shared values and concepts (Bowker, 2005). Just as a tool shapes the object of its creation so do technologies make their imprint on the information that is communicated through it. The increasing use and complexity of online digital platforms for communication within and between communities is here defined as such a technical innovation that shapes memory practices.

Research has shown that networked forums constitute important arenas for minority communities in the for example social, cultural, gender, medical or socio-economic sense (Af Segerstad & Kasperowski, 2014; Boyd, 2010; Marwick & Ellison, 2012; Wagner, 2018). In these online spaces, communities engage in many different memory-making practices for a variety of purposes and intents (e.g, Sköld, 2015, 2017). Such community memory-work also plays important roles of social support for its members. It also increases a sense of identification with, and inclusion in, the community itself (Assmann & Czaplicka, 1995), as shared memories of a community can be used to socialize new members into a group (often termed ‘mnemonic socialization’) in order for them to identify with the group’s past (Misztal, 2003). It is clear that digital community platforms consist an essential part of digital existence.

Here, an insight emerges with regards to participatory memory work in the memory institution-sector. The ability to support an inclusive and diverse public memory rests on key ongoings in the digital present being competently grasped, collected, and integrated into the collections of public memory institutions. Such an ambition can only be realized if it also includes the massively productive memory work communally conducted by online communities in the different spaces and services of the Internet. The memory work of online communities is however an understudied topic, and the opportunities and pitfalls present in the important endeavor to include the infrastructural complexity of shared memories of communities in the collections of public memory institutions are poorly understood (e.g., McDonough et al., 2010; Sköld, 2018a; Winget, 2011).

Aims, Materials, and Methods

The aim of this study is to explore the information infrastructural premises for the memory work of online communities and how public memory institutions can succeed better in their efforts to create diverse and inclusive collections by recognizing and supporting those premises.

The study is based on two case studies of memory work in online MMORPG videogame communities. Videogame communities offer an interesting case in relation to the aim of the study for several reasons. Firstly, videogames and videogaming are landmark features of digital culture today. Videogames and videogaming have impacted many arenas of contemporary life. Examples include technology development and adoption (Swalwell, 2007), management and organizational thought (Deterding et al., 2011), and the everyday interactions of many people by becoming sites of meaningful play and social interplay (Pearce, 2009), storytelling (Albrechtslund, 2010), learning (Barr, 2014), and knowledge production (Sköld, 2017). Secondly, and owning mainly to the ubiquity of the videogame phenomenon, videogame communities showcase many of the key issues and considerations that confront memory institutions aiming to build bridges between online-community memory work and institutional practices. Examples include ethical issues, legal and economic and ownership issues, and in the broader online space commonly occurring patterns of power relations and memory-making practices.

The aim of the study is met in two steps (RQ-1 and RQ-2), and is guided by a basic tenet of preservational work: successful curation rests to a significant extent on sufficient knowledge of the material in focus (Mortensen, 1999; Kirschenbaum, 2008).

– RQ-1. How are online communities conducting memory work, and what are the characteristics of the materials they produce as a result of this work?

– RQ-2. What are the potential results, pitfalls, and opportunities of efforts seeking to integrate the memory work

of online communities into the collections of public memory institutions?

The materials of the first case study consist of 40 World of Warcraft blogs collected in 2011 (Sköld, 2011). The second case is a study of 140 discussion threads (containing texts, images, videos, and audio) posted on a City of Heroes (CoH) discussion forum between 2012 and 2013 (Sköld, 2015). RQ-1 is answered by reporting on the WoW and CoH communities' practices of memory work, and typological analysis of the materials they produce. RQ-2 is met with guidance from theory of information infrastructures, the concept of institutionalization, and the results of previous research on videogame preservation (see e.g., Sköld, 2018b; Winget, 2011 for overviews).

Theoretical Framing and Discussion

The concept of information infrastructures makes visible the otherwise often transparent foundations for information and communication (Star & Ruhleder, 1996). It is employed in this study to distinguish between different information spheres and to understand the conditions and the challenges that has to be overcome when including material produced by online communities in collections of memory institutions. It is used to highlight the differences in settings and practices between the community sphere and the institutional sphere. Furthermore, this bridging process of information spheres is discussed in terms of institutionalization, which denotes the integration process of material created within the online community, to become part of institutional collections of archives, libraries, or museums.

As one of the many challenges and concerns for the institutionalization of online-community memory work, this paper argues that the organizational paradigms usually found in public memory institutions are among the most critical. For example, the multi-medial characteristics of online-community communicative memory may create difficulties for memory institutions whose collection management and mediation practices are centred on mono-medial materials. The benefit

that the institutionalization of (the often very diverse) online-community memory work offer to memory institutions seeking to support inclusive memory politics however makes it worthwhile to strive to overcome such hindrances.

The relevance of this paper extends beyond the issue of how and why online communities can and should be represented in the collections of memory institutions. It discusses the connectivities that can be built across communal and institutional practices of memory work and illustrates more broadly what challenges have to be met in order for other areas of contemporary digital culture and communication, like social media content, to potentially become a part of the cultural memories of our societies.

References

1. Af Segerstad, Y. and Kasperowski, D. (2014). A Community for Grieving: Affordances of Social Media for Support of Bereaved Parents. New Review of Hypermedia and Multimedia, 21(1–2):25–41.

2. Albrechtslund, M. (2010). Gamers Telling Stories: Understanding Narrative Practices in an Online Community. Convergence: The International Journal of Research into New Media Technologies, 16(1):112–124.

3. Assmann, J. and Czaplicka, J. (1995). Collective Memory and Cultural Identity. New German Critique, 65:124–133.

4. Barr, M. (2014). Learning Through Collaboration: Video Game Wikis. International Journal of Social Media and Interactive Learning Environments, 2(2):119–133.

5. Bowker, G. (2005). Memory Practices in the Sciences. Massachusetts: The MIT Press.

6. Boyd, D. (2010), Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications. In Z. Papacharissi (Ed.), A Networked Self (pp,47–66). New York: Routledge.

7. Deterding, S., Dixon, D., Khaled, R., and Nacke, L. (2011). From Game Design Elements to Gamefulness: Defining “Gamification”. In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, MindTrek ’11, Academic MindTrek 2011, Tampere, Finland (pp. 9–15). New York, NY: ACM.

8. Entertainment Software Association (2017). 2017 Essential Facts About the Computer and Video Game Industry. Retrieved from http://www.theesa.com/wp-content/up-loads/2017/09/EF2017_Design_FinalDigital.pdf

9. Hoskins, A. (2009). Digital Network Memory. In: A. Erll & A. Rigney (Eds.), Mediation, Remediation, and the Dynamics of Cultural Memory (pp. 91–106). Berlin: De Gruyter.

10. Kirschenbaum, M. G. (2008). Mechanisms: New Media and the Forensic Imagination. Cambridge, Mass: MIT Press.

11. Marwick, A. and Ellison, N. B. (2012). 'There Isn't Wifi in Heaven!' Negotiating visibility on Facebook memorial pages. Journal of Broadcasting & Electronic Media, 56(3):378–400.

12. McDonough, J., Olendorf, R., Kirschenbaum, M., Kraus, K., Reside, D., Donahue, R., Phelps, A., Egert, C., Lowood, H., & Rojo, S. (2010). Preserving Virtual Worlds Final Report. Retrieved from Illinois Digital Environment for Access to Learning and Scholarship: https://www.ideals.illinois.edu/handle/2142/17097

13. Misztal, B. A. (2003). Theories of Social Remembering. Columbus: McGraw-Hill Education.

14. Mortensen, P. (1999). The place of theory in archival practice. Archivaria, 47(1):1–26.

15. Star, S. L. and Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure> Design and Access for Large Information Spaces. Information Systems Research, 7(1):111–134.

16. Sköld, O. (2011). On social media and document theory. In Huvila, I., Holmberg, K., & Kronqvist-Berg, M., (Eds.), Information Science and Social Media: Proceedings of the International Conference Information Science and Social Media (ISSOME), vol. 1 of Skrifter utgivna av Informationsvetenskap vid Åbo Akademi, Åbo/Turku, Finland,(pp. 73–82). Åbo: Åbo Akademi University.

17. Sköld, O. (2015). Documenting virtual world cultures: Memory-making and documentary practices in the City of Heroes community. Journal of Documentation, 71(2):294–316.

18. Sköld, O. (2017). Getting-to-know: Inquiries, sources, methods, and the production of knowledge on a videogame wiki. Journal of Documentation, 73(6):1299–1321.

19. Sköld, O. (2018a). Documenting videogame communities: a study of community production of information in social-media environments and its implications for videogame preservation. Uppsala University, Uppsala, Sweden.

ID: 202
Long paper presentations
Topics: historical studies, linguistics, corpus linguistics, networks / relationships / graphs, visualisation
Keywords: Akkadian, Gephi, Korp, pointwise mutual information, fastText

Studying Semantic Domains in Akkadian Texts

Heidi Jauhiainen, Krister Lindén, Saana Svärd, Tero Alstola, Aleksi Sahala

University of Helsinki, Finland

In the Semantic Domains in Akkadian Texts project, we study semantic fields in texts written in Akkadian language. Akkadian is an East Semitic language that was spoken and written in ancient Mesopotamia, roughly in the area of modern-day Iraq, c. 2400 – c. 100 BCE. The texts were written in cuneiform script and the ones we are analyzing come from the Open Richly Annotated Cuneiform Corpus or Oracc (http://oracc.museum.upenn.edu). Oracc is an international cooperative undertaking providing free online editions of texts written mostly in the Akkadian and Sumerian languages. The text corpora in Oracc have been created by various projects and it is one of the largest electronic resources of cuneiform texts.

The snapshot of the corpus we have downloaded from Oracc contains 16,487 texts and almost 2 million words. About half of these texts have been tagged as having been written in the Akkadian language. The basic lexical unit in Oracc is the transliteration of the word, that is the representation of the cuneiform signs in Latin script. More than half of the words have been annotated with, for example, dictionary forms, word senses, and part-of-speech tags. No syntactic annotation of these texts has been published yet. Since Akkadian is an inflecting language, we have opted for using the dictionary forms of the words when analyzing semantic contexts.

Annotation of a text is always an interpretation by a scholar and, as Oracc is composed of a number of subprojects, the same word can be annotated in different ways in different

subprojects – despite available Oracc guidelines. Therefore, we had to do some preprocessing of the texts, such as normalizing the spellings of divine and geographical names. A word may also have several meanings, but homonyms are distinguished by their translation glosses in the annotation. As the annotation of the texts has been done by hand in many different projects, also the translation glosses are not always consistent, so we needed to unify synonymous translation glosses of the words that we wanted to study.

We use Pointwise Mutual Information (PMI) and fastText to find semantic contexts and relations of words in the Akkadian texts. PMI is a popular statistical association measure used in automatic collocation extraction (Church & Hank, 1989). PMI measures the ratio of observed word co-occurrence probabilities compared with their hypothetical co-occurrence if the word order of the corpus is randomized and all syntactic and semantic information is lost. PMI excels in discovering syntagmatic semantic relationships. FastText is a method that uses an artificial neural network model to generate word vectors (Bojanowski et al., 2017) which have been shown to model paradigmatic semantic regularities between words (Mikolov et al., 2013). FastText is a variation of a method called word2vec but, in addition to words, fastText takes subword information into account by representing a word as shorter sequences of characters. We, furthermore, use network analysis to study relations between clusters of words in context windows of approximately ten words. Since our dataset is quite small, we generally use at least two of these three methods together to study certain kinds of words and their semantic fields.

After preprocessing the data, we build a file with each document of the dataset on one line. We then extract collocations with PMI and may build word vectors with the similarity values produced by fastText. We have also experimented on building word vectors for each word with the collocation PMI values indicating a position in a multidimensional semantic space. We can look directly at the N most informative syntactic collocates using PMI or, using

fastText, we can extract the words in the semantic space that are closest to each of the words we are interested in. The results we get with computational methods are lists of words that are supposedly semantically similar to the words of interest as they occur in similar contexts.

Our dataset is not as big as the ones generally processed with such tools which affects the noise in the result, so it is imperative to manually evaluate the automatically proposed results, for which we have created a workflow. We start by visualizing the words and their clusters of similar words, for example, as networks with Gephi, an open source visualization tool (https://gephi.org). We typically build graphs for both the 10 and the 50 semantically closest words. The larger graphs of 50 are used for getting an idea of the wider contextual domain of the words of interest. In the smaller graphs of the 10 closest words, it is usually easier to spot the important links between words and, in some cases, the common contexts as well (see Figure 1). The analysis of the networks moves in a hermeneutic circle. After examining the graphs, a better understanding of the word contexts emerges.

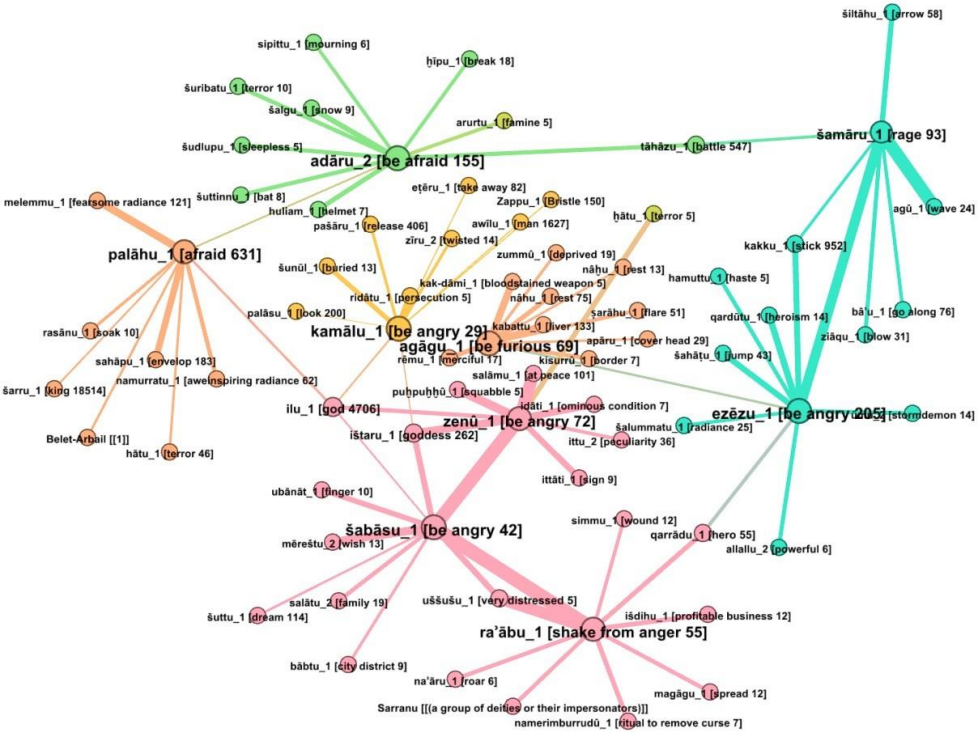


Fig.1. A selection of Akkadian words for anger and fear and their 10 closest collocates according to PMI.

Different possibilities can be explored by going back to the context of individual words. For this, we use the corpus search tool Korp to further analyze the contexts in which certain words appear together. Korp is an online service in the Language Bank of Finland provided by FIN-CLARIN (https://www.kielipankki.fi/support/korp/, Jauhiainen et al., 2019). In Korp, we can search for words that our methods indicate appear in similar contexts and the results of the query are presented as concordances. After examining the contexts of the chosen words in Korp, certain possibilities appear more likely than others. We then return to the graphs to distill the

idea further and use Korp at the same time to question and examine our analyses of contexts in graphs. The final results of this analytical process are then reviewed in light of previous lexical research as documented, e.g. in current lexical resources, on the Akkadian words under scrutiny.

In the presentation, we outline our methodology by describing the methods of our work combined with hermeneutically evaluating the results with the help of visualisation and the corpus search tool Korp. We then present some results we have reached by showcasing three recently concluded case studies. First, a network analysis on divine names demonstrating how the conquering Assyrian elite promoted their identity as rulers through the worship of their main deity Assur, and how he was integrated into the pantheon of established deities at the time (Alstola et al., 2019). The other two case studies are a study of verbs of seeing (Sahala & Svärd, forthcoming) and a study of emotion words (Svärd et al., in press), which both identify special usage contexts for synonyms which have not previously been documented in the available lexical resources for Akkadian. We present the main research results and describe how we arrive at our results by analysing semantically similar words in Akkadian texts.

References

Alstola, Tero, Shana Zaia, Aleksi Sahala, Heidi Jauhiainen, Saana Svärd, and Krister Lindén, 2019. Aššur and His Friends: A Statistical Analysis of Neo-Assyrian Texts. Journal of Cuneiform Studies 71: 159–180.

Bojanowski, Pjotr, Edouard Grave, Armand Joulin, and Tomas Mikolov, 2017. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5:135–146. https://arxiv.org/pdf/1607.04606.pdf (January 13, 2020).

Church, Kenneth Ward and Patrick Hanks, 1989. Word Association Norms, Mutual Information, and Lexicography. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 26–29 June 1989, Vancouver, British Columbia, Canada, 76–83. Stroudsburg, PA: Association for Computational Linguistics.

Jauhiainen, Heidi, Aleksi Sahala, and Tero Alstola, Open Richly Annotated Cuneiform Corpus, Korp Version, May 2019 [text corpus]. Language Bank of Finland. Retrieved from http://urn.fi/urn:nbn:fi:lb-2019060601 (January 13, 2020).

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013. Efficient Estimation of Word Representations in Vector Space. Last modified September 7, 2013. https://arxiv.org/abs/1301.3781 (January 13, 2020).

Sahala, Aleksi and Saana Svärd, forthcoming. Language Technology Approach to 'Seeing' in Akkadian. In Kiersten Neumann & Allison Thomason (eds.), Handbook on the Senses in the Ancient Near East, Routledge/Taylor and Francis.

Svärd, Saana, Tero Alstola, Heidi Jauhiainen, Aleksi Sahala, and Krister Lindén, in press. Fear in Akkadian Texts. In Shih-Wei Hsu and Jaume Llop-Raduà (eds.), The Expression of Emotions in Ancient Egypt and Mesopotamia. Culture and History of the Ancient Near East (CHANE), Brill.

ID: 208
Poster
Topics: medieval studies, philology, data modeling / knowledge representation, digital resources – publication and discovery, project design / organization / management
Keywords: lexicography, Old Norse

Legacy Data in a Digital Age

Ellert Thor Johannsson, Simonetta Battista, Tarrin Wills

University of Copenhagen, Denmark

In this poster presentation, we study the data used for the making of a historical dictionary and its development during three different periods. The focus is on A Dictionary of Old Norse Prose, which covers the language of Iceland and Norway in the Middle Ages. This dictionary has evolved from a rather straightforward collection of citations through a multi-volume, but incomplete, print publication to its current state as a dynamic online lexicographic tool, providing detailed information about the vocabulary of Old Norse and its textual foundation in medieval manuscripts and documents. Even though the dictionary is not finished, its wide scope is evident by the fact that its archive of around 800,000 example citations represents an estimated 7% of the entire 10 million word corpus of Old Norse. The long history of the project gives a unique opportunity to study the development of the data and how it has been used throughout the decades while the project has been in existence.

Background

Work on this dictionary began in 1939 long before computers and databases became available. The nature of the material and the editorial principles set out by the founders of the dictionary demanded a wide variety of data be collected and organized (cf. Widding 1964). This primarily involved excerpting the source material for examples of word use, which were then copied by hand onto slips and filed alphabetically in a physical archive. In addition to this, various other data were gathered about the medieval texts, such as the dating of manuscripts, bibliographic references to scholarly editions and secondary literature, information about foreign sources in case

of translations, and various other supplementary information. As with the example citations, all this information was also registered on paper through various filing systems.

The advent of computers opened up new ways to keep track of all this information. It was clear that the nature and scope of the material lent itself well to be organized in a database. The challenge became to convert all the existing information into a digital form and organize it in a database structure suitable for lexicographic work. This process gradually led to the development of an elaborate data structure and a tailor-made dictionary editorial system based on the information from the database.

The Data

The core of the dictionary is the collection of 800.000 example citations, each of which is provided with a sentence illustrating a specific form and/or meaning of the headword, a detailed reference showing the work of origin as well as page and line number. Even in the days of early computing, it was difficult to take advantage of the new technology because of the nature of the material, especially the widespread use of non-standardized characters and symbols.

Besides the dictionary citation archive, it was important to keep track of various information relating to the source material. Structuring this data involved creating an index of all the different medieval works, which had been excerpted for dictionary citations. The citations included a reference to scholarly editions as well as the manuscripts these were based on, so all this information had to be registered as well. This work was also done by hand.

The Database

In the 1980s the dictionary staff had realized the potential advantages of working with the data in a database structure. An evaluation report of the project from 1993 gives an insight into the thought process and considerations behind the design of the database (ONP 1993). The database needed to keep track of all the dictionary citations as well as the data related to the source material. This involved creating many different

tables where all the bits of information are stored in separate fields, such as wordlist table, headword table, definition table, citation table. Moreover, there are additional tables that hold references to the literature and other glossaries. The tables were then linked together through the headword field common to all tables. This allowed for additional information to be added relating to both the source material and each citation. The most important of those was noting the geographical provenience of the manuscripts and the grouping of the source material by literary genres.

The benefit of organizing the information in the database was immediate. Once the content of the hand-written index registry had been entered into the database the information was made available as a printed volume published in 1989 (ONP Registre). Even though this work is primarily designed to facilitate the use of the paper dictionary it stands alone as an independent reference work over Old Norse prose texts and their manuscript origins.

Online Dictionary

The database work facilitated the preparation and eventual publication of dictionary entries. Once the citations had been keyed into the database the editors could proceed with the structuring of dictionary entries, supplying extra grammatical information as well as information about collocations and syntactic relations.

After the publication of three printed volumes of the dictionary from 1995 to 2004, containing entries that cover the alphabet from a-em, the project underwent another restructuring process, which resulted in an online version made available in 2010. An important step in the conversion of the paper dictionary to a fully digital online dictionary was the scanning of ca. 500.000 non-typed paper slips, which were integrated into the database and linked to the same fields as the typed citations (cf. Johannsson 2019).

After this restructuring, the database became an essential part of the dictionary as online users could query the database directly and search the data in different ways being no longer

limited by the printed alphabetical list of dictionary entries. Besides headword search, the database structure makes it possible to search the data by several criteria, such as the dating of the original manuscript, country of provenance, work, literary genre, and so on.

Enhancing the Data

Since 2010 the online version has gradually expanded with new edited articles and it has been redesigned and improved with additional search options. The dictionary database is no longer the only source of information that is available to the user of the dictionary. The data have been enhanced through various ways of linking them internally and to other digital resources. There are now links to other dictionaries as well as digital editions of Old Norse texts (cf. Wills et al. 2018). We demonstrate how the data can be used in different ways and how they are displayed in the current online version of the ONP dictionary, e.g. by the reader feature which provides glossaries to scholarly text editions (cf. Wills & Johannsson 2019). In this way, ONP remains an important research tool for scholars in medieval Scandinavian language, literature, and culture.

The current study demonstrates how legacy data, originally only organized in a paper filing system, have been structured in a database and improved in various ways through three main periods in the project's history. We show that even though the original data still provide the basis of the dictionary they have been built upon and enhanced by innovative use of the information from a specialized database as well as by external digital sources.

References

Johannsson, E. (2019). Integrating analog citations into an online dictionary in C. Navarretta, M. Agirrezabal, B. Maegaard (eds.) Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, pp. 250–258.

Johannsson, Ellert Thor & Simonetta Battista (2014). "A Dictionary of Old Norse Prose and its Users – Paper vs. Web-based Edition", in Andrea Abel & al. (eds.): Proceedings of the XVI EURALEX International Congress: The User in Focus, 15–19 July 2014, Bolzano/Bozen, 169–179.

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

ONP = Degnbol, H., Jacobsen, B.C., Knirk, J.E., Rode, E., Sanders, C. & Helgadóttir, Þ. (eds.). Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose. ONP Registre (1989). ONP 1: a-bam (1994). ONP 2: banda (2000). ONP 3: de-em (2004). Copenhagen: Den Arnamagnæanske Kommission.

ONP 1993 = Evaluation of the Production Plan for the Dictionary of Old Norse Prose. Copenhagen: Ministry of Education and Research.

Wills, T., Jóhannsson, E., & Battista, S. (2018). Linking Corpus Data to an Excerpt-based Historical Dictionary. In J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.) Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 979–987.

Widding, Ole (1964): Den Arnamagnæanske Kommissions Ordbog, 1939–1964: Rapport og plan, Copenhagen: G.E.C.GADS Forlag.

ID: 210
Short paper presentations
Topics: cultural studies, literary studies, philology, data mining / text mining, natural language processing, stylistics and stylometry / authorship attribution
Keywords: topic modelling, literature, themes

Text Mining Themes of the Urban Night in Historical Literary Corpora

Hanne Emilia Juntunen

Tampere University, Finland

In this presentation, I will go over how I have used the text mining method of topic modelling to discover salient themes associated with the literary urban night using historical text corpora.

The study falls under the umbrella of digital literary studies. Its interest focuses on large-scale historical thematic trends which are difficult to study with traditional literary methods. As is usual in literary studies, the object of this study is a theme, rather than an era, certain authors, or places. Specifically, the interest lies with the subthemes, the themes that cluster around a larger theme, the urban night. Topic modelling was used for this discovery, supported with corpus linguistic methods to produce the most salient themes, or topics, associated with the urban night in the data used.

The topic modelling approach was chosen for the study of the literary urban night because it has been studied with qualitative methods applied to a relatively large number of texts to produce generalizable statements about its subthemes – a quantitative approach is therefore both relevant and lacking. The study has a large timeframe, looking at literary texts from 1500's to 1920's. This trajectory represents a historical time when the urban night evolved into the phenomenon we now recognize: before the sixteenth century, walking outside at night was illegal in most major European cities! The timeframe also foregrounds the historical thematic trends under consideration: idiosyncratic and short-lived trends get lost in the large mass of texts. Similar timeframes are, moreover, employed with qualitative methods in the studies of the literary

(urban) night more generally as well. As such, the questions the study set out to answer were whether the method would produce such consistent historical thematic trends, and whether the results would challenge or support the established understanding of the literary urban night.

The data used in this study is comprised of several full-text corpora of literary texts: Early English Books Online, Eighteenth Century Collections Online, Corpus of Late Modern English Texts, Corpus of English Novels, and Tampere Corpus of English Novels. Together, they span the years 1500–1923. These digital resources are all, except for the last one, freely available for research use. Some of them contain a mixture of genres and texts from different centuries. These have been sorted into centennial and literary-only and mixed-genre subcorpora (parts of a larger corpus) automatically, and randomly checked manually. American literature was hand-picked out of the corpora as well. The size of the corpus obtained by these operations is 7797 full literary texts.

In order to start applying the text mining method of topic modelling to the data, the literary urban night had to be operationalised. That is, the qualitative feature of story-telling, the theme of the night of the city, had to be transformed into a quantifiable and measurable variable. It was decided to focus on the explicit mention of the night in the context of the town as this was easiest to automatically detect. Furthermore, a comparison of the occurrences of words that signal a nocturnal setting in literature showed that these words (such as 'lamp', 'candle' and 'moon') occur in similar patterns or with notably less frequency. Taking other words into consideration was thus deemed unnecessary. A corpus tool was used to extract texts that contain the nodewords (‘night’, ‘nocturnal’ and ‘nyght’) in the desired context, the words 'town', 'city' and 'urban' within a 40-word window. These texts were then compiled into a corpus of their own, forming the urbannight-subcorpus of 1686 full texts.

The new urbannight-subcorpus was then used to extract the themes. The text mining method used was Latent Dirichlet Allocation (LDA). Topics were extracted from the full texts

which produced a set of topics that limited both in terms of its internal consistency and variety (3–5 topics per centennial subcorpus). These topics were very generic, and most likely reflect the global themes for the texts, i.e. themes that the texts as a whole thematise in many different contexts. However, the focus of the study was on subthemes, or themes local to the urban night. To get at the local level of the urban night, several chunking options were tested. The best solution was judged to be chunking based on the nodewords including 500 words to the left and 500 to the right. This enabled the modelling to focus on only the most relevant parts of the full texts. Lastly, these extracts were lemmatized, that is, the different word forms ('write/writes') were collapsed into the basic word form, using the topic modelling software's own WordNet lemmatizer, and analysed separately using the Bag of Words method with Euclidean distance. Each century produced a set of topics that varied slightly (10–14) with some variations in parameters (lower threshold at 0,05–0,10, and the upper at 0,15–0,25). The topics obtained in this way were more internally consistent as well as representing a wider variety of phenomena than both the full text and the sequentially chunked text, which were rather similar, and lacked some areas (e.g. entertainment) that were prominent in the nodeword-chunked data.

The justifiable criticism has been aimed at topic modelling that it is quite subjective. Indeed, as the labelling of the topics is decided by the researcher, the results of the method can be subjective, even spurious. To mitigate this, a common practice from qualitative thematic analysis was adapted for the purposes of this study: double-coding. This means simply that two different people assign themes (or 'code' themes) to the same data independently of each other. The result of this double-coding is then checked for intercoder agreement (ICA), that is, how often the same segment is considered by both to fall under the same theme. This is used specifically to counteract bias that results from the subjective theme assignment of one researcher. Double-coding is usually done by two humans, but in this case the words composing the topics obtained from LDA were subjected to semantic tagging

(assigning tags that indicate the meaning of the word using the USAS Semantic Tagger). The entire corpus could not be tagged in this way due two factors, one being a technical limitation, and the other that the tagger functions best with contemporary language. Only the resulting themes were therefore semantically tagged, and the final labels for the topics were based on the results of both the intuitive labelling and these tags. Despite the relative context-lessness of the tagged words, the tagging turned out to be a valuable and fruitful addition to human intuition, challenging and combating subjective bias in labelling.

The list of historical thematic trends was not yet complete, however. The themes obtained needed to make sense from a literary analytical perspective. Therefore, topics indicating e.g. reported speech and interaction between characters were dropped from the final listing – it is hardly a discovery that novels have characters who talk with each other. The final list of thematic trends does contain some discoveries. The final trends were body & experiences, entertainment, family, rulers, journey, military, and religion. These themes could be found in either four out of five or all five centennial subcorpora, forming the six main lines of historical thematic trends in data.

To minimize the effect of context knowledge, relevant research literature (i.e. cultural and literary analysis) into the urban night was consulted for comparison only after obtaining these final results. As noted, prior to this study, the urban night has been only analysed with traditional qualitative methods, and the themes covered in these studies include entertainment and religion quite prominently; rulers, journey, and family less so but still to a certain degree; and military not at all. The by far most prominent theme in these qualitative studies, lighting, was absent in the results of this study.

The pre-existing research literature relies on a different scale of data and analysis as the present study – at best, the data consists of some hundred texts (per study), whereas the results of this study were obtained from nearly 1700 texts with a wider variety of authors, cultural status and plot significance of the urban night than qualitative research could encompass.

As such, it can be preliminarily concluded that quantitative methods like topic modelling can present a significant contribution to the existing research done on the themes of the urban night, as it can produce meaningful thematic trends, and both confirms some aspects of pre-existing analysis and challenges others.

Lastly, the software programme used for the topic modelling was Orange. It has a graphical user interface and requires the minimum amount of coding skills. While certainly not foregoing learning the differences between qualitative and quantitative approaches, it is important to recognize especially in digital literary studies that if we want the field to grow and be contributed to by researchers with a classical literary studies training, the tools we use need to be truly available. Not only low-cost or completely free of charge, but most importantly such that they do not require extensive pre-existing skill sets to use. Orange fulfils these requirements.

ID: 128
Long paper presentations
Topics: linguistics, philology, corpus linguistics, crowdsourcing, data modeling / knowledge representation, diversity and multilingual / multicultural approaches, linking and annotation, natural language processing, computational science, big data, citizen humanities, citizen science
Keywords: annotation, translation projection, sentiment analysis, emotion analysis, crosslingual

Emotion Preservation in Translation: Evaluating Datasets for Annotation Projection*

Kaisla Kajava[1], Emily Öhman[1], Hui Piao[2], Jörg Tiedemann[1]

[1]University of Helsinki, Finalnd; [2]University of Tokyo, Japan

This paper is a pilot study that aims to explore the viability of annotation projection from one language to another as well as to evaluate the multilingual data set we have created for emotion analysis. We study different language pairs based on parallel corpora for sentiment and emotion annotations and explore annotator agreement. We show that the source data is a possible source for reliable L1 data to be used in annotation projection from high-resource languages, such as English, into low-resource languages and that this is a reliable way of creating data sets for fine-grained sentiment analysis and emotion detection.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 201
Long paper presentations
Topics: computational science, big data
Keywords: food, computing, complexity, topic, modelling

Tracing Complexity in Food Blogging Entries*

Maija Kāle[1], Ebenezer Agbozo[2]

[1]University of Latvia, Latvia; [2]Ural Federal University, Russia

Food consumption is a complex phenomenon involving many aspects that are easier or harder to be captured, one of them being eating habits that are partly determined by unconscious choice mechanisms (Mai et al, 2011). To large extent it presumes favoring nourishing, calorie rich foods over others, creating a situation that among the most urgent health issues to solve in this world are obesity, life-style determined Type 2 diabetes, cardiovascular diseases and other illnesses related to food and lifestyle in general. With a growing digital consumption that entails rich representation of food (food items, cooking lessons, food shows, plating aesthetics) humans are more and more exposed to food images and descriptions, which in turn impacts the way we think about the food and the way we consume it. One of the most interesting phenomena when discussing food related decisions is complexity (Spence et al, 2018). The more complex the food, the more likeable it is – and the more documented the story of complexity, the more expensive the food can get (Mccall and Lynn, 2008). Within this paper, we focus on complexity and how it is represented in food blogging entries. We turn specific attention to complexity capture when looking at healthy and unhealthy foods, testing the hypothesis that healthy foods are represented as less complex and, thus, less likable than unhealthy foods.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

With this paper we add to the knowledge body of food representation in digital media and adhere with the focus point of DHN 2020: Digital language resources (e.g. historical texts, parliamentary records, speech and multimodal corpora, social media data, etc.) and tools for digital humanities and social sciences (e.g. linking data across domains, media and languages).

ID: 165
Long paper presentations
Topics: communication studies, linguistics, corpus linguistics, data mining / text mining, discourse analysis
Keywords: affect studies

Modal Grammar and Metaphoricity as Vehicles of Affectivity in Political Newspaper Journalism

Antti Kanner[1], Anu Koivunen[2], Eetu Mäkelä[1]

[1]University of Helsinki, Finalnd; [2]Tampere University, Finalnd

This paper introduces an approach to derive computationally tangible markers and higher-level modes of social behaviour through the identification of pragmatic functions of language. In historical and social science studies utilizing text mining methods, the focus of the studies is seldom in the actual patterns of data that these methods yield, but some higher-level aspects of linguistic behaviour (such as conceptualizations or discourses) inferred from these patterns, which are predominantly based on lexical co-occurrences. While off-the-shelf methods have certainly their place in the workflow, we argue that the core of a methodological approach should start from the categories and analytical concepts of the study. The study presented here offers an example of such an approach.

The context of the study presented here is the project Flows of Power: Media as Site and Agent of Politics (Academy of Finland 2019–2022), which investigates the agency of journalistic media in the flows of information, public opinion and power. Through a large-scale empirical analysis of Finnish journalism between 1998 and 2018, it seeks to explore the strategies of journalistic news media in staging and managing political processes. In that context, the role of the present study is that of a pilot case, exemplifying political and economic polarization in contemporary Europe and, as such, a challenge for media attempting to resist being polarized. The focus is on the media coverage of a spectacular political conflict in 2015–2016 between the Finnish right-wing conservative government

and the trade unions. The conflict resulted in wage freezes, reduced pay for public sector employees, extensions of annual working time and increasing social security contributions from employees.

As a theoretical starting point, the study takes the notion of affective economy (Ahmed 2004) from cultural theory, used to analyse affectivity not as properties of subjects or objects but as qualities in movement and circulation themselves. Our approach is based on the idea that the dialogic nature of language imposes writers to pre-emptively adapt to the perceived affective atmosphere of their readers, proposed in the field of discourse oriented linguistics and pragmatics (eg. White 2003). In newspaper reporting, this manifests through what Gaye Tuchman (1972) has termed the ‘‘strategic ritual of objectivity’’, the desire to appear to stand outside or rise above the subject at hand: to be dispassionate, disembodied and impartial. Traditionally, journalistic genres have carefully distinguished between news and opinion, relegating judgement and emotionality to columns, editorials and other opinion pieces. However, as Karin Wahl-Jørgensen (2019) argues, there is also a strategic ritual of emotionality operating alongside that of objectivity, entailing conventions and codes for incorporating affects into storytelling – and of hiding and displacing emotion. Coming from this background, we thus start from the assumption that affective flows and intensities are circulated in news media not just by overt expressions of sentiment, but through a specialized convention which necessitate a form of affective linguistic labour from the part of writers. This paper thus seeks to develop a methodological approach and a corresponding workflow through which it becomes possible to recognize traces of such affective linguistic labour by grouping linguistic structures that correlate with structures that are known to perform affective functions.

The full dataset of the FLOPO project consists of the whole published material of key Finnish news agencies (STT), newspapers (Helsingin Sanomat), public service broadcasting (YLE) and daily tabloids (Iltalehti). The subset used in the pilot project (of which this paper forms a part) covers news reporting

on the topic of Competitiveness Pact from early 2015 until the end of 2016. The dataset used here is thus relatively stable concerning the themes discussed in the texts. The texts were assigned metadata categories relating to genre (news report, commentary, analysis and editorial), the temporal location concerning political events reported and outlet. Each text was also annotated to hold information about intratextual segments, allowing to observe whether a given word resides in quotations or the beginning of a text or not. The assumption was made, that journalistic conventions dictate how linguistic expressions mediating affective intensities may distribute across these variables. Presuming that adjectives with evaluative and emotive meanings are used in affective functions provided the analysis with a baseline against which other features could be compared. This presumption is not only corroborated by the wide use of emotive and evaluative adjectives in Sentiment Analysis but has also been explicitly observed in corpus-based studies in newspaper journalism (Huang 2018). Evaluative adjectives especially are a good entry point because their evaluative meaning is often their primary semantic component and is not based on interpretations of their use in context.

Samples of texts were manually close read by experts to extract other passages of texts which had heightened levels of affective intensity. From these passages, the affective lexical core was extracted manually. A considerable number of these expressions were used metaphorically in their context and hypothesis was made that affectivity of these passages was somehow based on or tied to their that metaphoricity. This hypothesis was tested by expanding the set of words that were used metaphorically with other words belonging to same conceptual domains (often of sports, war or physical pain) using pre-trained word embeddings and analysing whether this list of words exhibited similar patterns of distributions as adjectives.

According to the dialogic view of language, a considerable degree of the selections of linguistic structures in texts is not directly derived from their propositional content but has more

to do with how that content is framed and how the writers align their position and the position of their perceived audiences in relation to that content. All this gives reason to assume that, alongside perhaps more obvious emotionally loaded vocabulary, grammatical structures also play an integral part in how writers at the same time adapt to and reproduce the affective intensities. This motivated another hypothesis according to which two related grammatical categories, evidential structures and epistemic modals, contributed to the conventions of affective mediation in newspaper journalism alongside overt emotive and evaluative vocabulary. The use of these structures, especially in Academic writing, has been analyzed with the concept of hedging, a politeness strategy through which writers pre-emptively make their claims less threatening by reducing their level of certainty or assurance. As things like certainty and trustworthiness are highly affective in also journalistic practices it seemed reasonable to assume that these structures would also be relevant in the conventions through which journalistic writing mediates affectivity. A wide range of linguistic markers, identified by established linguistic scholarship in Finnish to function (among others) as evidentials were tested – modal verbs and verb constructions, modal adjectives and adverbs, connectives and so on (eg. ISK 2004, Kangasniemi 1990, Laitinen 1989), building up to around 90 distinct linguistic markers. These structures were also compared with the affective baseline provided by the evaluative and emotive adjectives.

The results seemed to confirm both hypotheses and point towards the interpretation that both metaphors and hedging strategies could be used as markers for identifying heightened affective intensity and concentrations of affective linguistic labour. Instead of cataloguing each affective expression in the data, the idea here was to chart the functional resources used in mediating affective intensities, as it is likely that observations about them retain their validity outside the studied case. While a metaphoricity of a given word depends on whether the news piece is about partisan politics, war or ice hockey, that metaphors, in general, have affective values in each of them is

likely to remain true. The upside of this approach, we argue, is not that it would produce readily-usable resources applicable to other cases, but instead offers a framework through which content-sensitive expertise can intrude the computational operation.

Thus our paper contributes to journalism studies, developing a theoretical and methodological approach to affectivity in news and actualities. Integrating discourse-oriented linguistics and pragmatics into affect studies entails re-introducing a linguistic model to a post-linguistic theory frame. This, we suggest, is necessary to understand affectivity as meaning-making, an important feature of news journalism beyond explicitly emotive storytelling. From the perspective of Digital Humanities, the study introduces two insights into how humanities expertise can be operationalized in the context of large scale computational analysis of complex discursive phenomenon, first by showing how focusing on presents opportunities not available in content-agnostic settings and, second, by showing how functional and abstract linguistic structures can become accessible by taking known example categories as a starting point.

References

Ahmed, Sara (2004) The Cultural Politics of Emotion. Edinburgh: Edinburgh University Press.

ISK 2004 = Alho, I. (2004). Iso suomen kielioppi. Helsinki: Suomalaisen Kirjallisuuden Seura.

Kangasniemi, Heikki (1992). Modal expressions in Finnish. Helsinki: Suomalaisen kirjallisuuden seura.

Laitinen, Lea (1989). Välttämättömyys ja persoona: Suomen murteiden nesessiivisten rakenteiden morfosyntaksia ja semantiikkaa. [Helsinki]: Helsingin yliopisto.

Tuchman, Gaye (1972) ‘‘Objectivity as Strategic Ritual: an examination of newsmen’s notions of objectivity’’, American Journal of Sociology 77.

Wahl-Jørgensen, Karin (2019) Emotions, Media, and Politics. Cambridge: Polity.

White, P. R. R. (2003): “Beyond hedges and modality: A dialogic view of the language of intersubjective stance”, Text 23(2) (2003), pp. 259–284.

ID: 148
Poster
Topics: historical studies, linguistics, philology, bibliographic studies, corpus linguistics, data mining / text mining
Keywords: vernacularization, modal expressions, public sphere, nation building

Becoming a State Language: Finnish Public Debate and Modal Grammar 1820ï 1917

Antti Kanner, Hege Roivainen, Tuuli Tahko, Jani Marjanen
University of Helsinki, Finalnd

Introduction

The early modern and modern periods in Europe entailed a transformation in linguistic geography, where diglossic systems (Ferguson 1959; Fishman 1967; Hudson 2002) that had prevailed for centuries started to erode and new national languages based on local vernaculars were adopted and promoted as languages of administration, law, politics, learning and the arts. This process is sometimes labelled as vernacularization. In a three-year project at the University of Helsinki, we combine theories of nationalism and the sociology of language to build our understanding of the complex interplay of state, language and public discourse in the dynamic linguistic environment of vernacularization. We do this by using newly available digitized historical corpora that can be accessed with computational text mining. The project brings together perspectives from 1) studying bibliographic information to look at changes in the publishing landscape, 2) historical analysis of changing past perceptions of different languages, and 3) linguistic analysis of how the features of a language change when it becomes more readily used in a written, more elevated form in public discourse.

This poster focuses on the third perspective by analyzing changes in the Finnish language in a diachronic corpus consisting of digitized newspapers from 1820–1917. In the Finnish context, vernacularization happened in two steps, first with Swedish being adopted more and more as a language of administration and science in the eighteenth century, and second, in a more powerful and rapid transformation, when

Finnish developed into a national language with state and nation bearing functions during the course of the nineteenth century. From having been an underdeveloped and underprivileged written language, Finnish had, due to active promotion and development of written standards, become a state-bearing language by the early twentieth century (Engman 2016; Huumo, Laitinen & Paloposki 2004). Our hypothesis regarding the development of modern standard Finnish is that once Finnish became more readily used in public debate, more nuanced and complicated structures emerged to countenance the newly arisen rhetorical needs. The increase in printed material required a more elaborate notion of the public, and the potential audience and readership of the texts – an audience that was largely unknown to the author.

In the study, we are especially interested in the linguistic structures expressing epistemic modality and other forms of evidentiality. We hypothesize that these specific linguistic resources, used to align authors' views with those of their perceived audiences, are robust markers of larger linguistic change that took place when language transcended from mostly agrarian spoken language to a literary and administrative language. Hence we study the emergence, frequency, and distribution of epistemic modal expressions. We examine this hypothesis by comparing it against two separate baselines. The first one of these establishes the general confidence interval of temporal variance and is made out of a random sample belonging to comparable grammatical categories. The second one seeks to establish the general temporal pattern of emerging public discourse, by matching the aforementioned grammatical constructions with key terms related to the overt discourse on the social shifts of the public sphere (especially the rise of terminology like julkisuus 'public sphere' or julkiso 'the public') and variables connected to the development of the concrete material limits of public discourse (ie. growth in book printing and newspaper publishing (Tolonen et al. 2018; Marjanen et al. 2019)), respectively.

Materials and Methods

Our data set consists of newspapers published in Finnish between the years 1771 and 1917 (with first Finnish-language publications from 1820 onwards). These have been digitized and made available for data and text mining by the National Library of Finland. The bulk of the newspapers are in Finnish and Swedish. They consist of 5.2B (in Finnish) and 3.4B (in Swedish) token words and provide a nearly complete record of newspapers and periodicals in the country published in this period. The newspaper corpus as such cannot be seen as representative of the Finnish language in general, but it is the best historical corpus available, as newspapers covered a wide range of topics and recorded new features in language by publishing everything from poetry to reports on political events and reflections on academic texts.

As linguistic markers, a catalogue of 92 central modal expressions has been selected based on established descriptive grammars of Finnish. Not all of them have epistemic or evidential uses in modern Finnish, but as we are building a scalable approach it is advantageous to look at a wider range of expressions. Furthermore, a possible implication of our working hypothesis – that epistemic and evidential expressions developed their fine-lined ecologies as an outcome of the vernacularization process – is that it is quite plausible that the division of labour between expressions devoted to deontic, dynamic and epistemic modalities was more fuzzy and dynamic before the onset of that process. Our catalogue of modal expressions includes morphological moods, modal verbs and verb constructions (voida ‘can’, on tehtävä ‘must be done’), modal adjectives (e.g. todennäköinen ‘probable’, ilmeinen ‘obvious’) and modal adverbs (e.g. oletettavasti ‘presumably’, tuskin ‘hardly’). As is often the case with vernacularization, the sources of newly developed expressions fall roughly under three categories. The first are expressions that already existed in the language, either in earlier written forms or in spoken dialects only but which undergo either a semantic or syntactic change (or both). The second group are translations and calques from languages that are further ahead in the vernacularization process (in the

case of Finnish, mostly from Swedish), these languages providing models for what kind of expressions are presumably required to fulfill literary, administrative and public functions. Finally, the third group are productive formations based on the language's own resources which have not been in use earlier nor have any obvious outside models.

In analysing the development of modal expressions the key objects of interest are 1) the overall frequencies of the modal expressions, relative to the amount of text in general and relative to each other, and 2) the scope of use of each expression. The hypothesis of the study dictates that there should be considerable changes in the ecology of epistemic modal expressions, these changes mainly taking the shape of the specialization of functions for a number of expressions and that shifts in these patterns should happen in concordance with other variables describing the emergence of the public sphere. The most robust signals for these changes are perhaps the expression's relative frequencies, mapping which is relatively trivial task (given the common technical reservations relating to unreliable OCR results and the temporally uneven distribution of the text mass). A more detailed picture is achieved by looking at the use of modal expressions as a whole, which is much more demanding undertaking: the expressions' grammatical, contextual and semantic features have to be tracked simultaneously and in correspondence to each other. The overall approach is akin to behaviour profile analysis, where occurrences of the studied linguistic items are examined across a wide range of variables and then subjected to scrutiny by univariate, bivariate and multivariate statistical tests (eg. Divjak & Gries 2006; Arppe 2008).

Concluding Discussion

Our study traces historical changes in language features and relates them to the development of the public sphere in Finland. Detailed analysis, based on large-scale historical corpora, of where and when modal expressions have come to be used in written Finnish is a major contribution to the study of the Finnish language, even if the central aim of our study resides in understanding the historical process of

vernacularization. We have already found that changes in the relative frequency of a selection of the modal expressions increase in conjunction with key stages in the development of the Finnish press, which supports our original hypothesis. However, a full analysis, including multivariate statistical tests, is still needed for making this argument in full. We further believe that testing our hypothesis on the Finnish case, which is rather straightforward with a relatively rapid vernacularization process, may lead to testing the hypothesis for other languages as well. If the growth of modal expressions is a feature that reflects the writers' increasing need to take into account an abstract notion of the general public, we should see similar developments elsewhere as well.

References

1. Arppe, A. (2008). Univariate, bivariate and multivariate methods in corpus-based lexicography. Publications of the Department of General Linguistics, University of Helsinki, No. 44.

2. Divjak, D. & Gries S. Th. (2006). Ways of trying in Russian: clustering behaviour profiles. Corpus Linguistics and Linguistic Theory, 2(1): 23–60.

3. Engman, M. (2016). Språkfrågan: Finlandssvenskhetens uppkomst 1812–1922. Helsingfors: Svenska litteratursällskapet i Finland.

4. Ferguson, C. (1959). Diglossia. Word, 15, 325–340.

5. Fishman, J. A. (1967). Bilingualism with and Without Diglossia; Diglossia With and Without Bilingualism. Journal of Social Issues, 23(2), 29–38.

6. Hudson, A. (2002). Outline of a theory of diglossia. International Journal of the Sociology of Language, 2002(157), 1–48. https://doi.org/10.1515/ijsl.2002.039

7. Huumo, K., Laitinen, L., & Paloposki, O. (Eds.). (2004). Yhteistä kieltä tekemässä: Näkökulmia suomen kirjakielen kehitykseen 1800-luvulla. Helsinki: Suomalaisen Kirjallisuuden Seura.

8. Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L., & Tolonen,M. (2019). A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917. Journal of European Periodical Studies, 4(1), 54–77. https://doi.org/10.21825/jeps.v4i1.10483

9. Tolonen, M., Lahti, L., Roivainen, H., & Marjanen, J. (2019). A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 52(1), 57–78. https://doi.org/10.1080/01615440.2018.1526657

ID: 155
Short paper presentations
Topics: linguistics, philology, cultural heritage collections, digitisation – theory and practice, digital resources – publication and discovery, image processing, open data, computational science, artificial intelligence
Keywords: OCR, neural networks, tesseract, Norwegian, fraktur

Targeted, Neural Re-OCR of Norwegian Fraktur

Andre Kåsen, Lars G. Johnsen
National Library of Norway, Norway

1. Introduction

Mass Digitization demands high precision tools right from the start. Since 2006, the National Library of Norway has had as its goal to digitize the whole national bibliography, a goal which was reached in 2018 with a digital collection of over 500 000 books and 1.2 million newspapers. Therefore, the time has come to take a look back and evaluate the result of the digitized material, particularly targeting the older novels with a fraktur typeface.

The Fraktur Problem or the pre-1900 problem is known to be detrimental to the quality of optical character recognition (OCR) of older fonts [1,3]. The problem is illustrated for fraktur in 1 where the letter æ is OCRed as either ce or cr.

Proprietary systems like ABBYY FineReader are every so often outper- formed by neural systems like e.g. Tesseract and OCRopus [7]. These systems are also open-source and are therefore preferable, because it invites technologists and digital humanities (DH) scholars to work together. Such systems also rely on high-quality data which all so often is produced at the hand of the DH scholar rather than in the technologist community.

2. Another Ground Truth for Fraktur

Norwegian began to diverge from Danish with respect to lexical variants in the mid-to-late 19th century. Especially with the alternate written standard Landsm˚alet a set of new words was introduced. Although both written Dan- ish and Norwegian are quite similar at this stage, the differences warrant a new training

material, a new ground truth, in order to systematically improve the quality of the OCR.

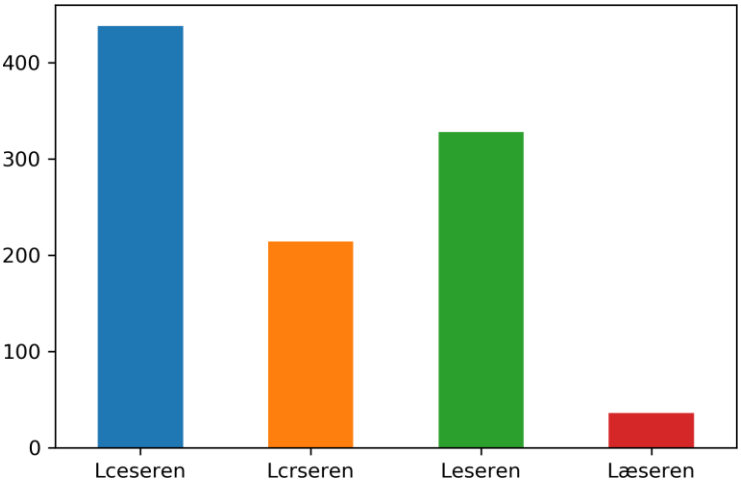


Fig. 1. Raw frequency counts of variation in the spelling of ‘reader’ in Norwegian in the period 1850 to 1900.

2.1 Relevant Work

Both Swedish, Danish and German fraktur is supported by Tesseract. These models, except the German one, was trained with an earlier version of Tesser- act[1] i.e. pre-neural Tesseract. There has also been substantial work for Finnish newspaper fraktur with Tesseract (see [4]).

2.2 The Collation Process

In order to create a new gold standard ground truth, scans of books and their textual counterpart have been aligned and checked by a human annotator. This was and is done using the editor jTessBoxEditor[2] a sub-module of the VietOCR program. The output of this process is box files i.e. text files where every

[1] https://github.com/tesseract-ocr/tesseract
[2] https://sourceforge.net/projects/vietocr/files/jTessBoxEditor/

char- acter in the text is assigned a line coordinate in the book scan like the following:

```
S 52 943 706 995 0
k 52 943 706 995 0
r 52 943 706 995 0
i 52 943 706 995 0
f 52 943 706 995 0
t 52 943 706 995 0
  52 943 706 995 0
```

Fig. 2. An example word from a box file.

2.3 The New Norwegian Ground Truth (1800-1900)

Our gold standard ground truth, per the time of writing, consists of parts of En Glad Gut (1868) by Bjørnstjerne Bjørnson, Symra (1863) by Ivar Aasen, Ferdaminni fraa Sumaren 1860 (1871) by Aasmund Olavsson Vinje, Mordet paa Maskinbygger Roolfsen (1840) by Maurits Christopher Hansen and En stemme fra England (1814) by Anders A. Feldborg.

The ground truth consists of about 167 pages, roughly 36k tokens in total.

As this is an iterative process new book pages will be successively added.

3. A Neural Network-based Tesseract OCR Engine for Norwegian Fraktur

3.1 Transfer Learning with Tesseract Models

Recently, transfer learning has been applied to OCR [2,8]. A somewhat different kind of transfer learning is possibile with Tesseract. In our cross-lingual transfer learning approach, we take the best existing (Tesseract) fraktur OCR engine which to our knowledge is the German model. This is then feed to Tesseact's training procedure together with our new ground truth.

One of the main motivations for using this approach is the fact that Norwegian fraktur contains the letters æ, ø and ˚a. A draw is the fact that our set of possible characters is somewhat overdimensioned.

3.2 Neural Network Design

There are also several attested architectures that work well for text recogni- tion, and the current state-of-the-art seems to be convolutional recurrent neural network (CRNN) (see [5]). Our initial model is akin to this aforementioned. Af- ter the input layer we have a 3x3 convolution which output 16 feature maps which then goes into a max pooling layer of similar dimansionality. After the convoluting we have four long short term memory layer where the two first are feed-forward, while the two last are feed-backward. The final layer is a fully connected one.

4. Re-OCRing the 19th Century

4.1 The Public Domain Books Corpora

The Public Domain Books Corpora is a corpora of books that no longer is pro- tected by copyright. The corpora are OCRed with ABBYY FineReader and is attributed with a OCR 'confidence.' We sampled all books more than one standard deviation below the average confidence and thereafter randomly selected a book to serve as an evaluation text.

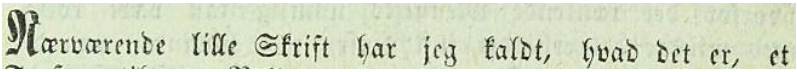


Fig. 3. Facsimile of first line in the evaluation data.

4.2 Evaluation

We evaluate our model on the randomly selected book from the corpora and this in principle could be included as training since this data also had to be corrected/checked by a human annotator. As our evaluation metric we used bag- of-words F1-measure which we compute with PRImA TextEval 1.5. For a whole page of which 3 is the first line, the original ABBYY FineReader measure was found to be 0.28, while our new

fraktur model gives 0.93 which is a substantial improvement with the original scan as Jrftorenbe Wt frift bar jeg falbt, l / oab bel er, and the new Mproarende lille Skrift har jeg kaldt, hvad det er, et. This shows that our new model indeed is superior, but, still, in need of fine-tuning.

5. Conclusion

In the present work we have presented a work flow for establishing a ground truth for historical prints in Norwegian fraktur. The ground truth will be made available. We have also trained a model for Norwegian fraktur that also will be made available for the general public. Lastly, we have begun re-OCRing The Public Domain Books Corpora that will increase and ease the scholarly works of those who work with old Norwegian fraktur prints.

A guiding ideal principle for the present work has been transparency. The open-source movement should also include DH and an important part of that is to be open about methods and data.

5.1 Further Work

As was evident in the sample in 3 there is still work to be done. Now that we have an open model for fraktur we will work towards a pipeline that can handle ornamental initial letters as well as different page formatting such as those found in news papers. We will also try to blend several font models and work towards what might be called a ‘font agnostic’ model, and we will train models with other systems than Tesseract like OCRpus and Calamari.

References

1. Evensen, N.: OCR-behandling av tekst i fraktur. Unpublished manuscript. Univer- sity of Oslo.

2. Reul, C., Wick, C., Springmann, U., and Puppe, F.: 2017. Trans- fer learning for OCRopus model training on early printed books. https://arxiv.org/pdf/1712.05586.pdf

3. Tanner, S., Muoz, T. and Ros, P. H.: 2009. Measuring Mass Text Digitization Quality and Usefulness. D-Lib Magazine 15 (7/8): 10829873. http://www.dlib.org/dlib/july09/munoz/07munoz.html

4. Kettunen, K. and Koistinen, M.: 2019. Open Source Tesseract in Re-OCR of Finnish Fraktur from 19th and Early 20th Century Newspapers and Journals – Collected Notes on Quality Improvement. in Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries Copenhagen, March 6-8 2019. CEUR Workshop Proceedings, no. 2364, pp. 270-282, 4th Digital Humanities in the Nordic Countries, Copenhagen, Denmark, 06/03/2019.

5. Baoguang Shi, Xiang Bai and Cong Yao.: 2017. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition.

6. Springmann, U., and Ldeling, A. 2016. OCR of Historical Printings with an Appli- cation to Building Diachronic Corpora: A Case Study Using the RIDGES Herbal Corpus. https://arxiv.org/pdf/1608.02153.pdf

7. Strbel, P. H., and Clematide, S. 2019. Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images. DH2019, Utrecht.

8. He, Y., Yuan, J., and Li, L. 2018. Enhancing RNN Based OCR by Transductive Transfer Learning From Text to Images. Thirty-Second AAAI Conference on Arti- ficial Intelligence.

ID: 104
Long paper presentations
Topics: historical studies, linguistics, crowdsourcing, cultural heritage collections, data mining / text mining, data modeling / knowledge representation, natural language processing, political science, computational science
Keywords: parliamentary proceedings, every man's rights, allemansrätten, semantic tagging, Finnish parliament

Digging Deeper into the Finnish Parliamentary Protocols ï Using a Lexical Semantic Tagger for Studying Meaning Change of Everymanô s (Allemansrätten)*

Kimmo Kettunen[1], Matti La Mela[2]

[1]University of Helsinki, National Library of Finland, Finalnd; [2]Aalto University, Semantic Computing Research Group, Finland

This paper analyses the protocols of the Finnish parliament 1907–2000. They have been digitised and published as open data by the Finnish Parliament in 2018 . In the analysis we use a novel tool, a semantic tagger for Finnish – FiST [1]. We describe the tagger generally and show results of semantic analysis both on the whole of the parliamentary corpus and on a small subset of data where everyman's rights (a widely used right of public access to nature) have been the main topic of parliamentary discussions. Our analysis contributes to the understanding of the development of this "tradition" of public access rights, and is also the first study utilizing the Finnish semantic tagger as a tool for content analysis in digital humanities research. Keyword search shows first that that the discussion of everyman's rights has had three different peak periods in the Finnish parliament: 1946, 1973, and 1992. Secondly, the contents of the discussions have different nature for all the periods, which could be clearly detected with FiST and keyness analysis.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 177
Short paper presentations
Topics: historical studies, history and theory of digital humanities
Keywords: First World War, ANNO, digitized newspaper collections, historical newspapers, Newseye

Can Umlauts Ruin Your Research in Digitized Newspaper Collections? A NewsEye Case Study on óthe Dark Sides 1918)*

Barbara Klaus

University of Innsbruck, University of Vienna, Austria

Digitized newspaper collections facilitate the access to historical newspapers. Even though they offer several useful possibilities regarding the research in historical newspapers and magazines, the (automatic) research in these collections is (still) full of limitations and pitfalls. Based on the research conducted on the platform AustriaN Newspapers Online (ANNO) for the NewsEye case study 'the dark sides of war', the main challenges of working with digitized newspaper collections will be discussed in this paper. Especially two aspects – the fire catastrophe at the munitions factory Wöllersdorf (1918/09/18) in Lower Austria and the Austrian press coverage about war widows during the First World War – will be used as specific examples. The discussed limitations include the Optical Character Recognition (OCR) quality, provided search options and metadata, as well as others. Furthermore, possible improvements regarding these challenges, e.g. Optical Layout Recognition (OLR), Named-entity Recognition (NER) and Named-entity Linking (NEL), will be presented in this paper.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 200
Short paper presentations
Topics: literary studies, interdisciplinary collaboration, computational science
Keywords: literature, data analysis, growing up, transition books

In Quest of Transition Books*

Denis Kotkov[1], Kati Launis[2], Mats Neovius[1]

[1]Åbo Akademi, Finland; [2]University of Turku, Finland

Literature read by a person not only reflects, but also affects that person. In fact, certain books (transition books) might trigger this process of becoming interested in grownup's literature and therefore mentally becoming a grownup. In this paper, we detect books that are likely to be transition books or transition book candidates based on a loan dataset provided to us by Vantaa City Library. With four methods applied to this dataset we show what books and why are likely to be the candidates. We found the following candidate books: Tähtiin kirjoitettu virhe by John Green, Punainen kuin veri by Salla Simukka and Luukaupunki by Cassandra Clare. Our findings also indicate a few other books that are less likely, but still good candidates for transition books.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 230
Short paper presentations
Topics: folklore and oral history, GLAM: galleries / libraries / archives / museums, open data
Keywords: digital archive, folklore, mapping, open data

Digital Mapping: Research Method and Data Representation Tool in garamantas.lv

Sandis Laime

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

Archives of Latvian Folklore (ALF) at the Institute of Literature, Folklore and Art (ILFA), University of Latvia is the main institution in Latvia with the aim of collecting, archiving, publishing and doing research on Latvian folklore. Since 1924 when the ALF was established it has become one of the biggest folklore archives in Europe and an important resource of records of Latvian traditional culture. The advancement of digital technologies provides new possibilities of structuring and analysing the archival data, on the one hand, and new ways of presenting the contents of the archive and the research results to general public, on the other hand. In 2014 an ambitious project of digitizing the contents of ALF and making it available online (garamantas.lv) was started in order to promote the accessibility of traditional culture resources in the digital environment both for the professionals of the field and the general public. The digital archive provides not only the possibility of storing the data but also analysing it. No other institution of the Latvian humanities is currently offering such a possibility to geocode, analyse and visualise the humanities sources.

In my paper I will focus on the mapping tool of the digital archive garamantas.lv and concentrate on the following aspects:

-) Digital mapping as a research tool and method. Mapping of folklore sources is one of the basic methods used in folkloristics and developed since the emergence of historical-geographical method or so called Finnish school in folkloristics. Tradition-geographical approach is also an important

methodological tool in research related to the history and historical typology of specific folkloristic / cultural phenomena. Although the method is not very often used nowadays, the possibilities of digital mapping have revived the mapping method with many interesting projects already finished and data made available in open access resources.

-) Digital infrastructure of garamantas.lv related to digital mapping. The infrastructure of the digital archive related to the mapping process and storage of the geocoded data consists of two main elements:

1) the mapping tool which aims to enable mapping of the metadata of the humanities sources included in garamantas.lv digital archive; quantity and quality analysis of the mapped geospatial information and visualisation of the results (the archive uses the basic visualisations – markers, heatmaps, clusters and circle charts); mapping tool is made available only for registered users and editors of the digital archive in the administrator environment of garamantas.lv;

2) to ensure storage of the mapped data in a single database wherein the information related to particular places is aggregated (alternative names, address, geographical coordinates, description, photographs etc.) along with the relations of the place with the data included in the other databases of garamantas.lv (persons, organisations, folklore and literary texts, audio and video recordings, illustrations). This enables examination of the diachronic complexity of those particular places. To date, in garamantas.lv digital archive, these digital tools are available to process the text corpus. Besides, guidelines have been developed and described for the editors of the digital archive to use those digital tools; the methodology has been standardised for using the mapping tool and entering of data in the Places database. A double-check system has been elaborated to ensure the quality of the data which is made available in open access database.

-) Map as a way of data structuring and representation to general public. The main principle of every digital archive is to provide the access to the collections in user-friendly and

structured way making the data easy searchable and really accessible. One of the ways of structuring data is according to the geospatial principle – if this kind of information is available. Depending on the specific needs of the researchers and general public there are two options to retrieve the geocoded data in open access digital archive garamantas.lv:

1) The universal map (http://garamantas.lv/en/map/index) is a map-based data retrieval option giving access to the whole corpus of geocoded localities contained in the database of places. The universal map provides advanced search options by different criteria: 1) by data repository, 2) by collection and 3) by database. If any of seven databases is chosen for data exploration it provides further database-specific data selection options. The search engine of the universal map is designed so that data retrieval is possible on both on very general and very specific subjects depending on the combination of selected search criteria.

2) Maps are also included in all databases of garamantas.lv containing geospatial data.

Garamantas.lv supports four basic visualisation options of the quantitative geocoded data: (1) clusters, (2) heat maps, (3) markers and (4) circle charts. The quantity of data can be retrieved by choosing one of the two options: 1) by place as the main principle of quantity and 2) by the quantity of related data of the place.

The mapping tool and the database of places has been elaborated within two projects carried out in ILFA: ERDF project “Empowering knowledge society: interdisciplinary perspectives on public involvement in the production of digital cultural heritage” (No. 1.1.1.1/16/A/040, project leader Sanita Reinsone) and the ERDF postdoctoral project “Latvian Folk Narrative Research: Elaboration of Geospatial Data Analysis Tool and Online Legend Motif and Type Index” (No. 1.1.1.2/VIAA/1/16/193, project leader Sandis Laime).

ID: 243
Poster
Topics: communication studies, linguistics, corpus linguistics, data mining / text mining, digital resources – publication and discovery, natural language processing, computational science, big data
Keywords: web-as-corpus, web data, web genre identification, online registers, online language resources

Towards Better Structured Online Data with the Project ñNews, Opinions or Something Else? Modeling Text Varieties in the Multilingual Internetò

Veronika Laippala[1], Saara Hellström[1], Sampo Pyysalo[1], Liina Repo[1], Samuel Rönnqvist[1], Anna Salmela[1], Valtteri Skantsi[1,2]

[1]University of Turku, Finland; [2]University of Oulu, Finland

Introduction and Objectives

The Internet has brought revolutionary potentials for many fields benefiting from textual data. The masses of text englobe new ways of writing and present unprecedented possibilities to explore, e.g., language, communication and culture (Berber-Sardinha, 2018; Biber and Egbert, 2019). Furthermore, thanks to the billions of words of data available online, the quality of many NLP systems, such as machine translation, can be improved tremendously (Tiedemann et al., 2016; Srivastava et al., 2016). Importantly, almost anyone can write on the Internet. Therefore, the web provides access to languages, language users, and communication settings that otherwise could not be studied.

Despite the potentials, the use of web data is currently very restricted. Above all, the diversity of different kinds of texts on the web imposes serious challenges. Currently, all the texts have a similar status in the web language resources, and there is no information on the origins of the texts, or, specifically, on their situational and communicative specificities - on their register (Biber 1988; Biber and Conrad 2009). Lack of understanding of register may lead to wrong conclusions about the text, as we do not know how to interpret it (see, e.g., Koplenig, 2017). For instance, we read news, discussion forum

messages and encyclopedia articles very differently. Furthermore, register is one of the most predictors of linguistic variation (Biber 2012) and critically impacts NLP: for example, methods developed to process legal texts perform poorly on texts from social media (Webber, 2009) Register would thus offer important information to develop web data from masses of raw text to organized collections that can serve specific purposes and research questions.

This poster presents the newly started project News, opinions or something else? Modeling text varieties in the multilingual Internet running at University of Turku, Finland. In this project, the objective is to analyze and characterize the full range of registers found on the Internet, and to develop a system that could automatically detect them from online language resources. As a practical outcome, the project applies the developed system to detect registers from Universal Parsebanks (UP), a collection of web corpora we have compiled in our research group in previous research. The project focuses on the French, English, and Swedish UP collections, as well as on the Finnish Internet Parsebank, which can be referred to as UP Finnish.

Raw Data

The raw online data analyzed in this project comes from Universal Parsebanks (UP), which is a collection of billion-word automatically collected web corpora, developed in the previous projects of our research group and widely used by linguists, lexicographers, and NLP researchers, among others (see Zeman et al. 2017). UP includes 64 languages and almost 100 billion words. The most frequently used of the language-specific collections is Finnish Internet Parsebank (Luotolahti et al. 2015), which was originally collected in a project funded by Kone foundation. UPs are freely usable online at http://bionlp-www.utu.fi/dep_search/. As a result of the current project, register information will be added to the UP collections in French, English, Finnish and Swedish.

How to Know what Registers Include?

As the project objective is to model the full range of registers found in the Internet, a key question is what these registers are. To this end, we profit from the online register taxonomy developed by Biber et al. (2015) for the Corpus of Online Registers of English (CORE). CORE is based on a near-random sample of the English-speaking web, and manually annotated for registers by four coders. The register taxonomy is created in a data-driven manner to cover the full range of register variation found on the web.

The CORE taxonomy is hierarchical and consists of altogether eight main registers and ~30 sub-registers. These are described below in Table 1.

Narrative	News reports/News blogs, Sorts reports, Personal blog, Historical article, Short story / Fiction, Travel blog, Community blog, Online article
Informational Description	Description of a thing, Encyclopedia articles, Research articles, Description of a person, Information blogs, FAQs, Course materials, Legal terms / conditions, Report
Opinion	Reviews, Personal opinion blogs, Religious blogs/sermons, Advice
Interactive discussion	Discussion forums, Question-Answer forums
How-to/instructional	How-to/instructions, Recipes
Informational persuasion	Description with intent to sell, News+Opinion blogs/Editorials
Lyrical	Songs, Poems
Spoken	Interviews, Formal speeches, TV transcripts

Table 1. Register taxonomy applied to model the linguistic variation on the Internet.

We use the CORE taxonomy to create similar, manually register-annotated corpora for the other project languages, Finnish, Swedish and French. These will be described in the next section.

Current Status: Creating Multilingual Online Register Corpora to Allow Register Detection

We are currently manually annotating registers in samples of the Finnish, Swedish and French UP collections. The objective is to develop similar corpora than CORE for these languages.

So far, we have made the most progress with FinCORE, the Finnish register collection. The first version of FinCORE consisting of 2,200 documents was already used in a study on cross-/multilingual register detection (Laippala et al. 2019). Currently, FinCORE covers 7,500 documents, of which 1,100 have been annotated as machine translations or machine-generated texts.

For French and Swedish, we have started the register annotations with some hundreds of texts annotated. As a positive outcome of this, we can already say that the CORE taxonomy suits relatively well the registers found in these languages. However, the distribution of registers seems to vary, which may complicate their cross-lingual modeling.

Once the annotations are ready, we can start to analyze the registers and develop classifiers to identify them in the raw UP datasets. We have already done experiments on the English CORE using Bert (Devlin et al. 2018). The task is relatively difficult, because the corpus consists of a near-random sample of the web, and the documents are not selected in the corpus to represent a certain category. However, the preliminary results are encouraging. In the poster, we will present the performance of the classifier.

References

Biber, D. 2012. Register as a predictor of linguistic variation. Corpus Linguistics and Linguistic Theory 8, 9–37.

Biber, D. and Egbert, J. 2018. Register variation online. Cambridge: Cambridge University Press.

Biber, D., and Conrad, S. 2009. Register, genre, and style. Cambridge: Cambridge University Press.

Koplenig, A. 2017. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets–Reconstructing the composition of the German corpus in times of WWII. Digital Scholarship in the Humanities, 32(1), Pages 169–188.

Laippala, V., Kyllönen, R., Egbert, J., Biber, D. and Pyysalo. S. 2019. Toward Multilingual Identification of Online Registers. Proceedings of Nordic Conference on Computational Linguistics (NoDaLiDa), September 2019.

Luotolahti J, Kanerva J, Laippala V, Pyysalo S. and Ginter, F. 2015. Towards Universal Web Parsebanks. Proceedings of the Third International Conference on Dependency Linguistics, Depling 2015. Uppsala: Uppsala University.

Srivastava, A., Rehm, G., & Sasaki, F. 2017. Improving Machine Translation through Linked Data, The Prague Bulletin of Mathematical Linguistics, 108(1), 355–366.

Tiedemann, J., F. Cap, J. Kanerva, F. Ginter, S. Stymne, R. Östling, and M. Weller-Di Marco 2016. Phrase-based SMT for finnish with more data, better models and alternative alignment and translation tools. Proceedings of the First Conference on Machine Translation, Volume

ID: 112
Poster
Topics: literary studies, big data
Keywords: reading culture, literary studies, library loan data

LIBDAT: Towards a More Advanced Loaning and Reading Culture and Its Information Service

Kati Johanna Launis, Erkki Sevänen

University of Eastern Finland, Finland

Our presentation is based on the interim results produced by the consortium LibDat: Towards a More Advanced Loaning and Reading Culture and its Information Service (2017–2021, Academy of Finland). The researchers in this consortium come from the University of Eastern Finland, Åbo Akademi University, the Technical Research Centre of Finland and Vantaa City Library in the Helsinki metropolitan area. The project not only works with the digitally born library loan data having been collected by Vantaa City Library since 2016 (seel also Neovius et al.2018), but also utilizes the digital data collected by the joint Helsinki Metropolitan Libraries (HelMet libraries, 17 million loans yearly).

The project consists of three tasks. First, we ask what kind of picture of current Finnish reading culture this data mediates. Second, we clear up the ways in which Finnish public libraries' information services function and how they can be elaborated on. Third, we develop concrete methods by which scholars can analyse and interpret the huge and mainly quantitative data concerning libraries' loans. We wish to be able to show how large digital material, computational methods and literature-sociological research questions can be united in the study of literary culture. The presentation at hand focuses on the first task. consequently, we attempt to show how digitally preserved huge data material can change and deepen our understanding of current literary culture.

Our data indicates that a clear change has occurred in Finnish reading culture since the 1970's and 1980s. In her well-known studies, Suomalaiset kirjanlukijoina (The Finns as Book

Readers, 1979) and Lukijoiden kirjallisuus Sinuhesta Sonja O:hon (Readers' Literature from Sinuhe to Sonja O, 1990), Katarina Eskola concluded that, in the 1970s and 1980s, Finnish readership was characterized by the uniformity of literary taste and the popularity of the "national classics". By "national classics" he meant mainly male authors such as Eino Leino, Mika Waltari, Väinö Linna and Kalle Päätalo which were central figures in the 19th and 20th centuries' Finnish literary culture. In Eskola's studies, a clear majority of Finnish readers named these authors as their favorite authors. Thus, both male and female readers were fond of them.

As we have shown earlier on the basis of the same loan data (Launis et. al 2018), Finnish reading culture IS CURRENTLY fragmented. The commonly known, widely read "classics", that is, books belonging to the literary canon of Finnish literature, no longer attract library users, and the literary taste of library users is more heterogeneous. A striking feature in current Finnish readership is the dominance of women. IT is middle-aged female readers who today maintain literary culture in Finland: between July 2016 and October 2017 there were about 1.5 million loans of fictive literature in Vantaa City library and 76% of these fiction loans were done by women. During this period, the most popular fiction book among women was an entertaining domestic historical novel Ruokarouva (2016, "The Housekeeper"), written by the popular female author Kirsti Manninen under the pen name Enni Mustonen. Loaners also favor novels published in series. In contrast, the young loaners between 15–19 years (mainly girls, 75% of all book loans) favor new translated Anglo-American Young Adult Fiction (John Green, Estelle Maskame), published also in series and adapted for film or television. They also read authors who are active in the social media (YouTube, Twitter) (Launis & Mäkikalli, forthcoming).

Thus, there is a clear difference between the literary taste of Finnish readership in the 1970's and today, as well as a generation difference between the adult and young readers today: middle-aged women prefer domestic entertaining historical fiction, whereas young female readers are fond of

translated Anglo-American YA-fiction. Even though the uniformity of the reading culture has turned into the diversity of it, some features in Finnish literary taste seem to be quite permanent. The depictions of Finnish history, narrated in a realistic manner and depicting hard work and the countryside (such as Ruokarouva mentioned above), still seem to tempt readers. On this part, our results are in line both with Kimmo Jokinen's (1997) and Katarina Eskola's (1979; 1990) studies. Those studies, in particular, Jokinen's Suomalaisen lukemisen maisemaihanteet (The Ideal Landscapes of Finnish Reading, 1997) emphasize that Finnish readers are fond of books that describe our common history and social world in a realistic manner that avoids form experiments and artistic inventions. Likewise, library users quite much favor the brand new first-rate domestic Finnish fiction, for example, the winners of the annual literary prize (Finland-prize).

In earlier studies of Finnish reading culture, methods such as interviews and queries have been widely used. Since that date, the attempts to introduce quantitative methods into the study of literary culture have been hampered by the lack of suitable data. The situation has changed radically along the rise of the digital humanism: nowadays big data – e.g. library loan data used in the LibDat-project – constitutes a different, significant resource for understanding literary culture from a new and wider perspective, for reading it distantly (cf. Moretti 2000). At the moment, we are applying the social network analysis, as well the clusters analysis to the library loan data. Analysis based on the integration of large "born-digital" material, new computational methods and literary-sociological approach open a possibility for posing new questions in the humanities. By uniting quantitative analysis and qualitative interpretation, this sort of research is able to reveal new features in literary culture.

Our project started in 2017. Currently, we attempt to integrate a truly comparative element into it. In this sense we have, for example, gOt in contact with French, American and Australian scholars who work also with comparable digital data material. It will be interesting to see, in what respects other countries'

reading cultures have changed during past decades and how much their reading cultures resemble current Finnish reading culture.

References

Eskola, Katarina (1979). Suomalaiset kirjanlukijoina. Helsinki: Tammi.

Eskola, Katarina (1990). Lukijoiden kirjallisuus Sinuhesta Sonja O:hon. Helsinki: Tammi.

Jokinen, Kimmo 1997: Suomalaisen lukemisen maisemaihanteet. Jyväskylä: Jyväskylän yliopisto.

Launis, Kati, Eugene Cherny, Mats Neovius, Olli Nurmi & Mikko Vainio 2018: Mitä naiset lukevat? Kirjallisuudentutkimuksen aikakauslehti Avain 4/2018, 4–21.

Launis, Kati & Aino Mäkikalli (2019, forthcoming): Mitä tehdä, kun Shakespeare ei vloggaa eikä Waltari twiittaa? Koulu, kirjasto ja nuorten uudistuvat lukemiskulttuurit.

Moretti, F. (2000/2013). Distant Reading. London & New York: Verso.

Neovius, Mats, Kati Launis & Olli Nurmi 2018: Exploring Library Loan Data for Modelling the Reading Culture: Project LibDat. Proc. of Digital Humanities in the Nordic Countries 3rd

Conference, Helsinki, Finland, March 7–9, 2018, CEUR-WS.org, online: ceur-ws.org/Vol-2084/short13.pdf

ID: 146
Short paper presentations
Topics: historical studies, cultural heritage collections, digital resources – publication and discovery, linked data / semantic web / ontologies
Keywords: linked open data, biographical dictionary, digital humanities, semantic web, named entity recognition

Linked Open Data Service about Historical Finnish Academic People in 1640ï 1899*

Petri Leskinen[1], Eero Hyvönen[1,2]

[1]Aalto University, Finland; [2]University of Helsinki, Finland

The Finnish registries “Ylioppilasmatrikkeli”' 1640–1852 and 1853–1899 contain detailed biographical data about virtually every academic person in Finland during the respective time periods.

This paper presents first results on transforming these registries into a Linked Open Data service using the FAIR principles.

The data is based on the student registries of the University of Helsinki, formerly the Royal Academy of Turku, that have been digitized, transliterated, and enriched with additional data about the people from various other registries.

Our goal is to transform this largely textual data into Linked Open Data using named entity recognition and linking techniques, and to enrich the data further based on links to internal and external data sources and by reasoning new associations in the data. The data will be published as a Linked Open Data service on top of which a semantic portal “AcademySampo” with tools for searching, browsing, and analyzing the data in biographical and prosopographical research are provided.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 207
Short paper presentations
Topics: corpus linguistics
Keywords: general corpus, text selection criteria, balance

The 10M Balanced Corpus of Modern Latvian (LVK2018)

Kristīne Levāne-Petrova

The Institute of Mathematics and Computer Science, University of Latvia, Latvia

Introduction

Nowadays the research of different scientific disciplines would not be possible without the use of corpora, especially a general corpus, that "aims to represent the universe of contemporary language" (Aston & Burnard 1998, 5).

A corpus is used in linguistics to conduct language research, create dictionaries and grammars, in sociology to analyze mass opinion and behaviour and in computer science to develop natural language processing components, such as machine translation, speech recognition and various text taggers.

This abstract presents The Balanced Corpus of Modern Latvian (LVK2018) – a new 10 million representative corpus of contemporary Latvian. It describes the design, composition and text selection criteria of LVK2018. The LVK2018 is available at www.korpuss.lv

Development of the LVK Corpus

The balanced corpus of modern Latvian has been developed in multiple rounds. The history of the LVK series goes back to 2007 when the first 1 million corpus was created. The LVK design, compilation and the text selection criteria were based on the Latvian Language Corpus Conception (Levāne-Petrova 2012). The experience from the designing of other general corpora was taken into account as well. The reviewed list of corpora includes British National Corpus (Burnard 2007; Aston, Burnard 1998), Czech National Corpus (Čermák 2002; Hnátková et al. 2014; Křen et al. 2016), Corpus of

Contemporary Lithuanian Language (Kovalevskaitė 2006; Rimkutė et al. 2010), and others. The same corpus design criteria were also used for the subsequent LVK series. The previous corpus from this series (LVK2013) was released on 2013 with 4.5 million words (Levāne-Petrova 2012). All corpora are morphologically annotated (Paikens 2007; Paikens et al., 2013; Paikens 2016) and with the texts also annotated with metadata. LVK2018 is an extended version of LVK series corpus so it contains all the data from previous corpus releases (Levāne-Petrova 2019).

Design of LVK2018

LVK2018 is designed as general-language, representative and balanced corpus that aims to cover the variety of existing texts in some estimated proportions. Therefore, the corpus contains five different sections:

journalism (60%)

fiction (20%)

scientific (10%)

legal (8%)

parliamentary transcripts (2%).

The corpus proportions are slightly modified from the previous edition of the Corpus, i.e., the Miscellaneous section has been incorporated in the Journalism section, because almost all corpus samples previously included in this section also might be included in the Journalism section, for instance, web articles on different topics, etc.

To cover different magazines and newspapers, subsequently the Journalism section also has been divided into the following subsections: nationwide media (41%), regional media (22%), leisure media (24%), popular science media (13%). This section also has been updated as compared with the previous LVK2013 corpus, for instance, there was a separate category for news articles published online, but now it is incorporated in the nationwide newspapers' subsection, because nowadays there is almost no distinction between the printed press and the online press. Subsequently the sources are categorized by

the genre, not publishing media like in the previous versions of the LVK.

Text Selection Criteria

To ensure quality and diversity and other desired aspects of the developed corpus, multiple text selection criteria were set.

(1) time – texts published from 1991 as this corpus aims to reflect the contemporary Latvian. Although the data sample as from 1991 might be chosen for the Corpus, it is also crucial to cover the last years events in the mass media and other fields. Besides, the sources from the last years are available in the digital format, which means that does not require any additional efforts to digitalize the sources.

(2) availability – the data sample in the digital format will be chosen primary for the corpus, not in printed media. For instance, for the Journalism section we choose the articles from the period 2013–2017 to cover the last years events in the mass media and they were also available in the digital format. The novels in the digital format will be preferably included in the Fiction section instead of the novels in the printed media.

(3) originality – LVK2018 contains just texts originally written in Latvian; therefore, the obvious translations of the different texts into Latvian will not be included in LVK2018. It is clear that some foreign news from the mass media are translations from other languages, but there are no specific criteria set to not include such news or other sources that might be translations or localizations in the Corpus.

(4) diversity – texts should cover as wide range of topics as possible. It is very important criteria for the Journalism and Science sections but applies to the Legal and Parliamentary transcripts sections as well. That means that the sources should be selected from the very different topics like foreign news, sports news, leisure, etc. The sources for the Science section should cover different branches of science like biology, mathematics, linguistics, etc.

text completeness – the selected documents should be included in full length if possible, but there is some exception

to this. To not dominate the terms from the particular source in the whole Corpus, the source may not exceed 5% of the particular section of the Corpus. Thus, novels or PhD thesis that exceed this amount were not fully included in the Corpus. The Corpus contains just the samples of these sources.

(5) uniqueness – in the previous Corpus development stages it was very important to not include the news about one and the same event from different sources in the Journalism section, for example, articles of various journalists about the newly elected President of Latvia in various editions. It is also crucial at this Corpus development stage, but it was even more important not to include any news or parts of the articles that could duplicate, as most portals republish the same news, even in the different time periods.

(6) quality – texts should only contain clean text. There should be methods developed for text quality validation because the text selection for the corpus is not the problem, but the conversion from the source digital format to the corpus text. The text sources that are suitable for the Corpus just to be selected, for instance, the source with many tables, formulas, etc. non-text parts would not be usable for the search queries and therefore these parts of the text will be removed from the particular source.

The text selection criteria also might be changed or updated during the different Corpus development stages. For instance, recently just sources with hard copies were included in the LVK, but nowadays when almost all sources are in the digital format, especially journalism, this text selection criteria is not relevant any more.

Although all of the criteria were taken into account in the development of the corpus, the way the criteria were applied for each section (journalism, fiction, scientific, legal, parliamentary transcripts) were different, besides the data was selected automatically in this corpus development stage, that makes this task even more difficult.

Annotation of LVK2018

LVK2018 metadata schema is fully revised and updated. Multiple metadata fields were standardized and normalized during the revision of the metadata. LVK2018 has three publicly visible metadata fields – unique identifier (id), section and reference. A different reference template was designed for each of the five sections to incorporate all the relevant metadata fields for that sections. For instance, the following metadata fields like author, title/chapter, publishing place, publisher, year are used for the fiction section.

LVK2018 contains morphosyntactic annotation by the IMCS morphological tagger. (Paikens, 2007; Paikens et al., 2013; Paikens, 2016) Morphosyntactic annotations contain PoS tag, lemma and other Latvian specific morphological and syntactic information.

A balanced subcorpus of LVK2018 (10 000 sentences), containing samples of texts from the different styles, domains and subdomains existent in the corpus, is also syntactically manually annotated (Rituma et al., 2019), using hybrid dependency-constituency grammar formalism developed in the previous Latvian Treebank pilot project (Pretkalniņa et al., 2011). Afterwards the hybrid annotation is automatically converted to Universal Dependencies to achieve the cross-lingual compatibility, as well as to provide training data for efficient and robust parsers (Gruzitis et al., 2018).

Availability

LVK2018 has been released in the framework of Latvian National Corpus. LVK2018 is freely available via the corpus query interface NoSketch Engine (Rychlý 2007) at http://nosketch.korpuss.lv/.

How to Quote the LVK?

The corpus material is to be quoted in the bibliography in the following way:

The Balanced Corpus of Modern Latvian – LVK2018 (beta). The Institute of Mathematics and Computer Science, University of Latvia. Riga, 2018. Available at: www.korpuss.lv

Acknowledgements

This work has received financial support from the Latvian Language Agency through the grant agreement No. 4.6/2019-029.

References

Aston, G., Burnard, L. (1997). The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press. Available at: http://corpus.leeds.ac.uk/teaching/aston-burnard-bnc.pdf

Burnard, L. (ed.). (2007). Reference Guide for the British National Corpus (XML Edition). Published for the British National Corpus Consortium by Oxford University Computing Services. Available at: http://www.natcorp.ox.ac.uk/docs/URG/

Čermák, F. (2002). Today's corpus linguistics. Some open questions. International Journal of Corpus Linguistics.

7(2). Amsterdam: John Benjamins, 265–282.

Gruzitis, N., Nespore-Berzkalne, G., Saulite, B. (2018). Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18), Miyazaki: ELRA.

Hnátková, M., Křen, M., Procházka, P., Skoumalová, H.. (2014). The SYN-series corpora of written Czech. Proceedings of LREC 2014. Reykjavík: ELRA, 160–164. Available at: http://www.lrecconf.org/proceedings/lrec2014/pdf/294_Paper.pdf

Kovalevskaitė, J. (2006). Dabartinės lietuvių kalbos tekstynas – 10 metų kaupimo ir naudojimo patirtis. Prace Bałtyszne 3. Język. Literatura. Kultura. Warszawa: Uniwersytetu Warszawskiego, 231–241.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P., Zasina, A. J. (2016).

SYN2015: Representative Corpus of Contemporary Written Czech. Proceedings of LREC 2016. Portorož: ELRA, 2522–2528. Available at: http://www.lrec-conf.org/proceedings/lrec2016/pdf/186_Paper.pdf

Hnátková, M., Křen, M., Procházka, P., Skoumalová, H. (2014). The SYN-series corpora of written Czech. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'14), Reykjavík: ELRA, pp. 160–164. Available at: http://www.lrecconf.org/proceedings/lrec2014/pdf/294_Paper.pdf

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P., Zasina, A. (2016). SYN2015:

Representative Corpus of Contemporary Written Czech. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 2522–2528. Available at: http://www.lrec-conf.org/proceedings/lrec2016/pdf/186_Paper.pdf

Levāne-Petrova, K. (2012). Līdzsvarots mūsdienu latviešu valodas tekstu korpuss un tā tekstu atlases kritēriji. Baltistica VIII priedas. Vilnius: Vilniaus Universiteto leidykla, 89–98. Available at: http://www.baltistica.lt/index.php/baltistica/article/view/2113/2019

Levāne-Petrova, K. (2019). Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. Valoda: nozīme un forma 10. LU Humanitāro zinātņu fakultātes Latviešu un vispārīgās valodniecības katedras rakstu krājums. Rīga: LU Akadēmiskais apgāds. 131–146. Available at: https://www.apgads.lu.lv/fileadmin/user_upload/lu_portal/apgads/PDF/Valoda-nozime-forma/VNF-10/vnf_10- 12_Levane_Petrova.pdf

Paikens, P. (2008). Lexicon-based morphological analysis of Latvian language. Proceedings of the 3rd Baltic Conference on Human Language Technologies (Kaunas, October 2007). Vilnius: Vytautas Magnus University, Institute of the Lithuanian Language, 235–240.

Paikens, P., Rituma, L, Pretkalniņa, L. (2013). Morphological analysis with limited resources: Latvian example. Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA). Linköping: Linköping University Electronic Press, 267–277.

Paikens, P. (2016). Deep neural learning approaches for Latvian morphological tagging. Human Language Technologies – The Baltic Perspective: Proceedings of the Seventh International Conference Baltic HLT 2016.

289. Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press, 136–143.

Pretkalnina, L., Nespore, G., Levane-Petrova, K., and Saulite, B. (2011). A Prague Markup Language profile for the SemTi-Kamols grammar model. In Proceedings of the 18th Nordic Conference of Computational Linguistics, pages 303–306, Riga, Latvia.

Rimkutė, E., Kovalevskaitė, J., Melninkaitė, V., Utka, A., Vitkutė-Adžgauskienė, D. (2010). Corpus of Contemporary Lithuanian Language – the Standardised Way. Human Language Technologies – The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010. 219. Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press, 154–160.

Rituma, L., Saulīte, B., Nešpore-Bērzkalne, G. (2019). Latviešu valodas sintaktiski marķētā korpusa gramatikas modelis. Valoda: nozīme un forma 10. LU Humanitāro zinātņu fakultātes Latviešu un vispārīgās valodniecības katedras rakstu krājums. Rīga: LU Akadēmiskais apgāds. 200–216. Available at: https://www.apgads.lu.lv/fileadmin/user_upload/lu_portal/apgads/PDF/Valoda-nozime-forma/VNF-10/vnf_10- 16_Nespore_Saulite_Rituma.pdf

Rychlý, P. (2007.) Manatee/Bonito – A Modular Corpus Manager. Petr Sojka, Aleš Horák (Eds.): RASLAN 2007 Proceedings (2007). Brno: Masaryk University, 65–70.

ID: 180
Long paper presentations
Topics: communication studies, historical studies, linguistics, sociology, bibliographic studies, corpus linguistics, data mining / text mining, digital resources – publication and discovery, interdisciplinary collaboration, open data, open science, project design / organization / management, software design and development, standards and interoperability, user studies / user needs, big data
Keywords: complexity, data issues, non-standard data, bias, interpretation, workflows

Wrangling with Non-Standard Data*

Eetu Mäkelä[1], Krista Lagus[1], Leo Lahti[2], Tanja Säily[1], Mikko Tolonen[1], Mika Hämäläinen[1], Samuli Kaislaniemi[3], Terttu Nevalainen[1]

[1]University of Helsinki, Finalnd; [2]University of Turku, Finland; [3]University of Eastern Finland, Finland

Research in the digital humanities and computational social sciences requires overcoming complexity in research data, methodology, and research questions. In this article, we show through case studies of three different digital humanities and computational social science projects, that these problems are prevalent, multiform, as well as laborious to counter. Yet, without facilities for acknowledging, detecting, handling and correcting for such bias, any results based on the material will be faulty.

Therefore, we argue for the need for a wider recognition and acknowledgement of the problematic nature of many DH/CSS datasets, and correspondingly of the amount of work required to render such data usable for research. These arguments have implications both for evaluating feasibility and allocation of funding with respect to project proposals, but also in assigning academic value and credit to the labour of cleaning up and documenting datasets of interest.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 239
Poster
Topics: cultural studies, literary studies, data mining / text mining, image processing, artificial intelligence
Keywords: digital humanities, children books, Yolo, CNN.

Object Recognition in Illustrated Children Books: Challenges of Applying Computer Vision Systems

Thomas Mandl[1], Im Chanjong[1], Helm Wiebke[2], Schmideler Sebastian[2]

[1]University of Hildesheim, Germany; [2]University of Leipzig, Germany

Introduction: Historical Children Book Research

Digital Humanities is having a considerable impact on humanities research related to text. Many text mining tools have been developed and are currently being applied to genuine research questions in the humanities. This trend is currently contributing to a larger variety of methods being used. But there has been no comparable paradigm shift in research related to visual material. Digital historical corpora allow the automatic access to images and their analysis in great numbers. This can lead to new innovative research questions and quantitative results. Especially, the analysis of digitized historical books with rich visual materials can be of a great value.

Research on historical children's and youth books has yet not often been the subject of digital humanities (DH) studies. This research requires processing for both text and images. Children books typically contain more images than adult books typically. As a consequence, they are of special interest for an analysis of images. In addition, they form a closed category on the one hand which contains sufficient variety on the other hand [1,2].

Illustrated books have played a significant role in knowledge dissemination. The declining production costs for printed images have led to a growing exposure of more and more people to rich visual resources. Research in this area can

identify trends in the objects depicted. The algorithmic analysis seems promising [3].

Fig. 1. Example of an illustration from the collection: Das Jahr und was es den Kindern bringt! Düsseldorf 1861. Elkan, Bäumer & Co. urn:nbn:de:gbv:084-09110211315

State of the Art

Few researchers have processed large amounts of book images to address issues of style or objects. The HBA data challenge for old books intends to improve algorithms for separating illustrations from text automatically [4].

One experiment in the art domain by Salah & Elgammel is dedicated to classify the painter of artistic work. Such work is highly dependent on the type of paintings in the collection [5]. An approach to identify objects within art work has also been

presented. Similar to our approach, it needs to deal with the domain shift and apply current technology to historic print [6].

A recent project is focusing on research on graphic novels. Current state of the art CNNs are applied to tasks like author identification with very good success [7]. In addition, the processing is aimed at measuring the drawing style of a graphic novel in order to find similar book titles. A study of modern children books based on information available in catalogues has analyzed market structures and book formats [8].

One of the first goals for the research of images in historical children books lies within the production technologies. As a classification problem with few classes, it seems like a challenge which could be solved with current technology. However, detailed analysis of production technology in the 19th century is still a hard task [9].

Data Collection of Children and Youth Literature

Our research is exploiting two collections of mainly German children books that are partly digitalized. The first collection is the Wegehaupt corpus maintained by the Staatsbibliothek in Berlin [10].

The second data collection is based on the Hobrecker collection. This collection of books is archived in the library of the Technical University of Braunschweig. A subset has been digitized and is available online [11].

Both collections are of great interest for cultural research. They contain a rich variety of different genres of children books mainly from the 19th century: e.g. alphabetization books, picture books, biographies, natural history descriptions as well as adventure and travel stories.

Results and Analysis

Convolution Neural Networks (CNN), the recent state of the art technology is known to be very effective in automated feature detection and subsequent classification in many domains [e.g. 12]. In the approach presented in this paper, CNNs have been used as the processing model for classification. First, we

classify whether a page contains an image and locate it on the page. Second, we apply the object recognition technology Yolo to the images and record the recognized object types.

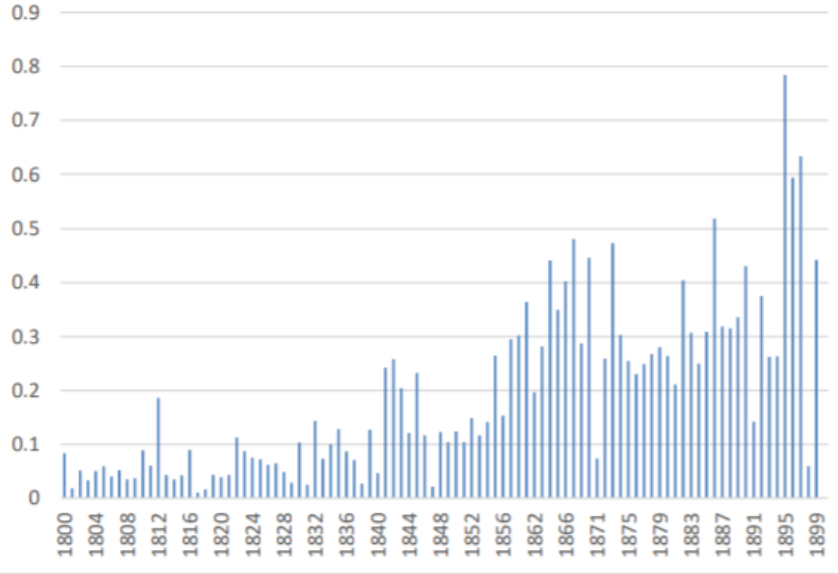


Fig. 2. Average Number of Illustrations per page during the 19th Century in the Wegehaupt corpus.

The first analysis is done on 678 books of the Wegehaupt collection, which is focusing on non-fiction books. Training data was generated by students. The classification is very reliable. This allows a look at the development of illustrations within the 19th century. In the second half, the improvement of technology for printing led to more images per page on average (see figure 2)

The results of Yolo [13] have been recorded for a subset of 321 books from the Hobrecker collection. The classification is not very reliable; a detailed evaluation is ongoing. However, for a statistical analysis, it seems sufficient. We manually classified the fiction books and non-fiction books. The analysis shows

that there is no difference in the overall number of illustrations for the two classes. However, the fiction books contain more images of humans and horses and thus a more limited scope of object classes than is the case in the non-fiction books.

Non-fiction displaying many different objects and animals seem to cause that difference (see table 1).

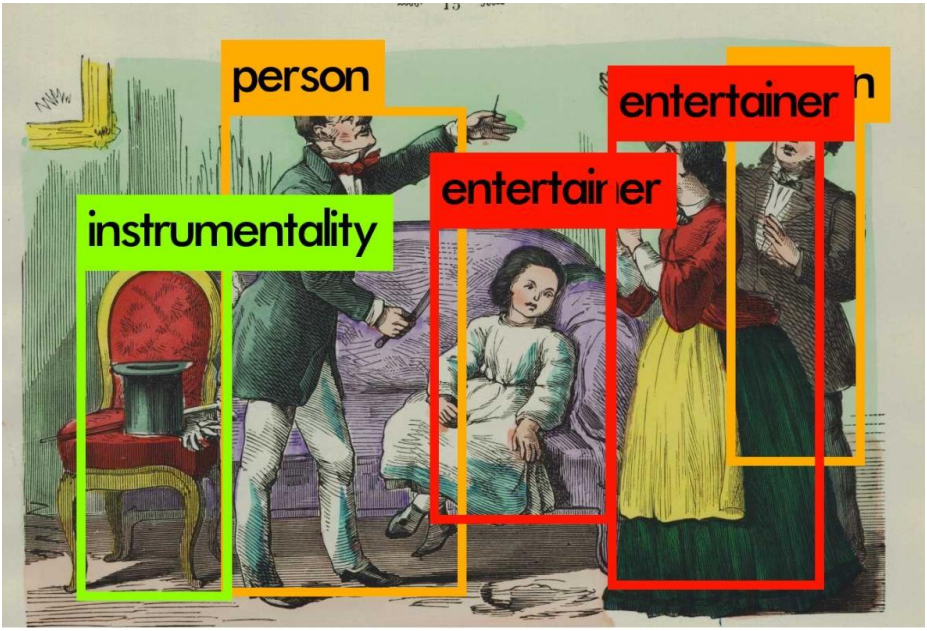


Fig. 3. Example of an illustration from the collection with Yolo results: Der Mummelsack: ein Sittenspiegel der Jugend; Berlin 1875. Kießling. urn:nbn:de:gbv:084-11052415001

Most of the research in image processing is currently being carried out for photographs. But these vary greatly from the non-realistic drawings and illustrations which can be found in children books. It is unclear how well the model transfer actually works. Often unrealistic objects occur like anthropomorphic animals or fairy tale figures or today unknown

objects (e.g. toys). These can obviously not be recognized with systems trained on contemporary photographs.

Median	*Illustr. per page*	*Person per illustr.*	*Horse per illustr.*	*Chair per illustr.*
fiction books	0.400	1.143	0.020	0.0
non fiction books	0.526	0.737	0.042	0.0
Average	*Illustr. per page*	*Person per illustr.*	*Horse per illustr.*	*Chair per illustr.*
fiction books	0.448	**1.446**	**0.103**	**0.072**
non fiction books	0.566	**1.151**	**0.128**	**0.035**

Table 1. Results of Yolo on the Hobrecker collection.

In addition, the type of material also leads to many other challenges. The distribution of the class frequencies is highly skewed. The most frequent classes are humans and a few animals. This does not allow a quantitative tracing of many different motifs through the century.

Outlook

For further analysis, other object recognition systems will be applied. Also transfer learning will be used. For that, a small number of object is labelled and the algorithm is re-trained.

Future work needs to also address stylistic and artistic aspects of the children books. We intend to analyze page patterns and their development over time. A deeper analysis of content on a page and in particular of frequent classes (primarily pictures of humans) offer great potential for advanced analysis tools for digital humanists.

Acknowledgements

We thank the Fritz Thyssen Foundation for their funding for the research project Distant Viewing. We thank the Staatsbibliothek Berlin for providing access to the Wegehaupt collection. We also thank the library of the Technische

Universität Braunschweig for facilitating access to the digitized Hobrecker collection.

References

1. Kiefer, B. (1994). The potential of Picturebooks: From Visual Literacy to Aesthetic Understanding. Prentice-Hall.

2. Schmideler, S. (2018). Lutherbilder. Ein Streifzug durch die Illustrationsgeschichte der Kinder und Jugendliteratur des 18. und 19. Jahrhunderts. Die Reformation in der Kinder und Jugendliteratur.

3. Im, C; Mandl, T.; Helm, W. and Schmideler, S. (2018). Automatic image processing in the Digital Humanities: A pre-study for Children Books in the 19th Century. Picture archives and the emergence of visual history of education. ISCHE 40 pre-conference workshop. “Pictura Paedagogica Online: educational knowledge in images”. Berlin. URN: urn:nbn:de:0111-pedocs-158145

4. Mehri, M.; Heroux, P.; Gomez-Krämer, P. and Mullot, R. (2017). Texture feature benchmarking and evaluation for historical document image analysis, International. Journal on Document Analysis and Recognition (IJDAR), pp. 325–364,

5. Saleh, B. and Elgammal, A. (2016). Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature. International Journal for Digital Art History, (2).

6. Crowley, E. and Zisserman, A. (2014). The State of the Art: Object Retrieval in Paintings using Discriminative Regions. Proc. British Machine Vision Conference. BMVA Press.

7. Dunst, A.; Hartel, R. (2018). Auf dem Weg zur Visuellen Stilometrie: Automatische Genre und Autorunterscheidung in graphischen Narrativen. Kritik der digitalen Vernunft. 5. Tagung „Digital Humanities im deutschsprachigen Raum” http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf

8. Steiner, A. (2019). Conservatism in an Innovative Field: Childrens Digital Books in Sweden. DHN 2019 Digital Humanities in the Nordic Countries 4th Conference. ceur-ws.org/Vol- 2364

9. Im, C.; Ghauri, J.; Rothman, J. and Mandl, T. (2018). Deep Learning Approaches to Classification of Production Technology for 19th Century Books. LWDA, pp. 150–158. http://ceur- ws.org/Vol-2191

10. Staatsbibliothek zu Berlin, Preußischer Kulturbesitz. Wegehaupt Digital: https://digital- beta.staatsbibliothek-berlin.de/suche?category[0]=Kinder- und Jugendbücher&queryString

=project%3A”wegehauptdigital”.

11. UB TU Braunschweig, Hobrecker Kollektion Online. https://publikationsserver.tu-braun-schweig.de/content/collections/childrens_books.xml

12. Goëau, H.; Bonnet, P. and Joly, A. (2019). Overview of LifeCLEF plant identification task 2019: diving into data deficient tropical countries. CLEF Workshop Proceedings. http://ceur-ws.org/Vol-2380/paper_247.pdf

13. Redmon, J.; Divvala, S.; Girshick, R. and Farhadi, A. (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

ID: 238
Poster
Topics: natural language processing, social media, political science
Keywords: content moderation, deep learning, evaluation resources, digital social science

Evaluating, Monitoring and Regulating the Identification of Offensive Content

Thomas Mandl[1], Prasenjit Majumder[2], Sandip Modha[2], Mohana Dave[3]

[1]University of Hildesheim, Germany; [2]DA-IICT, Gandhinagar, India; [3]LDRP-ITR, Gandhinagar, India

Introduction

The amount of offensive and unacceptable content posted online poses a challenge to modern societies. Offensive language and Hate Speech directed from one person to another person and open for others undermines objective discussions. There is a growing need for research on the classification of Hate Speech on platforms of social media without human assistance.

In October 2019, the European Court of Justice decided that platforms need to take down content worldwide even after national decisions. In the particular case the EU court debated defamatory posts on Facebook. Also posts similar in tone need to be addressed and the ruling explicitly mentions automatic systems. This shows that automatic systems are of high social relevance.

Such content moderation can be considered as rating systems according to the European GDPR and as such need to show a certain level of transparency for its users and also for analysis. These algorithms also require monitoring of their quality and legislative regulation. It is necessary to check for potential bias and allow insights into their effects.

The Shared Task HASOC

In order to contribute to research in the identification of offensive content, the HASOC (Hate Speech and Offensive Content Identification in Indo-European Languages) initiative created a testbed from data on Twitter and Facebook.

Datasets were generated for German, English, and Hindi. The dataset provided for training contains 17.000 tweets altogether. The entire dataset was annotated and checked by the organizers of the track. HASOC consists of three tasks, a coarse-grained binary classification task, and two fine- grained multi-class classifications. The main task was the classification of Hate Speech (HOF) and non-offensive content. Example for tweets are shown in the table 1. The use of supervised learning with the annotated dataset is a key strategy for advancing such systems. There has been significant work in several languages in particular for English. The objectives of HASOC are to stimulate research for further languages and compare performance in one language to that in English. Other data collections for hate Speech include GermEval [1] and SemEval [2].

NOT	In case you missed this! You're uniting the country alright Ireland
HOF	America is the only country in history to fight a war to end slavery. Democrats won't fight the predominantly black, Muslim run countries that continue slavery today. Nor do they care about kids made sex slaves by the drug cartels. Democrats are dog

Table 1. Examples for Posts form the HASOC Dataset.

The identification of Hate Speech within a collection or a stream of tweets is a challenge because systems cannot rely solely on the content. Hate text can be about many issues and hate often has no clear signal words and word lists as in sentiment analysis are expected to work less well.

The performance of the best system for identification of Hate Speech for English, Hindi, and German was a Marco-F1 score of 0.78, 0.81 and 0.61, respectively. Most approaches model language as a sequential signal. That means they encode the se- quence within which a word occurs as input.

The approaches used most often were bi-directional Long-Short-Term memory (Bi- LSTM) networks processing word embedding input. The best system was the recently developed BERT (Bidirectional Encoder Representations from

Transformers). Further details and results can be found in the overview article [3].

Performance Analysis and Monitoring

HASOC provides an opportunity to analyze the behavior of algorithms. One core issue is that many experiments deliver very similar performance. We need to explore how much they differ and whether it is possible to combine the output of several systems into one fusion system. Such meta-experiments can also help to monitor and regulate systems. Having several algorithms available is an opportunity for comparative analysis.

However, the distribution of predictions for one document is highly diverse. We have shown that few systems agree on documents. We further developed majority vote runs from experiments submitted at HASOC. We selected five runs of top performing teams. In order to increase diversity of approaches, we selected the best experiment of the five top teams (MajVoteTop). For comparison, we also created majority vote runs for all systems, including the less successful ones (MajVoteAll). By changing the number of required votes for a document, we created different artificial experiments and compared their performance to the best runs. The performance metrics are shown in tables 2 and 3.

	Vito	IITG	QMUL	UIUC	YNU
Recall HOF	0.605	0.642	0.498	0.7023	0.689
Precision HOF	0.64	0.602	0.716	0.5707	0.671
Recall NOT	0.881	0.851	0.931	0.815	0.882
Precision NOT	0.864	0.872	0.841	0.8866	0.89
Weighted F1	0.820	0.784	0.877	0.742	0.827

Table 2. Results of Some Top Runs for HASOC English.

	MajVote3 Top	MajVote2 Top	MajVote30 All	MajVote35 All	MajVote40 All
Recall HOF	0.746	0.645	0.709	0.789	**0.853**
Precision HOF	0.613	0.707	0.639	0.587	0.507
Recall NOT	0.835	0.906	0.859	0.806	0.71
Precision NOT	0.904	0.88	0.894	0.916	**0.932**
Weighted F1	0.768	0.856	0. 877	0.798	0.620

Table 3. Results of Ensemble Runs Created Post Hoc.

It becomes obvious, that concerning the combined F1 metric the top runs cannot be outperformed by the fusion systems presented in table 2. However, for the important value Recall in the class Offensive and Hate, a fusion of all systems delivers the best performance (85%). In a realistic scenario, the preference for metrics needs to be clearly defined. Regulation can demand a better trade-off between recall and precision which fits the need to encountering enough hate posts (recall) while not labelling too many posts which could be acceptable (precision). The analysis of many algorithms can help in finding this trade-off.

Transparency

Hate Speech detection algorithms might unduly limit expressions of citizens, they potentially affect the right for Free Speech and should be carefully monitored and evaluated in order to investigate whether they are free of hidden bias. Within the taxonomy of Lipton [4] the most adequate way to achieve transparency seems to be “Explanation by Example”. Current research also investigates “Post-hoc Interpretability”.

An experiment found that textual description and explanation can improve acceptance of a system [e.g.5]. However, the relation between explanation and system functions are unclear. A future challenge will be the delivery of similar examples which are acceptable for a user. This is difficult due to the intransparency of deep learning systems. E.g. the similarities of some based on word embeddings are hard to interpret for humans. Because word embeddings are developed based on the sequence of words, even words with

very different meaning (good, bad) which appear nevertheless in similar contexts do not have a very low score.

Conclusion

Of course, freedom of speech needs to be guaranteed in democratic societies for future development. Nevertheless, the offensive text which hurts others' sentiments needs to be restricted. As there is such an increase in the usage of abuse on many internet plat- forms, technological support for the recognition of such posts is necessary. HASOC contributes to research in classification. We showed how a benchmark can be used to explore further issues relating to the realistic application of content moderation systems. Such challenges also appear in many other areas in digital social sciences.

References

[1] Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., & Klenner, M. (2019). Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language.

ids-pub.bsz-bw.de/files/9319/Struss_etal._Overview_of_GermEval_task_2_2019.pd

[2] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., ... & Sanguinetti,

M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proc 13th Intl Workshop on Semantic Evaluation (pp. 54-63).

[3] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019). Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th Forum for Information Retrieval Evaluation (pp. 14–17). ACM.

[4] Lipton, Z. C. (2016). The mythos of model interpretability. ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY. arXiv:1606.03490.

[5] Brunk, J., Mattern, J., & Riehle, D. M. (2019). Effect of Transparency and Trust on Acceptance of Automatic Online Comment Moderation Systems. In 2019 IEEE 21st Conference on Business Informatics (CBI) (Vol. 1, pp. 429-435).

ID: 142
Long paper presentations
Topics: historical studies, data mining / text mining, natural language processing
Keywords: intellectual history, international relations, digital text analysis, cultural treaties, methodology

Keeping It Simple: Word Trend Analysis for the Intellectual History of International Relations*

Benjamin G. Martin[1,2]

[1]Uppsala University, Department of History of Science and Ideas; [2]Umeå University, Humlab

In my current research on the intellectual history of international relations, I aim to use digital methods of text analysis to explore conceptual content and change in diplomatic texts. Specifically, I am interested in the sub-set of bilateral treaties explicitly related to cross-border cultural exchange – cultural treaties – some 2000 of which were signed in the twentieth century. What methods and workflows seem most appropriate for this task? Our answer thus far has been to keep it simple. Inspired by recent work by Franco Moretti, Sarah Allison and others, we apply a straightforward form of quantitative word trend analysis, integrated with analysis of metadata about the corpus and tested (and expanded) through full-text searching. By formulating this approach in a specific relationship to the nature of the corpus and the historical questions I want to ask of it, we are able to get quite a lot out of this simple method. In this paper, I describe this approach, share some provisional findings, and offer some methodological reflections.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 197
Short paper presentations
Topics: literary studies, corpus linguistics, data mining / text mining, open data
Keywords: poetry, Russian literature, topic modeling, genre studies, corpus

What Is Russian Elegy? Computational Study of a Nineteenth-Century Poetic Genre

Antonina Martynenko

University of Tartu, Estonia; Institute of Russian Literature, Russian Academy of Sciences, Russia

The presentation will be dedicated to the computational approaches to study a poetic genre, namely Russian elegy. I will try to show that quantitative and computational approaches to poetic corpora will give significant results in studying literary genres in their development.

In the beginning of the 19th century elegies were largely elaborated in Russian poetry as a result of European literatures' influences. While first writings were translations of English, French and Latin elegies [1], since the 1810s poets produced large number of original elegies in Russian language. Thus, the period between late 1810s and early 1830s is considered to be the most important stage in the development of Russian elegy. However, despite the importance of the genre as a whole, most literary scholars had analyzed only canonical elegiac poems (e.g. Pushkin's elegies) and given small attention to the large population of elegies published by minor authors or anonymously [2].

In order to examine the history of the elegy on macro-level, a corpus of Russian poems named as 'an elegy' and published between 1815 and 1835 was compiled for this study [3]. The corpus includes 509 poetic texts and retains punctuation, line division and rhymes according to historical sources. Texts are provided with metadata such as year of publication, bibliographical references, and verse characteristics (meter and number of feet, rhyme scheme). 390 out of 509 elegiac poems were gathered from periodicals, so that they could be dated more precisely than the ones in poetry collections; the analysis below will be based only on well-dated texts. In

addition, a part of the poems in the collection is digitalized and introduced as a research object for the first time. As a result, besides canonic elegiac poems mentioned above, elegies of minor writers are well represented in the corpus (for example, these are poems written by D.P. Glebov, P.A. Pletnev, V.I. Tumanskij, A.S. Norov, I.P. Borozdna, V.N. Grigorjev, and many others). Hence, the elegiac poems gathered in the corpus aim to represent the historical meaning of the genre title more properly than preceding collections of canonical elegies [4].

The corpus metadata overview leads to important conclusions about the authors of the elegies. In the beginning of 1820s, the most remarkable young poets were engaged in writing elegies (particularly, Alexander Pushkin and Yevgeny Baratynsky). This implies that initially the elegy was a promising genre elaborated by renowned Russian poets. However, already in early 1830s this genre seemed to be popular mostly among novice non-professional poets who started to represent themselves as romantic elegists. It is assumed that the latter had great influence both on the decline in the prestige of elegies and on content of the poems itself.

The corpus provides the opportunity to study the content of elegies and their formal features in order to test the hypothesis that elegies had changed significantly during the 1820s. The analysis of lexical frequencies shows that the key notions for the elegy genre are love and melancholy. Then the lexical features of the corpus of elegies were analyzed in comparison with general Russian poetic language of this period (the poetical subcorpus of Russian National Corpus was used as a contrast corpus). The most distinctive words for the elegies were detected using the log odds ratio: these words are nouns that express emotions and abstract notions mostly connected with the theme of love (such as “love”, “tear”, “heart”, “dear”, etc.) and loss (“sorrow”, “sad”, “wither”, “vainly”), as well as the words (collocations) for parting (“the last time”, “everything disappears”, “tears of heart”).

However, the themes detected by the lexical analysis are presented differently in the elegies published in different

periods. As it was mentioned above, the changes could be connected with novice authors' attention to the elegy in the late 1820s. To test the hypothesis of thematic change in elegies between 1810s and 1830s a topic model was created (LDA, R package 'topicmodels') [5]. The model shows that different themes are distributed unequally during the period under consideration. In elegies published between 1815 and 1825 thematic diversity is higher than in late 1820s. For instance, in the end of the 1810s elegies were likely to describe historical events; the importance of the historical theme in elegies is explained by the influence of Napoleonic wars poetry. Also, in the elegies published in the early 1820s pastoral scenes appear more as well as the scenes of mourning one's death. Both these themes are connected with exemplary Latin elegies that, according to the corpus, had lost their influence in the mid-1820s. Based on the distribution of topics in the model, the period between 1825 and 1835 should be described as the emergence of the theme of romantic love. The variety of themes in corpus decreases significantly in the end of the period under consideration and ultimately the elegy became a short love poem close to a madrigal.

The latter conclusion is supported by quantitative analysis of the elegies' formal features. The study of the metrical repertoire shows that decrease in thematic diversity happens simultaneously with the reduction of metrical variations in the corpus. In the elegies published before 1825 number of different meters were used. Above all, these are free iambic verse, iambic hexameter, and iambic verse with regular alternation of iambic hexameter and iambic pentameter ('iamb-65'). However, already in the end of the 1820s more than a half of the poems in the corpus belong only to iambic tetrameter.

Another formal feature worth considering is the length of elegies in lines. Statistical analysis shows that the length of a poem is strongly correlated [6] with the year of publishing: both mean and median lengths of poems aggregated for each year demonstrates significant decrease in elegies' length roughly from 60 to 30 lines during the period from 1815 to 1835 [7].

Thus, the computational study of the elegy genre leads to the following conclusions: between 1815 and 1835 the elegies became more thematically homogeneous and shifted from various meters to iambic tetrameter; at the same time a poem's size have reduced significantly. These findings proven by quantifiable results make visible the processes specific to elegy in 1820s, the last period of massive elaboration of this genre. Moreover, the prepared corpus could be used as training data for further genre classification. So, in conclusion, the results of such classification will be presented. Based on the elegiac features gained from the corpus, poems close to elegies will be extracted from the poetical subcorpus of Russian National Corpus and then compared to existing compilations of elegiac texts.

References

[1] See: Frantsuzskaia Elegiia XVIII–XIX vekov v perevodakh poėtov pushkinskoj pory [Russian Elegy of the 18th and 19th Centuries Translated by the Poets of Pushkin's Time], edited by Vadim Vatsuro and Vera Mil'china. Moscow, 1989.

[2] See, for example, important studies on the development of Russian elegy by Irina Semenko (Semenko, Irina. Poėty Pushkinskoj Pory [Poets of Pushkin's time]. Moscow, 1970) and Vadim Vatsuro (Vatsuro, Vadim. Lirika Pushkinskoj Pory [“Lyrics of Pushkin's time”]. Saint-Petersburg, 1994) both focused on the poetry of Pushkin's closest associates.

[3] The corpus is available on github repository: github.com/tonyamart/rus_elegies (texts' id-s do not correlate with actual number of texts in the corpus).

[4] Cf.: Russkaia elegiia XVIII – nachala XX veka [Russian elegy between the 18th and the beginning of the 20th century], edited by Leonid Frizman. Leningrad, 1991.

[5] The problems regarding the application of LDA topic modeling to poetical corpora were discussed in: Navarro–Colorado, Borja. “On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry.” Frontiers in Digital Humanities, 5:15, 2018.

[6] r = –0.6 and –0.8 for correlation between year and mean / median lengths respectively.

[7] See the similar conclusion about poems' decline in length based on the study of all-genre corpus of Russian poetry: Shelya, Artjom, and Oleg Sobchuk. “The shortest species: how length of Russian poetry changed (1750–1921)”. Studia Metrica et Poetica, 4.1, 2017, 66–84. The reduction of elegies’ lengths, therefore, proves that the corpus of elegies adequately represents the processes that had been happened in poetry of this period.

ID: 225
Poster
Topics: communication studies, data mining / text mining, history and theory of digital humanities, social media, discourse analysis
Keywords: lexicometry, Twitter, politics

Starting Points in French Discourse Analysis: Lexicometry to Tweets*

Marge Käsper, Liina Maurer

University of Tartu, Estonia

In the Nordic countries, French Studies is not probably the first field one would look at when thinking about digital humanities. This, however, might be a mistake. From its very beginning in the 1960s until today a part of what is called the French School in Discourse Analysis has been using various machine-based methods to measure the social impact of words in discourse. Since Michel Pêcheux's (1969) theories about an imaginary automatic tool to detect ideology and the first works in political lexicometry at St. Cloud (Maldidier 1969, Marcellesi 1971), the methods have been discussed, developed and diversified (for these discussions, see Guilhaumou 2002), to create various "textometric" (Salem 1987), "logometric" (Mayaffre 2004) or "ideometric" (Longhi et al 2017) analyses. We will examine some examples of these using Lexico 5, a key tool in the field today, to construct our analysis of a corpus of tweets by the French president Emmanuel Macron (#EmmanuelMacron).

Today some of the most significant work in lexicometry is produced by Damon Mayaffre (2004; 2007, etc.), who has analyzed comparative recurrences and vocabulary patterns in the speeches of all French presidents from de Gaulle to Emmanuel Macron. Mayaffre (2004) points out, for example, the most frequent words of the presidents ("problème" for Giscard d'Estaing, "civilisation" for Pompidou, "naturellement" for Chirac, etc.). He complements the quantitative analysis with a discussion of the significance of these vocabulary "over-

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

uses" on the qualitative level. Mayaffre has written extensively about the nature of the corpora and the methods for exploring them (for example Mayaffre 2007) but the core of his analysis is always based on the speeches of French presidents.

We claim, however, that an important part of today's political communication is performed through social media like Twitter, Facebook, etc. Julien Longhi (2013), indeed, considers the tweet as a sub-genre of political discourse like an interview or a speech and does not see it just as the "reduction of thought to 140 characters". Thus, during the French presidential elections in 2017 Longhi et al have launched a platform called # Idéo2017 to analyze what the candidates say on social networks with the help of lexicometric tools to identify the lexical fields and dominant themes of the different candidates and to establish thus "the linguistic profile" of the candidates and what differentiates them. We will use this platform to test Macron's discursive profile in comparison to other candidates in 2017, and also in comparison to his actual tweets. In this textometric approach it is possible, for instance, to follow, by a "section map" option, the topography of lexemes selected in corpora. In our tweet corpus, we can thus follow the continuity of the lexicon used by Macron in his tweets.

In general, however, the aim of our analysis is not a definitive analysis of Emmanuel Macron's ideological positions but a better understanding of the functioning of political communication. Thus, in a further analysis, we plan to compare the tweet corpus also to a media corpus of comments to analyze (by the detection of repeated segments) the extent to which the tweets form the core of the press commentaries.

References

Guilhaumou, Jacques 2002. Le corpus en analyse de discours : perspective historique. Cor¬pus, 1. [En ligne] URL: http://corpus.revues.org/8.

Longhi J., Marinica C, Hassine N., Alkhouli A., Borzic B. 2017. The #Idéo2017 platform, 5th conference CMC and Social Media Corpora for the Humanities, Bolzano, Italy, 3rd and 4th October 2017 – Conference proceedings, pp. 46–51. halshs-01619236.

Longhi, Julien 2013. Essai de caractérisation du tweet politique, L'Information grammaticale, 136, pp. 25–32. halshs-00940202.

Longhi Julien 2017. Humanités, numérique: des corpus au sens, du sens aux corpus, Questions de communication, 2017/1, 31, pp. 7–17. URL: https://www-cairn-info.ezproxy.utlib.ut.ee/revue-questions-de-communication-2017-1-page-7.htm

Longhi Julien 2018. Tweets politiques : corrélation entre forme linguistique et information véhiculée, in Mercier A. et Pignard-Cheynel N. (dirs.), #Info. Partager et commenter l'info sur Twitter et Facebook, Paris : Editions de la Fondation MSH, pp. 295–314.

Lorriaux, Aude 2017. Le « je » d'Emmanuel Macron. Interview avec Damon Mayaffre. Sciences Humaines. Août-septembre 2017.

Maldidier, Denise 1969. Analyse linguistique du vocabulaire politique de la guerre d'Algérie d'après six quotidiens parisiens. Thèse de doctorat. Disponible sur : http://classiques.uqac.ca/contemporains/maldidier_denise/analyse_linguistique/

ana¬lyse_linguistique.html.

Marcellesi Jean-Baptiste 1971. Éléments pour une analyse contrastive du discours poli¬tique. Langages, 23 « Le discours politique », pp. 25–56.

Mayaffre, Damon 2004. Paroles de président. Jacques Chirac (1995–2003) et le discours présidentiel sous la Vème République. Paris : Champion.

Mayaffre, Damon 2007. L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan : Retour sur les travaux actuels de topographie/topologie textuelle. Lexicometrica, André Salem, Serge Fleury, 2007, pp. 1–12. hal-00551468.

Salem, André 1987. Pratique des segments répétés. Essai de statistique textuelle. Paris : Klincksieck.

ID: 174
Long paper presentations
Topics: historical studies, library & information science, literary studies, stylistics and stylometry / authorship attribution
Keywords: authorship attribution, stylometry, bootstrap consensus network

Exploring the Potential of Bootstrap Consensus Networks for Large-Scale Authorship Attribution in Luxdorph's Freedom of the Press Writings*

Florian Meier, Birger Larsen, Frederik Stjernfelt

Aalborg University Copenhagen, Denmark

Authorship attribution (AA) is concerned with the task of finding out about the true authorship of a disputed text based on a set of documents of known authorship. In this paper, we investigate the potential of Bootstrap Consensus Networks (BCN) – a novel approach to generate visualizations in stylometry by mapping similarities of authorial style between texts into the form of a network – for large-scale authorship attribution tasks. We apply this method to the freedom of the press writings (\textit{Trykkefrihedsskrifter}), a corpus of pamphlets published and collected in Denmark at the end of the 18th century. By conducting multiple experiments, we find that the size of the constructed networks depends heavily on the type of variables and distance measures used. Furthermore, we find that, although a small set of unknown authorship problems can be solved, in general, the precision of the BCN method is too low to apply it in a large-scale scenario.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 118
Long paper presentations
Topics: literary studies, data mining / text mining, information retrieval, natural language processing
Keywords: fractal analysis, Hurst exponent, sentiment analysis, story arcs, text analysis

Implications of Multifractal Theory for Fictional Narratives ï a Dynamic Perspective on Sentiment-Based Story Arcs Exemplified by Ishiguro's Never Let Me Go

Kristoffer Nielbo[1], Qiyue Hu[2], Bin Liu[5], Mads Rosendahl Thomsen[6], Jianbo Gao[2,3,4]

[1]Center for Humanities Computing, Aarhus University, Aarhus, Denmark; [2]Center for Geodata and Analysis, Faculty of Geographical Science,\\ Beijing Normal University, Beijing, China; [3]Institute of Automation, Chinese Academy of Sciences, Beijing, China; [4]International College, Guangxi University, Nanning, Guangxi, China; [5]Business School, Guangxi University, Nanning, Guangxi, China; [6]School of Communication and Culture, Aarhus University, Aarhus, Denmark

The moods, feelings and attitudes represented in a novel will resonate in the reader by activating similar sentiments. It is generally accepted that sentiment analysis can capture aspects of such moods, feelings and attitudes and can be used to summarize a novel's plot in a story arc. With the availability of a number of algorithms that automatically extract sentiment-based story arcs, new approaches for their utilization becomes pertinent. We propose to use nonlinear adaptive filtering and fractal analysis in order to analyze the narrative coherence and dynamic evolution of a novel. Using Never Let Me Go by Kazuo Ishiguro, the winner of the 2017 Nobel Prize for Literature as an illustrative example, we illustrate how: 1) nonlinear adaptive filtering can extract a story arc that reflects the tragic trend of the novel; 2) the story arc displays persistent dynamics as measured by the Hurst exponent at short and medium time scales; and 3) the plot's dynamic evolution is reflected in the time-varying Hurst exponent. We argue that these findings are indicative of the potential that multifractal theory has for computational narratology and large-scale literary analysis. Specifically, that the global Hurst exponent of a story arc is an

index of narrative coherence that can identify bland, incoherent and coherent narratives on a continuous scale. And, further, that the local time-varying Hurst exponent captures variation of a novel's plot such that the extrema have specific narratological interpretations.

ID: 105
Short paper presentations
Topics: communication studies, film and media studies, historical studies, data mining / text mining, political science
Keywords: word distribution, keyword change, swedish politics, hierarchical cluster modeling

Evolving Political Keywords 1945ï 1989: Clustering Word Distributions in 3100 Swedish Governmental Official Reports

Fredrik Norén, Roger Mähler

Umeå University, Humlab, Sweden

The Swedish welfare state is usually associated with a number of political keywords: equality, liberalism, rationalism, internationalism and folkhemmet [the people's home]. Such keywords are by no means static. On the contrary, they tend to change in frequency, and sometimes new terms replace old ones. For instance, some researchers argue that the 1960s and 1970s was a period of radical change that also affected the vocabulary to describe the shifting trends and modes in society (e.g. Bjereld & Demker, 2018). Similarly, some media historians have argued that people during this period saw a need for a new language that could describe the shifting media and information landscape at that time (Hyvönen et al, 2017). Such historical research tends to present changes in language use through qualitative or even anecdotal methods – whether it regards specific ruptures or slowly evolving temporal developments in society. Today, computational methods offer an opportunity to empirically examine the rise and fall of central keywords – as well as to detect new or forgotten ones – on a larger scale (e.g. Guldi, 2018). In our presentation we use this approach in order to examine such previously mentioned research assumptions, and to study changes in the political vocabulary in general, and changes related to media and communication in particular.

In our presentation, which is part of a work in progress case study, we thus focus on changes in frequency of words within the political sphere during the Swedish Post-War era. The case study calculates and clusters word distributions over time in a

political corpus of government reports, with a specific focus on words with a higher degree of fluctuation over time. Our paper aims to answer the following questions: Are there particular time periods when keywords appear and disappear in the political vocabulary in Sweden between 1945 and 1989? If so, when? And more importantly, what previous knowledge can be discarded and confirmed about the political landscape in general, and about media and communication issues in particular, by studying aggregated changes of word distributions over time?

Empirically, we focus on the entire corpora of the Swedish Governmental Report series (Statens offentliga utredningar, SOU) from 1945 to 1989, in total about 3 100 reports. These reports, based on a system of commission inquiries, are used to provide the government with knowledge and alternatives before submitting a proposal for new legislation. The vast diversity of scrutinized topics after 1945 make the series a valuable historical source for broader studies of the Swedish government's view on various political matters, and which keywords that were of importance in order to navigate in the political landscape, as well as for more narrow studies, for example about the political vocabulary concerning issues of media and communication (Norén & Snickars, 2017).

Methodologically, this paper builds on David McClure's work, in which he traces aggregated word distributions across narrative time (from the first page to the last page) in 27 000 American novels (McClure, 2017). But instead of narrative time we focus on word distributions across historical time in 3 100 SOU reports – from 1945 to 1989.

As a first methodological step, we focus on (non-lemmatized) words that appear at least 10 000 in our corpus (cf. McClure, 2017). Due to the bureaucratic genre of our corpus it is expected that many of these words – "report", "scrutiny", "government", "evaluate" and so on – perform a relatively uniform distribution over time. However, since we are interested in changes in the vocabulary over time, and in particular if there were specific periods when such changes did occur, we need to focus on words with a higher frequency

fluctuation across this time period. As our null hypothesis, we assume that a word distribution over time will result in the same normalized frequency value each year, and thus generate a complete flat curve with no fluctuation in frequency across time. We then use the chisquared test to determine the difference between the expected word frequency (i.e. the null hypothesis, with a flat frequency curve over time) and the observed word frequency (i.e. the actual normalized frequency, with a fluctuating trend over time). From this result we limit our following study to a list with words that perform a high variance score (i.e. high fluctuation of normalized word frequency across time).

Then, as a second step to investigate whether some time periods are more sensitive towards change in the political vocabulary, we will cluster our chosen list of words based on how similar their relative frequency distributions are across time (and not by their similarity in meaning). We use basic hierarchical cluster modeling on our list of chosen words with the highest degree of normalized variation. An advantage of hierarchical clustering is that you do not have to base the clustering on too many subjective decisions – just metrics used to compute the distance between the distributions. Here, we will apply different distance metrics to test stability of the hierarchical clustering. We will start with a lower threshold and increase it until the hierarchical model delivers a result with a relatively limited number of word distribution clusters. This will leave us with a better overview of vocabulary changes during the political Post-War period – from the perspective of the Swedish government official report series. The result from the hierarchical cluster model will thus constitute the base for our analytical work: changes in political language across during the Post-War era, and what new knowledge can be generated about the political landscape in general, and about political issues concerning media and communication in particular.

This presentation is part of the research project “Welfare State Analytics. Text Mining and Modeling Swedish Politics, Media & Culture, 1945–1989” (WeStAc), that both digitizes and curates three massive textual datasets – in all almost four

billion tokens – from the domains of newspapers, literary culture, and Swedish politics during the second half of the 20th century.

References

Demker, M. & Bjereld, U. 1968: När allt började, 2018.

Hyvönen, M., Snickars, P. & Vesterlund, P, " The Formation of Swedish Media Studies, 1960–1980", Media History, published online 10 Feb 2017.

Guldi, J. "Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora", Cultural Analytics, published online Dec 2018.

McClure, D. "Distributions of words across narrative time in 27,266 novels" (https://litlab.stanford.edu/distributions-of-words-27k-novels/) and "A hierarchical cluster of words across narrative time" (https://litlab.stanford.edu/hierarchical-cluster-across-narrative-time/), 2017.

Norén, F. & Pelle, S. "Distant reading the history of Swedish film politics in 4500 governmental SOU reports", Scandinavian journa of Cinma, no 2 2017.

ID: 178
Short paper presentations
Topics: cultural studies, library & information science, literary studies, networks / relationships / graphs, visualisation, big data
Keywords: social network analysis, library loan data, book clusters, reading culture

Detecting Social Structures Using Library Loan Data

Olli Nurmi[1], Kati Launis[2], Erkki Sevänen[3]

[1]VTT, Finland; [2]University of Turku, Finland; [3]University of Eastern Finland, Finland

Finland is a country with high PISA rankings and a well-functioning publicly funded, free-of-charge library system. About 80% of the Finnish population use public libraries regularly, and during the last two decades 35–50% of Finnish people have loaned something at least once a year (books, journals, films, cd-records) from the public libraries. Against this background, it can be stated, that Finland has active reading culture. However, radical changes in time use, digitization, as well as attitudes towards reading have influenced our reading habits substantially.

In this article, we study the current Finnish reading culture by analysing the loan data collected by Vantaa City Library in Finland's metropolitan area. In earlier studies of the Finnish readership, methods such as interviews and queries have been widely used (see, for example, Eskola 1979). Since then attempts to introduce quantitative methods into the study of literary culture have been hampered by the lack of suitable data. The situation has changed radically along the rise of the digital humanism: nowadays big data – e.g. library loan data used in this article – constitutes a significant resource for understanding literary culture from a new and wider perspective. Integration of large "born-digital" material, new computational methods and literary-sociological research questions open a possibility to find new knowledge within the qualitative approach in humanities.

Our method is to apply social network analysis to the data concerning public libraries' loan activities in the Helsinki

metropolitan area. Firstly, we draw a co-occurrence network based on the paired presence of books within a specified loan cart. We then apply the modularity maximization method to detect book clusters. Visual representations of book clusters are drawn to reveal associated cultural and literary phenomena. This paper shows that current Finnish reading culture is heterogeneous and consists of several sub clusters. It also shows that the library users favour the newest literature and typically borrow multiple books of the same series and the same writer. Our methodological contribution is to demonstrate how social network analysis and clustering technique can be applied to library loan data to characterize reading culture.

Introduction

Starting from a previous work done in the field of digital studies of cultural trends through quantitative analysis of digitized texts (Michel et al., 2010), we use social network analysis as a method for detecting changes in book reading culture and identifying reading subcultures. In literary research, social network analysis and community detection has been a popular method used to visualize certain structural features of a text or a corpus. A common usage is the visualization of relationships between the texts based on the similarities of the textual contents, and relationships between textual entities such as words (Jänicke, Franzini, Cheema, & Scheuermann, 2015).

In this article, we use the visualization to disclose relationships between the books in the library collection. Firstly, we draw a co-occurrence network based on the paired presence of books within a specified loan cart. We then apply the modularity maximization method to detect book clusters. Visual representations of book clusters are drawn to reveal associated cultural and literary phenomena. This paper shows that current reading culture is a heterogeneous cultural phenomenon consisting of several different sub clusters. The position of national classics (such as Väinö Linna), popular among Finnish readership some decades ago, has radically weakened.

Data Source

The largest public library network in Finland is Helsinki Metropolitan Area Library network (Helmet) consisting of the city libraries of Helsinki, Espoo, Kauniainen, and Vantaa. In this work, we had access to anonymized Vantaa City Library loan data. The Helmet collection, consisting of 3.4 million items, is available for the Vantaa City Library users through this network. Our data sample includes all the loan interactions of Vantaa City Library users during 20th July 2016 – 22nd October 2017 containing about 1.5 million records.

We build our understanding on the library loan data because it gives an accurate, actual and much wider picture that interviewing a limited number of book readers. This work provides a reliable evidence basis for decision-making and development of effective policies in libraries.

Results

The analysis shows that the library users typically borrow the multiple books of the same series and the same writer: four of the six of the largest clusters are formed around contemporary female authors, writing entertaining fiction in series and under a pseudonym. This can be explained by the increased use of branding where a set of marketing and communication methods are applied to distinguish the author from competitors, aiming to create a lasting impression in the minds of the readers. An author brand is, in essence, a promise to its readers including emotional benefits. When readers are familiar with an author's brand, they tend to favour it over competing others.

The type of analysis used in this article, can also facilitate new ways to create book recommendations or place the books in the libraries. The books can, for example, be placed in libraries in clusters, which then may be sorted alphabetically. This facilitates the library users' ability to shift smoothly from one cluster to another when a library user is searching and selecting new books. In addition, book series should be marked to enable the readers to locate them easily.

The analysis may also help obtain the ‘market intelligence’ for a better understanding about the different book genres and subcultures performance and evolution. Several algorithms can be used to calculate the importance of any given node in a network. In libraries’ case, we can use these algorithms to identify books with influence over the whole network. By promoting these influential books, the librarians could increase their effect on the reading culture.

The library collection consists of tens of thousands of books and no one is able to read through them all to get the “whole picture” of the literature available for the loaners and the reading culture based on it. The distant reading (Moretti 2000) of the contemporary reading culture – based on the big, digitized, daily loan data during the 1.5 years – is the method that makes this kind of definition possible. Using data analytics methods and social network analysis we can focus on a manageable piece of information and enable literary scholars to make surprising discoveries, generate new hypotheses or suggest further research.

References

Eskola, K. (1979). *Suomalaiset kirjanlukijoina*. Helsinki: Tammi.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. *Eurographics Conference on Visualization (EuroVis) (2015)*, 1–21. https://doi.org/10.2312/eurovisstar.20151113

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., … Aiden, E. L. (2010). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Moretti, F. (2013). *Distant Reading*. London & New York: Verso.

ID: 115
Short paper presentations
Topics: art history, cultural studies, geography, historical studies, literary studies, cultural heritage collections, geospatial analysis – interfaces & technology, ethnography
Keywords: Arctic, indigenous people, visual, literature, mapping

Arctic Visible: Mapping the Visual Representations of Indigenous Peoples in the Nineteenth-Century Western Arctic

Eavan Fiona O'Dochartaigh

Umeå University, Sweden

This paper presents the data and early digital stages of the new postdoctoral project, ARCVIS (2019–2021), based at Humlab, Umeå University, with the support of the Arctic Research Centre (Arcum) and the Department of Language Studies. The project gathers, contextualizes, and maps over a thousand nineteenth-century visual images of local communities in the western Arctic (primarily Alaska, Canada, and Greenland). The core data includes sketches, paintings, and photographs, as well as engravings and lithographs in periodicals and in published travel narratives created by 'explorers.' Through the display and analysis of picture and text, the project will make visible the indigenous people who were key to the success or failure of expeditions from the south, disrupt the grand narrative of heroic polar exploration, and counteract the popular imaginary of the Arctic as empty and desolate.

The critical focus of the sciences and the humanities pertaining to the Arctic has traditionally been on ice and hostile environments. It is not surprising, then, that the dominant and enduring imaginary of the Arctic in more southern latitudes is of a space devoid of people. Numerous literary studies of the Arctic exist while the rich visual records created by travellers to the region in the nineteenth century have been surprisingly overlooked. Previous studies of the visual representations of the Arctic focus almost exclusively on the public representations, failing to consider the layers of visual archival material that often exist behind or beyond them.

Yet, hundreds of rich visual and textual sources, particularly from this period, attest to a peopled Arctic. In contrast to enduring images of ice, the project will show the Arctic as a peopled environment with a rich history and heritage.

ARCVIS will ultimately make available an open-access online platform. This will contain a geospatial database of visual works designed to encourage use by a general audience (as well as by scholars across disciplines with an Arctic focus). Much of the relevant data is currently held by various archives around the world, primarily outside the Arctic itself. Some of these images are available online but a significant portion cannot be easily accessed and their inclusion in the database will be prioritized. The project will communicate with Arctic NGOs and seek the input of local communities, encouraging a greater sense of ownership over the material. By spatially connecting little-known archival materials (held in repositories worldwide) to their places of origin in the Arctic, the project seeks to virtually 'return' sketches and illustrations to their rightful 'homes.' In this way, the research strives to give 'voice' to the indigenous people who were key to the success or failure of expeditions from the south.

Close analysis of the pictures and their associated texts is informed by the disciplines of literature, art history, and geography, and theoretical frameworks of postcolonialism, semiotics, iconography, and space and place. These combined approaches allow for a more cohesive understanding of the cultural material from Arctic expeditions, ensuring the analysis of the visual, textual, and spatial aspects of the archive simultaneously.

The geospatial database is being designed in a user-friendly and visually appealing manner, using bespoke design, mapping, illustration, and archival material in order to encourage users to explore the material, both deepening their knowledge and increasing their interest in the Arctic. The spatial display of information is critical for the research, particularly for the open-access geodatabase. The power to show the specific places that pictures depict, or originate from, is key to understanding the variation and the differences within

the Arctic environment and will show how certain places became central locations to which expeditions returned repeatedly.

Detailed information will be recorded and created to inform the metadata, which will include the title, artist, date, medium, subject, placename, season depicted, associated text, dimensions, and archival details, as well as what each picture contains. The project will aim to establish the geographic extent or origins of pictures through examination of inscriptions, associated texts, and comparative material. The data will be visualised on modern, historical, and bespoke maps of the Arctic, which will allow simple searches to be performed by zooming in on a location. The material will also be searchable by other parameters such as subject, place, artist, or voyage, making the information relevant to a variety of users. No similar or comparable database currently exists and this trans-disciplinary resource will make the (otherwise remote) archives accessible and relevant for local Arctic communities. By creating visually engaging and interactive online material, the project will extend its potential impact beyond academia.

This paper specifically looks at the process of building the ARCVIS database and outlines the challenges encountered when gathering data from a variety of sources in order to create a new, unified, interrogable, and evolving dataset. Preliminary data, database structure and interface design possibilities, as well as emerging problems, questions, and gaps in the data will also be explored in this paper. Ultimately, the research project investigates the importance of the role of indigenous people for expeditions and interrogates the public representation of the nineteenth-century Arctic as an empty, frozen space to be conquered.

ID: 127
Short paper presentations
Topics: communication studies, linguistics, philology, crowdsourcing, data modeling / knowledge representation, digital resources – publication and discovery, games and meaningful play, interface & user experience design, linking and annotation, user studies / user needs, big data, citizen humanities, citizen science
Keywords: annotation, crowdsourcing, user experience, sentiment analysis

Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task*

Emily Öhman

University of Helsinki, Finalnd

With the prevalence of machine learning in NLP and other fields, an increasing number of crowd-sourced data sets are created and published. However, very little has been written about the annotation process from the point of view of the annotators. This pilot study aims to help fill the gap and provide insights into how to maximize the quality of the annotation output of crowd-sourced annotations with a focus on fine-grained sentence-level sentiment and emotion annotation from the annotators point of view.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 110
Short paper presentations
Topics: linguistics, corpus linguistics, digitisation – theory and practice, image processing, natural language processing
Keywords: dialect texts, Swedish, transcription, handwritten text recognition, phonetic alphabets, digitisation

Handwritten Text Recognition and Linguistic Research*

Erik M. Petzell

Institute for language and folklore, Department for dialectology, onomastics and folklore research in Gothenburg, Sweden

In this talk, I describe my ongoing work with automatic transcription of handwritten Swedish dialect texts from the 19th century, and relate it to my linguistic research on enclitic pronouns in North Germanic. Enclisis is a linguistic phenomenon that balances on the border, as it were, between syntax and morphology. For instance, enclitic pronouns fill syntactic slots just like free pronouns and larger noun phrases. However, clitics are prosodically dependent on another word, in effect being unable to bear stress. In that respect, they are more like inflectional endings than independent phrases.

Enclisis of any kind is hard to investigate in texts, since orthography, both in the past and the present, normally does not mark it. Audio recordings of dialect speakers may contain relevant data for historical linguists, but this type of material is very time consuming to work with. However, there is a third type of archival language data, which constitutes an intriguing source of linguistic structure of old: dialect texts, handwritten in the 19th century using a traditional phonetic alphabet. Dialect texts of this sort exist in archives all over Scandinavia, and through them, we are granted access to the phonetic subtleties of an era that is too distant to have been caught on audio tape.

In my talk, I will address two such texts (both written in the 1890s) from the south-west of Sweden: the first one is a

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

compilation of dialectal expressions, collected in the parish of Fagered (in the province of Halland); the second one is a collection of narratives from of the island of Orust (in the province of Bohuslän). I refer to the alphabet used in these texts as LMA, a label based on the name of the Swedish dialect alphabet (viz. LandsMålsAlfabetet, 'the alphabet for rural dialects'). Nowadays, the LMA is used very marginally (and almost never outside of traditional onomastics). As a rule, linguists of today instead use the International Phonetic Alphabet, IPA (https://www.internationalphoneticassociation.org), when there is need for phonetic detail in written form. However, as soon as corpus-based linguistic research targets non-phonological issues, the fine phonetic details are superfluous. In fact, such detail only makes word- and phrase-based searches more complicated. Consequently, in order to make the old dialect texts useful for different sorts of linguistic research, it does not suffice to simply transform the text of the images to a digital correlate. In addition, there is need for several conversions of the original text into different more or less simplified formats, which, in turn, can be useful also for non-linguists (both other researchers and members of the general public).

The tool I use to analyse and transcribe the dialect texts is Transkribus (https://transkribus.eu/Transkribus). The first step was to decide how to write the LMA with a standard keyboard. As mentioned, the LMA is hardly used anymore, and only very few of the LMA symbols have a Unicode status. Although all IPA symbols indeed do, they are difficult to produce with a standard keyboard. In order to reach an acceptable transcription speed, I have instead created a SAMPA based transcription key. SAMPA stands for Speech Assessment Methods Phonetic Alphabet (https://www.phon.ucl.ac.uk/home/sampa) and it resorts only to the 128 characters that a standard (i.e. English) keyboard can produce. These characters, either in isolation or combined with others, are then given a specific phonetic value. Although the underlying principles for creating phonetic symbols are the

same, my dialect SAMPA is a digital version of the LMA and is therefore quite different from standard Swedish SAMPA, which is IPA-based.

To begin with, I made a SAMPA transcript of roughly 100 pages of the Fagered collection. This amount of manual transcription is what is needed to train a so called HTR engine (where HTR stands for handwritten text recognition). Once the HTR engine is integrated in the Transkribus platform, it is capable of automatically generate transcriptions of more text of the same hand. How well the engine works of course depends on an array of factors. One factor that often (according to the Transkribus crew) turns out to be complicating is super- and subscripted diacritics of the sort that occur abundantly in the dialect texts. Still, the HTR engine managed to handle the rest of the Fagered collection almost flawlessly; only a handful of minor manual corrections (concerning individual segments or diacritics) per page (16 lines) was required to perfect the transcription.

Transcription accuracy naturally decreases dramatically when the HTR engine is run on other LMA texts, written by other field linguists. When the Fagered engine handles text from Orust, only about a third of the LMA words are represented correctly in the SAMPA format. However, by adding some 50 pages of manual transcription of Orust text to the training sample of the existing HTR engine, the resulting SAMPA output becomes as satisfactory as with the Fagered collection.

Apart from dealing with the actual transference process (i.e. LMA image □ SAMPA transcript), I have also experimented with conversions from SAMPA to other more or less simplified formats, in order to make the texts accessible for a wider circle of users. Only quite recently have I become aware of the models for dialect transliteration developed by the Text Laboratory in Oslo. These models transform dialectal forms to standard language, which opens up for automatic lemmatization and annotation, in turn enhancing searchability radically. My ambition is to learn from the Norwegian project and to add transliteration to standard Swedish to the list of formats that the SAMPA transcripts can be converted to.

Finally, I will show how my linguistic research into clitics has been facilitated by the digitization of dialect texts. Since the SAMPA output contains both phonetic and prosodic details, it is fairly easy to extract those instances of prosodic dependencies (marked _ in the SAMPA format) in a text that represent potential enclitic pronouns. A somewhat prosaic effect hereof is simply that I am now able to sort and quantify relevant data in a way that I could not do before. A more intriguing consequence is that I have actually discovered linguistic variation that has previously fallen under the radar. For instance, descriptions of the traditional Bohuslän dialect mention only one masculine and one feminine object clitic: (e)n and (n)a respectively. However, my Orust text reveals a hitherto unnoticed gender asymmetry: the feminine form (n)a in fact competes with a reduced form of the full pronoun hener (viz. ner), whereas (e)n remains the only masculine option, reduced forms of the full pronoun ham being unattested.

ID: 150
Long paper presentations
Topics: cultural studies, historical studies, digital resources – publication and discovery, GLAM: galleries / libraries / archives / museums, museology, web research, archiving
Keywords: virtual museum, multimedia, web archive, museum studies

Digital History of Virtual Museums: The Transition from Analog to Internet Environment*

Nadezhda Povroznik

Perm State University, Russia; University of Luxembourg, Luxembourg

Many thousands of virtual museums exist on the Internet, demonstrating very diverse and significant museum resources, showcasing the treasury of humankind. These resources have come a long way in their evolution over past decades. The history of virtual museums began long before they appeared on the Internet, and the concept of virtual museums needed to be established in order to become an essential and effective means of accomplishing new museum functions in the digital age. Through the designing of such a concept, the creation and development of museums' information resources, websites and various digital initiatives have become the keys to the success of museums in the digital environment today. This article considers the concept of a virtual museum, traces the transition of virtual museums from analog and interim multimedia formats to the online environment. The author surveys the crucial moments in the history of virtual museums and the stages of their development from the digital turn to their appearance on the Internet and subsequent transformation after this transition. In this article examples of museum information resources from North America and Europe, Japan and Australia are traced back to the first virtual museums online in the 1990s. Based on the analysis of materials from web archives, strategies for creating the first virtual museum resources on the WWW are identified.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 182
Poster
Topics: literary studies, data mining / text mining
Keywords: verse structures, verse mining, literary prose, brazilian literature

Computer-Based Identification of Metric Verse Structures in Literary Prose of Portuguese Language

Joao Queiroz[1], Ricardo Carvalho[2], Angelo Loula[3]

[1]Federal University of Juiz de Fora, Minas Gerais, Brazil; [2]State University of Feira de Santana, Bahia, Brazil; [3]State University of Feira de Santana, Bahia, Brazil

Metric verse structures in Portuguese prose are still a phenomenon unexplored by philosophy, theory, and history of literature, and the automatic mining of such structures has not been tried in Computational Linguistics and Digital Humanities. The MIVES (Mining Verse Structure) system was developed for computational scansion of metric verse structures in Portuguese language prose (Carvalho, Loula and Queiroz, 2019). Unlike many computational systems already developed for scansion of metric poems, MIVES was designed to scansion metrical structures in prose, an operation that Augusto de Campos (2010, p.14) called “verse-spectral reading.”

MIVES extracts and processes sentences from the text, identifies and classifies metric structures that are searched by the user, and provides a view of the results obtained. The greatest challenge lies in the process of identifying metric structures, since a single, unambiguous, context-independent result does not arise from scansion. And considering that we are mining prose, there is no clear demarcation of the beginning and the end of structures, such as those easily found in the verses of a metric poem. In prose, the metric structure can be formed by a complete sentence, or by a sentence segment.

The processing begins with the extraction of sentences from a text file. Each sentence is then segmented into words for syllabic separation and identification of tonic syllables.

Although protocols for separating poetic and grammatical syllables do not necessarily produce coincident results, grammatical separation of syllables is an initial step towards poetic separation. At this stage, a sentence such as “Hipóteses sobre a sua gênese.” Can be scanned as “Hi/p#ó/te/ses/ s#o/bre a s#u/a g#ê/ne/se”, where / indicates a syllable separator and # a tonic syllable marker. But the scansion does not end in the phase of syllabic separation of the words. The sentence is subjected to scansion, which considers normatively accepted variations of syllabic separation, considering intervocabular phenomena and intravocabular phenomena, which can fuse or intervocabular and separate syllables. The scansion process does not produce unambiguous results and the intervocabular and intravocabular phenomena can be considered or not, thus multiple scansion possibilities are performed. It is then determined whether the sentence, or an excerpt from it, has a metric pattern and possible alternatives with different syllable counts are indicated. The text, whose metric sentences and variations have been identified, is sent to an interface for visualization, navigation and analysis of results.

The search for metric structures is not restricted to complete sentences. Initial or final sentences segments can be evaluated according to user decision. For a beginning or end of sentence, punctuation marks such as semicolon, colon, ellipsis, and exclamation are used as delimiters. For the identification of excerpts, the sentences are scanned until a metric structure is found that is adequate to the standards designated by the user.

The system is able to identify, classify and compare, frequency, density, and dispersion of heterometric verse structures, distributed at different scales of observation, from one work or author to aesthetics movements. Here we present preliminary results analyzing three works by Euclides da Cunha (Os Sertões, À Margem da História, Contrastes e

Confrontos). They were selected because constitute the main corpus of one of the most important Latin-American writers from XX century, and because Os Sertões was the object of a manual “verse-spectral reading” by the Brazilian poet and translator Augusto de Campos (2010, p.14). Such operation by Augusto de Campos revealed “more than 500 decasyllables in the book”, among sapphic and heroic structures, and more than two hundred dodecasyllables. On the other hand, MIVES processed 8564 sentences and found 652 (7,6%) full sentences, 1746 (20,4%) initial segments of sentences and 1728 (20,2%) final segments of sentences with structure between 10 and 12 metric syllables, a surprising rate, with much higher density when compared to results exhibited by Augusto de Campos.

Similar high density is also present in the other works by Euclides da Cunha, as reported by MIVES. A total of 1066 sentences were processed from À Margem da História, and the system found 76 (7,1%) full sentences, 227 (21,3%) initial segments of sentences and 219 (20,3%) final segments of sentences with metric structure. From 1598 sentences processed from Contrastes e Confrontos, MIVES found 82 (5,1%) full sentences, 282(17,6%) initial segments of sentences and 25 (14,7%) final segments of sentences with metric structure.

To evaluate the distribution of metric structures along the books, the distance between sentences with metric structures was measured from results obtained by MIVES. The average distance between such full sentences with metric structures in Os Sertões was 12,63±12,88 sentences, i.e. there were on average 12,63 (with standard deviation 12,88) non-metric sentences between each occurrence of metric structures of full sentences. In the book À Margem da História, the average distance between such full sentences with metric structures was 14,62±16,89, and in the book Contrastes e Confrontos, such distance was 18,38±19,61. Overall, these distances along with graphs of local density values along the books reveal that the distribution of metric structures throughout the books has a great variance, with regions of greater

concentration and regions of scarce presence of such structures.

Obviously, MIVES can perform scansions in much larger quantities than any human agent. But even more interesting, as a tool, is the ability that MIVES inaugurates to identify, quantify, and display patterns of distribution of versification structures throughout the text, numerically, with descriptive statistics and distance attributes, and visually, through dispersion and frequency throughout the works. It is not an exaggeration to say that the system is capable of opening a new direction in the investigations on literary prose, in Portuguese language.

References

Augusto de Campos (2010) Transertões. In: CAMPOS, A.; ALMEIDA, G. Poética de Os Sertões. São Paulo: AnnaBlume.

R. S. Carvalho, A. Loula, J. Queiroz (2019) . Identificação computacional de estruturas métricas de versificação na prosa literária de Euclides da Cunha. Revista de Estudos da Linguagem, aop14918.2019.

> MIVES is available at http://sites.ecomp.uefs.br/lasic/projetos/mives (project) and https://github.com/lasicuefs/mives (source code)

ID: 152
Short paper presentations
Topics: historical studies, linked data / semantic web / ontologies, visualisation
Keywords: linked data, Finnish Civil War, war history, semantic web

Building a Linked Open Data Portal of War Victims in Finland 1914ï 1922*

Heikki Rantala[1], Ilkka Jokipii[2,3], Mikko Koho[1], Esko Ikkala[1], Jouni Tuominen[1,2], Eero Hyvönen[1,2]

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

This paper presents first results from a project that aims to publish data about the war victims in Finland in 1914–22 as a Linked Open Data service and to create a portal of tools called WarVictim-Sampo 1914–22 to explore and analyze the data. At the same, time the data is extended with new information and cleaned from mistakes when found. The project is based on the database War Victims of Finland 1914–22 (“Sotasurmat 1914–22”) of the National Archives of Finland and related data compiled during the project. The core of the data includes information about roughly 40000 war victims. Most of these deaths are due to the Finnish Civil War but some are related to the First World War and the Kindred Nations Wars.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 114
Short paper presentations
Topics: archaeology, historical studies, linguistics, medieval studies, data mining / text mining, diversity and multilingual / multicultural approaches, geospatial analysis – interfaces & technology, interdisciplinary collaboration, visualisation
Keywords: personal names, Russia, Finnic, history

Personal Names as Mirrors of the Past in Medieval Northwestern Russia

Jaakko Raunamaa, Antti Kanner

University of Helsinki, Finalnd

Name is a linguistic universal that occurs in all known languages of the world. Names are used to identify individual people, places, and other referents. Furthermore, names are connected to the culture surrounding them. For example, Bedouins living in North Africa have different ways of naming places and people than the Finns living in Northern Europe. Similarly, many Finnic and Sami (Finno-Ugric language groups) place names occurring in Northern Russia prove that Finno- Ugric tribes inhabited these areas earlier. In other words, names preserve information about their users and can give researchers clues on what has happened in the past (Ainiala et al. 2012: 13– 29.)

This paper introduces the personal name system used at the end of the 15th century in Northwestern Russia. More precisely, the study focuses on the personal names attested in the census books of Novgorod (AD 1499–1563). These contain over 10 000 personal names and cover large areas in Northwestern Russia. The aim is to examine what kind of personal names were used in the area and what kind of regional differences can be found in the name usage. The study concentrates in particular on the northern areas of Novgorod Republic that supposedly had Finnic population. The goal is to learn if personal names used in Finnic areas differ from other ones. Last, the results are compared to archaeological, genetic and linguistics researches and a broader overview of the settlement history in medieval Northwestern Russia is presented.

Since Northwestern Russia, and especially its northern part, has been remote and loosely populated before the modern era, there are only limited amount of historically important sources, such as archeological finds or written documents. Thus, the history of Northwestern Russia is full of questions and uncertainties. For a long time already, researchers interested in history have used linguistics and onomastics in order to create a more comprehensive picture of the past (e.g. Rjabinin 1997 and Sedov 1982). However, the usage of names as a source material is, in many cases, small scale and limited. Either the studies are often regionally restricted or they have only limited amount of analyzed names. In addition, many history-oriented studies rely only on contemporary name data.

To some extent, the above mentioned problems can be explained by the methods and materials that have been used in the past. More precisely, the analogical materials, such as written documents or hand-drawn maps, have not allowed researchers to create a compressive studies based on names. The situation is now different since digital methods can be used to overcome the problems that earlier studies had. Many tasks that were previously considered as too time-consuming, like collecting thousands of names from documents, can now be done on a computer.

This study relies on methods that development of digital humanities have made possible. First, the research material is compiled from the editions of Novgorod census books by scanning the pages and using OCR-reading to create editable copies of texts. The census books from the area of the Novgorod Republic were a product of a certain order coming from the Grand Prince of Moscow. The Grand Duchy of Moscow had subjugated the city-state and its belongings before the end of the 15th century. The ruler wanted to know how much income the Grand Duchy of Moscow should acquire from the newly conquered area, and thus the Moscovites ordered the tax documentation after the conquest had been finalized (Nevolin 1851: i–xii). The documents are written in (old) Russian. Sources chosen for this study are edited versions of 15th and 16th century census books (NPK III, IV;

POKV; PKOP). These transcriptions were mainly done at the turn of the 20th century. The study area is presented on a map below. Material contains approximately three thousands pages, in which there are around 10 000 villages and over 20 000 homesteads. Tax payers are grouped into homesteads (in Russian dvor). One homestead usually contains one owner but sometimes there are other people named as well, such as the brother(s), adult son(s), nephews and other relatives of the owner. All the census books are divided into parishes (in Russian pogost), which are typically named after the location of the main church or after the monasteries or local nobles, who had the rights to collect the taxes.

The structured pattern of census books simplifies the process of collecting taxpayers' personal names. For example, in census book POKV, which covers the areas of Karelian Isthmus and western shores of Lake Ladoga, the pattern is almost always following: "Деревня Дуброва, (д) Фомка Ивашковъ, (д) Онтушко Ивашковъ;" ('Village Dubrova, (d)vor Fomka Ivaškovŭ, (d) Ontuško Ivaškovŭ;'). A Python script was written to exploit the systematic formalities of this record to harvest the personal names mentioned. The output is a data matrix that contains frequencies of person names for each parish, including main names (e.g. Ontuško) and bynames, such as patronyms (Ivaškovŭ) or descriptive ones (Volkŭ 'wolf').

This allows for a systematic statistical measurements of similarity across the parishes. Classification of names makes it possible to evaluate how the measured similarities are caused by names belonging to, for example, different parishes. Comparing for similarities of naming practices is not a straightforward task, since there is no straightforward definition for naming practice. However, simply applying different distance measurements highlights different aspects of the use of person names. Cosine similarity for highlights of the widest overall trends, Jaccard index for selection of names. Hierarchical clustering algorithm enables to cross-reference the similarity of naming practices with geographical data to see whether area based clusters emerge. Together these

approaches contribute in forming a holistic interpretation of how names expressed linguistic and ethnic identities in northern areas of Novgorod Republic.

One of the main aims of this study is to focus on the northern areas of Novgorod Republic that supposedly had Finnic population. This area was bordered in the northwest by the Diocese of Åbo that was eastern part of the Realm of Sweden. Mostly Finnic speaking tribes, such as Ingrians, Karelians and Savonians, inhabited the border area. The emergence of these groups is a continuously discussed question among scholars but it is known that they share many similarities in archaeological finds dated into Late Iron Age (AD 1000–1200) (Uino 2003: 300–400) and in linguistics as well (Frog & Saarikivi 2012). Thus, it is worthwhile to compare the personal names attested in the Novgorod census books to those that are attested in the Swedish taxation documents concerning the border area. The reference material, altogether approximately 2000 names, consists of personal names used in 1561 in parish Juva from Savo region and of names used in 1545 in parish Kivennapa located in Karelian Isthmus. Finnic names, such as main names or clan names, are particularly interesting because they have been used on both sides of the border: e.g. in Kivennapa Kaupi Nousia and in Kir'jažskij pogost (in Finnish Kurkijoki parish) Kiridko Novzejevъ.

Measuring and evaluating the census book data and comparing it to material collected from Swedish documents creates many new valuable perspectives into the history of Northwestern Russia. The results demonstrate how different personal names were distributed and used in the study area. This outcome is compared to the latest archeological, linguistic and genetic research, which allows us to create a comprehensive picture of the directions of cultural impacts and settlement movements in medieval Northwestern Russia. In addition, the results reveal those areas that were inhabited by people using Finnic names or Finnic forms of the Christian names in the end of the 15th century.

Fig. 1. The thin dotted line depicts the borders of the study area.

References

Ainiala, Terhi, Minna Saarelma & Paula Sjöblom 2012: Names in Focus. An Introduction to Finnish Onomastics. Studia Fennica. Linguistica 17. Helsinki: Finnish Literature Society.

CHR = Maureen Perrie (ed.) 2006: The Cambridge History of Russia: Volume 1, From Early Rus' to 1689. Cambridge: Cambridge University Press.

Nevolin 1853 = Неволин, К. А.: О пятинах и погостах новгородских в XVI веке, с приложением карты. Из Записок Императорского русского географического общества, Кн.

VIII. Санкт-Петербург: Тип. Имп. Акад. наук.

NPK III = Новгородские писцовые книги. Переписная окладная книга Водской пятины 1500(7008) года. Часть 1. Санкт-Петербург: Археографицеская Коммиссия. 1868.

NPK IV = Новгородские писцовые книги. Переписная оброчная книга Шелонской пятины. 1498, 1539, 1552–1553. Санкт-Петербург: Археографицеская Коммиссия. 1886.

PKOP = Писцовые книги Обонежской пятины : 1496 и 1563 гг. Ленинград: Академия наук Союза Советских Социалистических Республик. Археографическая комиссия. 1930.

POKV = Переписная окладная книга по Новугороду Вотьской пятины : 7008 года. Москва: Временник Московского общества истории и древностей. 1851.

Rjabinin 1997 = Рябинин, Е. А.: Финно-угорские племена в составе Древней Руси : к истории славяно-финских этнокультурных связей. Историко-археологические очерки. Санкт-Петербург: Изд-во С.-Петербург. унта.

Ronimus, J.V. 1906: Novgorodin vatjalaisen viidenneksen verokirja v. 1500 ja Karjalan silloinen asutus. Helsinki: Suomen historiallinen seura.

Sedov 1982: Седов, В.В.: Восточные славяне в VI–XIII вв. Москва: Российская академия наука.

ID: 251
Keynote speaker

A Vaccine Against Fake News

Jon Roozenbeek

Cambridge Social Decision-Making Lab, University of Cambridge

Jon Roozenbeek will be talking about online misinformation and what to do against it. The problem has been highly pervasive, and governments, social media companies, think tanks and civil society have found it difficult to find sustainable, scalable solutions. Jon will discuss what he and his colleagues have been doing to combat online misinformation, by combining insights from social psychology with gamification. He will present the game Bad News, an online browser game in which players take on the role of a fake news creator and must spread as much fake news as they can in order to win. We will then discuss the research that has been conducted on the effects of the game: does it actually make people better at spotting misinformation? And if so, how can this solution be used at scale?

ID: 169
Poster
Topics: historical studies, corpus linguistics, data mining / text mining, information retrieval, natural language processing, discourse analysis, political science
Keywords: political grammar, conceptual history, parliamentary proceedings, text mining, argument mining

The Grammar of Politics. Modelling Technocratic Speech and Argumentation in Parliamentary Debate 1918ï 2017

Ruben Ros

Utrecht University, The Netherlands

The rise of populism, the decline of ideology and the challenges of transnational governance have brought the concept of technocracy back to the centre of academic attention (Esmark, 2016). The concomitant narrative of “technocratization” involves the erosion of democratic debate, following from either the (institutional) displacement of decision-making powers or the diffusion of technocratic ideas. Technocracy, however, is hard to trace. It operates not so much on the level of what is said, but how it is said. This project therefore studies technocracy as a specific mode of thinking by computationally modelling argumentation in Dutch and British parliamentary proceedings between 1918 and 2017. The computational identification of different types of argumentation and the examination of connections between these different types, political actors and specific concepts allows a systematic analysis of the rise of technocratic modes of thinking in this period.

Data

This project employs the digitized parliamentary proceedings of the Dutch First and Second Chambers (Handelingen der Staten-Generaal, 1814–2017) and the transcripts of the debates held in the House of Commons and the House of Lords (Hansard Corpus, 1803–2005 & HanDeSet, 1997–2017). Both datasets consist of full-text proceedings accompanied by metadata on speakers, topics and the structure of debates (Alexander and Davies; 2015, Marx et al., 2012). The data has

been semantically and syntactically enriched, which allows the division of the data based on topics and linguistic features (Nanni et al., 2019). All datasets are fully accessible through APIs.

Parliamentary debate has long been regarded as a mere veil obscuring power interests. In the wake of the linguistic and cultural turns of the 1990s, parliamentary sources have been rediscovered as objects in the study of political culture (Mergel, 2002). The 2000s saw a greater focus on parliamentary language, involving such approaches as discourse analysis and cognitive linguistics (Ihalainen et al., 2018; Bayley, 2004).

This project considers Dutch and British parliamentary proceedings from the period between 1918 and 2017 because in this era, modern democracies formed. Universal suffrage was established in the late 1910s in both nations. During the twentieth century, parliamentary culture developed in different ways in the Netherlands and the United Kingdom. Nevertheless, they shared important political events and developments: from the postwar period of reconstruction and the consensual nature of parliamentary debate to the conservative turn in the 1980s and the resurgence of populism in the 2000s.

Approach

To connect arguments and argumentative types to the topic of this research, I identify three elements of technocracy/technocratization that relate to argumentation. First, technocratic thinking can be understood as a way of managing contingency and the denying the existence of political alternatives (Sánchez-Cuena, 2017). Second, technocratic thought is marked by frequent references to external actors and factors (McKenna & Graham, 2000). Third, this generalizing dynamic leaves the politician with a smaller “space” in which to operate (Palonen, 2003). This tripartite definition of technocracy, combined with the insights derived from the literature leads to a hypothesis with the following components:

Parliamentary debate slows a gradual increase in technocratic argumentation starting in the 1930s and accelerating in the 1950s.Technocratic argumentation is originally found in debates on macro-economic policies and welfare state provisions. Technocratic argumentation is not restricted to specific political parties. In the 1970s and 1980s, technocratic argumentation “spills over” into other policy areas, leading to convergence in the form of technocratic argumentation and thus a technocratization of political debate.

Methods

The project studies the topic of technocracy and technocratization from the perspective of political grammar. It does so by focusing on a specific unit of analysis: the argument. Arguments can be defined as linguistic expressions “aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or refuting the proposition expressed in the standpoint” (Van Eemeren and Grootenhorst, 2004). Philosophers distinguish between various argumentative forms and composed many so-called “argument schemes” (Lumber, 2016). The proposed research uses a recently proposed scheme: The Periodic Table of Arguments (PTA). This classification has been formulated as a synthesis between existing approaches and is particularly suitable for computational analysis (Wagemans, 2019). Compared to other frequently-used schemes, such as the Toulmin Scheme or Waltons scheme, the PTA focuses on the linguistic aspects of arguments, which results in a scheme that is more easily translated to models and text classification algorithms.

Recently, argument schemes have been fruitfully applied to “argument mining”: the computational extraction and classification of arguments from textual data. Argument mining has been highly successful in detecting and classifying arguments in for example online discussions and scientific papers (Janier and Saint-Dizier, 2019). The present project breaks down the process of argument mining into a specific series of “subtasks” (Stede & Schneider, 2018). First, argumentative sentences have to be identified and segmented

into so-called “Argumentative Discourse Units” (ADUs). This entails the identification of a conclusion (“the statement that is doubted”) and its premise(s) (“the statement that is supposed to take away that doubt”) (Wagemans, 2019). Subsequently, the nature of the relationship between these ADUs needs to be determined. Premises must be tied to conclusions and different arguments must be separated. Based on the classification of ADUs, the relations between them and the specific configuration of linguistic entities such as subjects and predicates, an argument type will be assigned (Figure 1).

Performing the subtasks computationally requires the generation of a data-driven model for argument classification, which is dependent on manually annotated training data. For this project, manual annotation is considered an added value: it facilitates a first iteration of close reading and hypothesis testing. Annotation guidelines will be developed based on the existing documentation on the PTA and other argument mining projects (Stab and Gurevych, 2014; Visser et al., 2018). After establishing a sufficient level of Inter-Annotator Agreement, the annotations will be exported in the standard Argument Interchange Format (AIF) (Lawrence et al., 2016).

The data structure that consists of argumentative texts, conclusions, premises and metadata will subsequently be used to investigate three components of technocratization. First, I investigate the general change in parliamentary argumentation. This entails mapping the changing prominence of argument types, looking at concepts in specific argument types and investigating the order of arguments in debates. Second, I examine whether arguments and types related to technocratic reasoning originate in specific policy areas. Lastly, I relate my findings to parties and politicians and ask to what extent technocratic argumentation runs through the traditional ideological divides or correlates with specific political actors.

This project looks at political language from a radically new perspective. By applying argument mining to historical data, new light is shed on political language and the changing ways in which politicians have argued. As such, this project

contributes to the integration of the rapidly innovating field of argument mining and political history.

References

Alexander, W. and Davies, M. (2015). The Hansard Corpus 1803–2005 <http://www.hansard-corpus.org/>

Budzynska, K., & Reed, C. (2019). Advances in Argument Mining. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 39–42.

Esmark, A. (2016). Maybe it is time to rediscover technocracy? An old framework for a new analysis of administrative reforms in the governance era. Journal of Public Administration Research and Theory, 27(3), 501–516.

Ihalainen, P., Ilie, C., & Palonen, K. (eds.). (2016). Parliament and Parliamentarism: a comparative history of a European concept. London/Frankfurt: Berghahn Books.

Janier, M., & Saint-Dizier, P. (2019). Argument Mining: Linguistic Foundations. London: John Wiley & Sons.

Lange, M. van & Futselaar, R. (2019). Debating Evil. Using Word Embeddings to Analyse Parliamentary Debates on War Criminals in the Netherlands. Contributions to Contemporary History, 59(1).

Lawrence, J., Duthie, R., Budzynska, K., & Reed, C. (2016). Argument Analytics. Proceedings of the Sixth International Conference on Computational Models of Argument 2016. Amsterdam: IOS Press.

Victor, J. N., Montgomery, A. H., & Lubell, M. (Eds.). (2017). The Oxford Handbook of Political Networks. Oxford: Oxford University Press.

Marx, M. Doornik, J. van, Nusselder, A. and Buitinck, L. (2012). Dutch Parliamentary Proceedings 1814–2012, nonsemanticized, Distributed by DANS EASY.

McKenna, B. J., & Graham, P. (2000). Technocratic discourse: A primer. Journal of Technical Writing and Communication, 30(3), 223–251.

Mergel, T. (2002). Überlegungen zu einer Kulturgeschichte der Politik. Geschichte und Gesellschaft, 28(4), 574–606.

Nanni, F., Menini, S., Tonelli, S., & Ponzetto, S. P. (2019). Semantifying the UK Hansard (1918–2018). Proceedings of the ACM/IEEECS Joint Conference on Digital Libraries (JCDL'19). New York.

Stede, M., & Schneider, J. (2018). Argumentation Mining. Synthesis Lectures on Human Language Technologies, 11(2), 1–191.

Visser, J., Lawrence, J., Wagemans, J. H., & Reed, C. (2018). Revisiting Computational Models of Argument Schemes: Classification, Annotation, Comparison. Proceedings of the Sixth International Conference on Computational Models of Argument 2018. Amsterdam: IOS Press.

Wagemans, J. H. (2019). Four basic argument forms. Research in Language, 17(1), 57–69.

ID: 215
Short paper presentations
Topics: linguistics, corpus linguistics, data mining / text mining, linking and annotation
Keywords: sentiment analysis, opinion mining, aspect-based sentiment analysis, annotated corpus

Creating an Annotated Corpus for Aspect-Based Sentiment Analysis in Swedish*

Jacobo Rouces, Lars Borin, Nina Tahmasebi

University of Gothenburg, Sweden

Aspect-Based Sentiment Analysis constitutes a more fine-grained alternative to traditional sentiment analysis at sentence level. In addition to a sentiment value denoting how positive or negative a particular opinion or sentiment expression is, it identifies additional aspects or `slots' that characterize the opinion. Some typical aspects are target and source, i.e. who holds the opinion and about which entity or aspect is the opinion. We present a large Swedish corpus annotated for Aspect-Based Sentiment Analysis. Each sentiment expression is annotated as a tuple that contains a one among 5 possible sentiment values, the target, the source, and the existence of irony. In addition, the linguistic element that conveys the sentiment is identified too. Sentiment for a particular topic is also annotated at title, paragraph and document level. The documents are articles obtained from two Swedish media (Svenska Dagbladet and Aftonbladet) and one online forum (Flashback), totalling around 4000 documents. The corpus is freely available and we plan to use it for training and testing an Aspect-Based Sentiment Analysis system.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 101
Long paper presentations
Topics: linguistics, cultural heritage collections, data mining / text mining, natural language processing, computational science
Keywords: named entity recognition; evaluation; historical newspapers; Finnish; OCR data

Name the Name ï Named Entity Recognition in OCRed 19th and Early 20th Century Finnish Newspaper and Journal Collection Data*

Teemu Ruokolainen, Kimmo Kettunen

University of Helsinki, National Library of Finland, Finalnd

Named Entity Recognition (NER), search, classification, and tagging of names and name like frequent informational elements in texts, has become a standard information extraction procedure for textual data. NER has been applied to many types of texts and different types of entities: newspapers, fiction, historical records, persons, locations, chemical compounds, protein families, animals etc. Performance of a NER system is usually quite heavily genre and domain dependent. Entity categories used in NER may also vary. The most used set of named entity categories is usually some version of three partite categorization of locations, persons, and organizations.

In this paper we report evaluation results with data extracted from a digitized Finnish historical newspaper collection Digi using two statistical NER systems, namely, Stanford Named Entity Recognizer and LSTM-CRF NER model. The OCRed newspaper collection has lots of OCR errors; its estimated word level correctness is about 70–75%. Our NER evaluation collection and training data are based on ca. 500 000 words which have been manually corrected from OCR output of ABBYY FineReader 11. We have also available evaluation data of new uncorrected OCR output of Tesseract 3.04.01.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

Our Stanford NER results are mostly satisfactory. With our ground truth data we achieve F-score of 0.89 with locations and 0.84 with persons. With organizations the result is 0.60. With re-OCRed Tesseract output the results are 0.79, 0.72, and 0.42, respectively. Results of LSTM-CRF are similar.

ID: 181
Short paper presentations
Topics: film and media studies, crowdsourcing
Keywords: metadata, digital humanities, film history, audiovisual culture, audiovisual heritage

Crowdsourcing Metadata for Audiovisual Cultural Heritage: Finnish Full-Length Films, 1946ï 1985*

Hannu Salmi[1,4], Kimmo Laine[2,3], Tommi Römpötti[2], Noora Kallioniemi[1], Elina Karvo[1]

[1]University of Turku, Department of Cultural History, Finland; [2]University of Turku, Department of Media Studies, Finland; [3]University of Oulu, Department of Art Studies and Anthropology, Finlanf; [4]Turku Group for Digital History, Finland

This paper is based on a crowdsourcing project which was realised at the School of History, Culture and Arts Studies of the University of Turku between the years 2013–2018. The idea was to develop a format through which long-term crowdsourcing could be integrated into the humanities curriculum. The project was realised in close cooperation with the National Audiovisual Institute (KAVI) in Finland. The aim was to help KAVI in developing its open database for Finnish cinema, Elonet, by engaging both graduate and postgraduate students in producing keywords, genre characterisations, plot summaries and other relevant fields of information for Finnish cinema. In total, the project produced metadata for 572 full-length films, both fiction films and long documentaries that had their theatre release between the years 1946 and 1985. The amount is substantial considering that, to date, around 1,600 full-length films have been released in Finland. At the same time, it produced a successful model for drawing on crowdsourcing in the classroom.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 188
Long paper presentations
Topics: folklore and oral history, cultural heritage collections, diversity and multilingual / multicultural approaches, stylistics and stylometry / authorship attribution
Keywords: runosong

Using Word Statistics in Studying Variation of Folksongs

Mari Sarv

Estonian Literary Museum, Estonia

Within the field of digital humanities various methods and tools based on word statistics, like stylometry, topic modeling and sentiment analysis have become popular to answer different research questions on the basis of literary texts, written documents or everyday writings. With digital text corpora created on the basis of voluminous folklore collections methods based on word statistics have the potential to help us to gain better understanding on the essence of folkloric variation.

Variation is an inherent feature of folklore, emerging as a result of folkloric transmission process, where the transmitted knowledge is constantly re-created and adapted. At the same time the process is hard to be catched and surveyed in real-life situations. Statistical analysis of large text corpora enables us to get insight into the essence and details of variation, tradition flows and regional peculiarities.

My paper explores the possibilities of using the word statistics in studying the variation of Finnic runosongs on the basis of the material in Estonian and Finnish runosong databases. In studying the variation of textual folklore we always have to keep in mind that linguistic variation always underlies folkloric variation. In addition to underlying dialectal variation, runosongs use specific poetic register with archaisms and specific word forms instead of colloquial language. Due to the extreme linguistic variativity of runosongs the wordforms can not be automatically lemmatized nor grammatically analyzed. Nevertheless, use of computational methods based on word statistics like stylometry and topic modeling give us a valuable

overview on the regional and topical variation of runosongs. Thanks to existence of large corpora we can for the first time ever draw data-driven outlines on the content and regional division of the tradition, but without being able to distinguish linguistic and contentual layers in the analysis, the results also always include both aspects.

ID: 218
Short paper presentations
Topics: cultural studies, film and media studies, gender studies, corpus linguistics, data mining / text mining, games and meaningful play, natural language processing
Keywords: game studies, gender studies, LIWC, video games, video game magazines

Towards an Analysis of Gender in Video Game Culture: Exploring Gender-Specific Vocabulary in Video Game Magazines*

Thomas Schmidt, Isabella Engl, Juliane Herzog, Lisa Judisch

Media Informatics Group, University of Regensburg, Germany

We present preliminary results of a project examining the role and usage of gender specific vocabulary in a corpus of video game magazines. The corpus consists of three popular video game magazines with 634 issues from the 1980s until 2011 and was gathered via OCR-scans of the platform archive.org. We report on the distribution and progression of gender-specific words by using word lists of the LIWC for the categories “male” and “female”. We can indeed show that words of the type male are considerably more frequent than words of the type female, with a slight increase of female words during 2006–2010. This is in line with the overall development of gaming culture throughout these years and previous research in the humanities. Furthermore, we analyzed how the usage of negatively connoted words for female depictions (e.g. chick, slut) has evolved and identified a constant increase throughout the years reaching the climax around 2001–2005, a timespan that coincides with the release and popularity of games encompassing rather sexist concepts. We discuss the limitations of our explorations and report on plans to further investigate the role of gender in gaming culture.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 247
Poster
Topics: film and media studies, literary studies, audio / video / multimedia, linking and annotation
Keywords: sentiment annotation, sentiment analysis, movies, movie annotation, Arduino

Live Sentiment Annotation of Movies via Arduino and a Slider*

Thomas Schmidt, David Halbhuber

Media Informatics Group, University of Regensburg, Germany

In this late breaking poster, we present the first version of a novel approach and prototype to perform live sentiment annotation of movies while watching them. Our prototype consists of an Arduino microcontroller and a potentiometer, which is paired to a slider. We motivate the need for this approach by arguing that the presentation of the multimedia content of movies as well as performing the annotation live during the viewing of the movie is beneficial for the annotation process and more intuitive for the viewer/annotator. After outlining the motivation and the technical setup of our system, we will report upon studies we plan to validate the benefits of our system.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 167
Long paper presentations
Topics: sociology, data mining / text mining, natural language processing, social media, visualisation, religious studies
Keywords: religious studies, distant reading, Reddit, sentiment analysis, computational social science

Distant Reading of Religious Online Communities: A Case Study for Three Religious Forums on Reddit*

Thomas Schmidt, Florian Kaindl, Christian Wolff

Media Informatics Group, University of Regensburg, Germany

We present results of a project examining the application of computational text analysis and distant reading in the context of comparative religious studies, sociology, and online communication. As a source for our corpus, we use the popular platform Reddit and three of the largest religious subreddits: the subreddit Christianity, Islam and Occult. We have acquired all posts along with metadata for an entire year resulting in over 700,000 comments and around 50 million tokens. We explore the corpus and compare the different online communities via measures like word frequencies, bigrams, collocations and sentiment and emotion analysis to analyze if there are differences in the language used, the topics that are talked about and the sentiments and emotions expressed. Furthermore, we explore approaches to diachronic analysis and visualization. We conclude with a discussion about the limitations but also the benefits of distant reading methods in religious studies.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 213
Short paper presentations
Topics: corpus linguistics, digital resources – publication and discovery, infrastructure, interdisciplinary collaboration, linking and annotation, natural language processing, open data, sustainability and preservation, big data
Keywords: Latvian language, CLARIN, research infrastructure, language resources and tools, toolchain for language processing

CLARIN in Latvia: from the Preparatory Phase to the Construction Phase and Operation*

Inguna Skadiņa, Ilze Auziņa, Normunds Grūzītis, Arturs Znotiņš

The Institute of Mathematics and Computer Science, University of Latvia, Latvia

Qualitative language resources and tools for natural language processing are key elements for research in digital humanities (DH). Several research infrastructures, e.g., CLARIN, DARIAH, provide access to digital research objects for researchers around the Europe. Although these are pan-European research infrastructures, availability of content and the readiness of the particular node varies from country to country. This paper aims to present current status of the CLARIN research infrastructure in Latvia – key language resources and tools identified, development of the technical infrastructure, collaboration with DH researches and initiatives on user involvement and education. Being active participant of the CLARIN initiative during its preparation phase, Latvia joined CLARIN ERIC only four years after its establishment. This four-year gap puts Latvia's node in a construction phase, while in many countries CLARIN is already operational. Although many Latvian language resources and tools are currently not included in CLARIN repository, researchers of Latvia already now can benefit from the language resources and tools from different members of CLARIN ERIC through single sign-on.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 120
Poster
Topics: folklore and oral history, data mining / text mining
Keywords: folk legends, topic modelling, computer-assisted text analysis

Adapting a Topic Modelling Tool to the Task of Finding Recurring Themes in Folk Legends*

Maria Skeppstedt, Rickard Domeij, Fredrik Skott

The Institute for Language and Folklore, Sweden

A topic modelling tool, which was originally developed for performing text analysis on very short texts written in English, was adapted to the text genre of Swedish folk legends. The topic modelling tool was configured to use a word space model trained on a Swedish corpus, as well as a Swedish stop word list. The stop word list consisted of standard Swedish stop words, as well as 380 additional stop words that were tailored to the content of the corpus and therefore also included older spelling versions and grammatical forms of Swedish words. The adapted version of the tool was applied on a corpus consisting of around 10,000 Swedish folk legends, which resulted in the automatic extraction of 20 topics. Future versions of the tool will be extended with text summarisation functionality, in order to retain the text overview provided by the tool also when it is applied on longer folk legends.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 253
Keynote speaker

Digital Emotions: Hybrid Structure of Emotional Impacts

Jurģis Šķilters

Laboratory for Perceptual and Cognitive Systems, University of Latvia

After a brief overview of the current state of art in the field of research on emotions, I will explore the way emotions impact users in interface systems in virtue of affective reactions to colors and shapes. I will provide evidence from the work conducted in my lab. Further, I will discuss effects of emotional reactions (e.g., emotional contagion) in large-scale digital social networks and will argue for a more universal approach to emotional patterns in digital environments.

ID: 221
Short paper presentations
Topics: communication studies
Keywords: safety, security, critical discourse analysis, parliamentary corpus

Discourse on Safety / Security in the Parliamentary Corpus of Latvian Saeima

Ilva Skulte, Normunds Kozlovs

Riga Stradins University, Latvia

The discourse on (public) safety and (social) security in the political communication has an impact on constructing national identity and community feelings through the ideas of risk and emergency. Indeed, the many aspects of unsecurity / unsafety make this to be elaborated in speeches as a rather manifold and complex concept. How is this conceptual nexus used and perceived in the speeches of MPs of Latvian parliament, and what impact it may have had on (re)formed national identity? These are the main issues in the proposed paper. Methodologically here the analysis combines critical discourse analysis (CDA) and corpus analysis and is based on the Corpus of Debates in Latvian Saeima (1993–2017 (http://saeima.korpuss.lv/)). By means of corpus analysis tools the categories and frames of representation of safety and security in the speeches of Latvian MPs are selected and described, and the qualitative analysis is carried out to understand and interpret differences and similarities in understanding and treating different aspects of safety adn security by MPs in the parliamentary discourse in Latvia and the changes in it during the time period after regaining independence.

ID: 199
Poster
Topics: literary studies, philology, digital resources – publication and discovery, scholarly editing
Keywords: scholarly editing, versions, genetic reconstruction, digital archive, Maironis

The New Possibilities for Philological Research in the Digital Archive: The Case of ñThe Voices of Springò by Maironis

Magdalena Slavinska

Vilnius University, Lithuania

It is widely understood that digital media may allow users of a digital archive to thoroughly examine its digitized objects and their data. Moreover, connecting separate objects (witnesses) into a hyperlink network may provoke the reader to reflect upon these connections as well. The latter might be visualized to the reader in different ways: by hypertextual links or by side-by-side comparison on the screen. Connecting visible elements may be even more important in some type of archives, e. g. the genetic archive of a literary work shows different versions of the same text to be perceived by the user as a continuous process.

The aim of this paper is to present the first digital genetic archive in Lithuania – a digital scholarly edition of “The Voices of Spring” (http://www.pb.flf.vu.lt/) by Lithuanian poet Maironis (1862–1932). The 1st and the 5th authorial editions of “The Voices of Spring” are separated by thousands of textual variants, and the number of poems in the collection increased from 45 to 131. Development of the poems took four decades (since the first published verses in 1885 until 1927) and it demonstrates the effort to modernize the language of the verses accordingly to the processes of linguistic modernization, which coincided with the period of Maironis creative activity.

One may ask, how such numerous differences between the versions of the poems might be clarified in an archive. The striking amount of different changes between the 1st and the 5th edition calls for step-by-step demonstration of continuous writing and editing processes. There is also a need for a

conceptualized presentation that would allow to summarize the genetic process. This scholarly edition aims to carry out a genetic reconstruction of Maironis' poetry by presenting the facsimiles, the XML-encoded versions and commentaries.

The user of the genetic archive encounters with two types of visualizing connections between the objects: the spatial and the hypertextual. The two objects connected spatially might be compared and examined visually on the screen. Hypertextual links attached to an isolated object signalize the reader that there is another object which can be compared to, but the comparison between equally visible objects will not be possible until the user makes another move to examine the signalized object. The archive offers two types of comparisons to reflect upon. The user might examine the identities and significant distinctions between different digital representations of the same witness, e. g. between the facsimile (image) and the xml file (text), and deepen his understanding of one particular witness. The archive could also differently visualize the genetic connections between the versions (various witnesses).

By collecting all the witnesses of the text creation, a genetic archive seeks to visualize the process of writing for the reader, that is, offers him a convenient platform for viewing and analyzing the processes involved in the creation of the text. Therefore, it is equally important to allow the user, on the one hand, to deepen the knowledge about each version (separately taken from the linear or more complicated genetic sequence), and, on the other, to more thoroughly reflect on the transitions between the witnesses (authorial revisions of the poems).

In a genetic dossier one looks for the genetic relations between the variants. The identified genetic relations are further investigated and interpreted; there are attempts being made, to represent their structures in the archive. Finding the solution for the best visualization might be viewed as a two-way process. One attempts to comprehend the concept of genetic relation in the case of a particular archive. Furthermore, one must consider several possible graphical forms that would withstand the problem of conveying the process of changes in the text.

Highlighting the elements that differ in each version might serve as a point of departure for the genetic reconstruction. The sequence of those elements might form a certain logic, e. g. it might resemble the more general historical-linguistic process. However, the sequence of linguistic changes viewed separately from the texts may be treated as a process involved in the creation, but not a complete representation of the genesis. Linguistic changes might not cover all the elements in the text that provoked the revisions. Therefore, it is useful to literally keep in sight both the text and those elements involved in the genesis that have been stated so far. In the digital scholarly edition of “The Voices of Spring” these functionalities are achieved through slightly modification of EVT 2 (developed by Edition Visualization Technology Project team, the leader – Roberto Rosselli Del Turco). The user reads and interprets the genetic archive partly by means of computational methods. Digital representations of the textual witnesses and the results of computational processing are displayed to be grasped together. The computer-generated results of such functionalities as word concordances and visual comparison of different versions may augment the perception of the textual variation.

The archive aims to provide a platform for scholarly research, that could be developed in the future, when the new sets of data and commentaries could be added. Visualization of the linguistic tendencies involved in the genetic process of “The Voices of Spring” can serve as a material for further investigation. E. g. syntactical changes made by Maironis may not only indicate improving poetry, but also signify the transition from foreign syntactic structures – a process fully documented through his five authorial editions.

Finally, a genetic archive, of which concern is to collect all the indices of the author’s thought, may take flexible forms unfamiliar to the previous readers of „The Voices of Spring”. The archive also incorporates significant versions of the poems from periodicals and books which provide more details on the chronology of changes. Between two subsequent editions several poems were published together as a part of another

book, and some of them were further changed before the publication in the new edition of „The Voices of Spring". Therefore, a distinct genetic trajectory might be demonstrated through each of the poems in the collection, though they all have been published in the final authorial edition in 1927.

References

Bleier, Roman, et al., editors. Digital Scholarly Editions as Interfaces, Norderstedt: Books on Demand, 2018.

Drucker, Johanna. "Performative Materiality and Theoretical Approaches to Interface." Digital Humanities Quarterly, Vol. 7, Issue 1, 2013. 13 Jan. 2020 <http://www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html>

McGann, Jerome. A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction. Cambridge, MA: Harvard University Press, 2014.

Rosselli Del Turco, Roberto. "Designing an advanced software tool for Digital Scholarly Editions: The inception and development of EVT (Edition Visualization Technology)." Textual Cultures, Vol. 12, No. 2, 2019. 15 Jan. 2020
<https://scholarworks.iu.edu/journals/index.php/textual/article/view/27690>

ID: 140
Long paper presentations
Topics: sociology, cultural heritage collections, digitisation – theory and practice, artificial intelligence
Keywords: museum collections, digitization, machine learning, Python, OpenRefine

Digital Analysis and Machine View on Latvian National Catalogue of Museum Collections

Maija Spuriņa

Latvian Academy of Culture, Latvia

Latvian National Catalogue of Museum Collections (Latvijas Nacionālais Muzeju Krājuma kopkatalogs) is an on-line national digital data base of museum collection (www.nmkk.lv) that currently contains information on more than 1.2 million objects stored in 129 state certified museums. For the past fifteen years enormous efforts and financial means have been invested to create this one digital gateway to the cultural heritage stored in Latvian museums. The database was centrally designed and is administered top-down by a government agency – Culture Information System Center. The data has been input bottom-up by each museum. The resulting database can be seen as a digital representation of the national cultural heritage preserved in Latvian museums. In my presentation I will show preliminary results of my analysis of publicly available part of the database using OpenRefine and machine learning algorithm for object recognition (Python). I will discuss the challenges faced in the process of analysis due to the design of the database, as well as point out possibilities and limits of using machine learning algorithm for analysis of museum collections.

ID: 212
Short paper presentations
Topics: sociology
Keywords: digital citizens

Impact of Technologies on Political Behaviour: What Does It Mean to Be ñGood Digital Citizenò

Ieva Strode

University of Latvia, Latvia

Although the opinion that political activity of society has declined is relatively common (also confirmed by studies), there exist objections that activity may not have diminished, but its forms have changed, replacing traditional forms of participation to others. Technological development, which offers a new environment and new forms for participation in the political community, also plays a particular role in these changes. Both “digital citizens” that have emerged as a result of the technological transformation and political institutions now need to redefine what does it mean to be a “good citizen” in terms of rights, duties and participation in this digital world.

Some changes in political behaviour means just the “movement” of existing norms and traditional behaviour patterns to the digital environment: interest in politics, expression of political views, participation in some political activities may essentially preserve traditional content, while activities take place using new tools, platforms (e.g. through social networks, new political communities etc.). However, the new digital environment also requires and creates new norms related to “good digital citizens’ “ duties (including behaviour (e.g. following etiquette, obeying laws and rules that are specific for digital world), obtaining specific education and knowledge (e.g. digital literacy etc.)) and rights (e.g. access to digital resources, equality within society regarding access to these resources etc.). It is also important to assess activity of digital citizens in the political events of non-digital world (e.g. elections if e-voting is not available).

In my proposed presentation, I will discuss how traditional rights and duties of “good citizens” are transferred to digital world, what kind of new rights and duties have emerged due to digital environment, and analyse Latvian population survey data regarding citizens’ opinions about norms of good digital citizens and their actual behaviour.

ID: 143
Short paper presentations
Topics: data mining / text mining, speech processing, political science, artificial intelligence
Keywords: automatic speaker identification, elections, deep neural networks, weakly supervised training

Analyzing Candidate Speaking Time in Estonian Parliament Election Debates*

Siim Talts, Tanel Alumäe

Tallinn University of Technology, Institute of Software Science, Estonia

In this paper, we analyze the amount of speaking time by each candidate and political party during the election debates that aired in broadcast media during the Estonian 2019 parliament election campaign, using automatic speaker identification. We use automated methods for retrieving speech recordings from publicly available sources that are likely to contain speech by the target speakers, and apply a weakly supervised method for training speaker recognition models from such data. The experiments show that the resulting models have high precision but they lack in recall. While most candidates from large and established parties could be identified automatically, candidates from smaller and newer parties could not be recognized, due to their little prior media presence. Our system allows to identify candidates who spoke for the longest time in the election debates. The analysis shows that the election debates were not biased from the speaking time point of view: all major political parties received relatively similar speaking time across the debates.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

ID: 132
Short paper presentations
Topics: literary studies, open science, visualisation
Keywords: Gulag literature, Estonian literature, camp prose

Gulag Literature: Looking Through the Glass of Digital Humanities

Kseniia Alexandrovna Tereshchenko

ITMO University, Russia

There are several Digital Humanities projects that are dedicated to Stalin's terror such as «Это прямо здесь» ("This is right here", URL: https://topos.memo.ru), «Открытый список» ("The open list" URL: https://ru.openlist.wiki), «Бессмертный барак» ("Immortal Gulag" URL: https://bessmertnybarak.ru), Gulag online (URL: http://www.gulag.online) etc. These projects mostly focus on the Russian history, providing information about Russians, Russian cities and so on. Also, they bring to light the phenomenon of Stalin's terror by means of history and historical facts. In my research I suggested taking a look at the same topic from a different angle.

First of all, in my project I emphasized that not only Russians suffered from Stalin's terror, but also citizens of other Union republics. Secondly, I researched this topic by applying mainly cultural rather than historical approach, as there is not only historical data available but also cultural artefacts (eg books) covering the theme of Stalin's purges. In the case of my project, I focused on the literature. No similar projects were discovered during the research.

Considering all this, the main goals of ongoing project is to propose a new approach to presenting the results of literary studies research, as well as attract more attention to Gulag literature studies overall and Estonian Gulag literature in particular.

These goals are to be achieved by:

- collecting the information about the Estonian camp prose authors;

- collecting their books that are dedicated to this theme;
- providing a brief literary analysis of the books;
- preparing illustrative materials;
- connecting the means of storytelling, illustration and literary studies;
- presenting all of the above digitally (as an interactive website).

As a bachelor in Finno-ugric philology who have dedicated a thesis to Estonian camp prose, I continue researching this genre. My thesis was one of the first papers that are dedicated to Estonian camp prose as this topic remains not very well researched. This among other things makes current project relevant.

In the process of working on the thesis I have compiled an overview of the camp prose, including most noted authors, books, genre features etc. List of Estonian camp literature consists of books by such authors as J. Kross, A. Viirlaid, A. Kask, A. Helm, A. Uustulnd, R. Kaugver and some others.

When it comes to the definition of camp prose, it can be seen that often this genre is defined as one of literary movements in the history of Russian literature. In my paper though, camp prose is considered to be a literary genre that includes all books dedicated to life in Soviet camps, no matter what language they were written in and to which national literature they belong. This approach allows to analyze this genre most efficiently, as it allows to avoid excluding books that were written by authors of non-Russian descent.

During the research camp prose was analysed on the basis of a novel “Forty Candles” (“Nelikümmend küünalt”, 1966) and a short story collection “Letters From the Camp” (“Kirjad laagrist”, 1989), both written by Raimond Kaugver, one of the most well-known Estonian authors who have dedicated several books to the Gulag topic. Analysis of the stated books has led me to believe that Estonian camp prose is characterized by the following features: autobiographism; simplicity of language; fragmented composition, retrospection. It was also noted that not all of the books are wholly dedicated to the life in a camp;

some of the texts were modified due to censorship; character's emotions are often neglected. Camp prose portrays events that actually took place, but elements of fiction are also being used for various reasons. Some of those statements can be confirmed and/or illustrated by means of digital technologies. Therefore, I propose the usage of digital technologies to continue researching this genre.

In the process of research I have made a decision not to concentrate on computational methods at this stage of a project due to the small amount of available digital versions of chosen books. Instead, the focus of the project has switched to visual aspects and storytelling as tools of presenting manual literary analysis.

As it was already stated, one of the key features of R. Kaugver's style is fragmented composition: short chapters, unexpected endings, lack of connection between the ending of a chapter and the beginning of a new one. Therefore, as a tool to visualize such texts it was chosen to create a number of collages that illustrate episodes of the book, as this form of visual art is also determined by creating something whole out of separated fragments.

Storytelling part of the project is presented in a form of annotation of said collages. Certain parts of illustrations are being connected with a piece of related text (an excerpt of literary analysis).

As a result, project's goal of presenting a literary analysis in a modern way is achieved through illustrations (collages) and storytelling (literary analysis presented as collages' annotations). At this point, a prototype of described website is available.

I believe that usage digital technologies as well as illustrative material and storytelling can be useful as tools to refresh the way we perceive literary analysis and allow to demonstrate research results not only to fellow researchers, but to a wider audience. Therefore, this project can become an example of modern approach to literary studies, for the form in which the literary analysis is presented is not conventional. To make it

more suitable for common use and not only to scholars, the literary analysis was divided into short paragraphs that give some clues for interpreting the text. Such form does not provide in-depth literary analysis, but attracts attention to the topic and encourages people to continue analyzing the text themselves. Artworks are also helpful as means to illustrate events of the books, because most texts chosen for this project are not translated to English which makes them not available to foreign readers. Hopefully it will also attract more attention of Estonian researchers, for at this point few papers covering Estonian Gulag literature topic are available.

ID: 220
Long paper presentations
Topics: historical studies, linguistics, sociology, bibliographic studies, cultural heritage collections, data mining / text mining, linked data / semantic web / ontologies
Keywords: bibliography, demography, language community, historical sociolinguistics, metadata

Estonian Language Community ca. 1900: Learning from Linked Metadata

Peeter Tinits[1,2,3]

[1]Tallinn University, Estonia; [2]University of Tartu, Estonia; [3]National Library of Estonia, Estonia

The expansion of digital resources has provided new avenues for historical research on language in a number of ways. Based on digital text collections, corpus linguistics has become one of the core disciplines for language researchers over the last few decades. Enriched texts allow one to extract facts, people, or geographic locations from texts and allow us to better understand what people were writing about.

An intriguing option that has come to be explored more recently is the study of collection metadata themselves (e.g. see Lahti et al. 2019). That is, with the study of collections and registries themselves, it may be possible to say something substantial about the historical events too. This naturally entails a more careful consideration of how the data points end up in the archives and whether they can be considered representatitive of an era, or may be biased in some way (e.g. by focussing on authors that were later canonized by critics, cf. Algee-Hewit et al. 2016).

In this presentation, I will present explorations of historical bibliographic data (i.e. Estonian National Bibliography), that aims to give a complete and comprehensive overview of publications in Estonian or related to Estonia. It is an aggregate of various bibliographies collected by book scholars over generations, and has been now made available in a digital structured format. I explore it from two angles: 1) printed books in the context of community demographics; 2) individuals involved in writing and publishing books and their backgrounds.

This study can be seen to contribute the discipline of historical sociolinguistics that aims to study the relations between society and language use in past communities. This research area is a victim to the problem of 'bad data': linguistic interactions preserve very poorly, and even for written language, a significant part of the sources have not been preserved. As a result the researchers turn to 'informational maximalism' (Janda & Joseph 2003), looking to integrate whichever data possible to research. Turning to demographic and bibliographic datasets thus offers a natural complement to the data in the field, that has not been explored much yet. While the conclusions also depend on the quality of the data, population-level characteristics allow an alternative point of view to the past linguistic communities. A systematic use of this may in fact extend the characteristics typically observed in historical sociolinguistics, and allow new concepts to be developed that to interpret data in terms of the linguistic community.

Estonian written language community provides a particularly interesting case for sociolinguistic study. While Estonian had been written regularly from the mid 16th century, this was done mostly by Baltic German clerics and used for religious purposes. The active language use in the native Estonian community was largely oral until the mid 19th century, and during a short time until the start of 20th century, written communication established a strong position within the community. At the same time, the area saw heavy urbanization, technological improvements like the railroad and improved printing, and a growing attention paid to language. The migration within the language community and the influx of a large number of new writers could influence both the shape of the language as it is used and the position that it is seen in. Here I present explorations of the growth dynamics of the written language community at the time. Particularly focussing on spoken and written language contacts as they can be reconstructed from demographic and bibliographic data, and some administrative events in language policy and education and whether their influence can be seen in the data.

Data and Methods

The study relies on the Estonian National Bibliography (ENB), which is publically available on data.digar.ee. The data was harmonized with some heuristics and custom dictionaries. Demographic data for the period was aggregated from various published primary and secondary sources. The individuals involved in the language community were retrieved from publication data in ENB, by taking all names associated with the publications (excluding original authors of translated works). Finally, for enrichment ENB data links with VIAF were relied upon, adding biographic information to the authors based on Wikidata and DNB collections, and adding a few more sources (ISIK, VEPER). As a result, bibliographic data combined with demographic data was established, and an enriched dataset of individuals actively involved with print publications.

The urbanization component of migration was estimated based on residual migration in cities, comparing the natural growth rates estimated by demographers (Ainsaar 1997) and recorded growth across the censuses. The migration pathways were estimated on the basis of 1922 census that recorded the birthplaces of the people by county level. The native dialects of geographic areas or people were based on a published dialect map (Kyrölainen & Uiboaed 2013). The estimated population data was used to calculate per capita publishing rates over time. Additionally, aggregated data from school history (Andresen 2003) were utilized to investigate the reasons behind different rates for the emergence of authors across the counties.

Case Studies

The demographic data show that due to internal migration within the Estonian population, most cities consisted of around 50% of immigrants around the turn of the century which has been described as heavy dialect contacts. However, only a minority of them were born in a different dialect area, due to which practical influence of dialect contacts on language can be expected to be marginal in terms of spoken language.

The publication record shows an exponential growth in both the number of printed works, and number of printed works per capita, as well as number of authors per capita. This provides a foundation for a steady rise in the written language community, that is mediated a bit by political events. Publication record shows rather abrupt changes in the relative roles of competing Estonian, German, and Russian languages as result of administrative policies.

The birthplaces of the associated individuals show a dominance of Livland in the late 19th century, to the extent that cumulatively, Livland comes to dominate the intellectual population over Estland. This trend can be understood in terms of administrative policies, that resulted in Livonian communities gaining affluence and also good coverage of public schools a decade or two earlier. However, between the north-south split of Estonian dialects, northern dialects are also dominant in large parts of Livland, so among the written language community, speakers with a native northern language still dominate.

Conclusions

The study of collection metadata provides an intriguing avenue for humanities research. It also opens up new discussions, particularly on the potential representativeness of the collections and the different ways that data points could be harmonized or generalized from. While these discussions may take a while to take place, opening up collections as datasets, and making them structured and machine-readable, is a clear step towards exploring these possibilities. In the case studies presented here, the encyclopedic metadata was used to study the shape of an emerging language community more than a 100 years ago. The same datasets, and other datasets like it, could be used to study many different questions relevant to humanities scholars of different fields. The more datasets become interlinked to each other, the more their investigative as well as critical potential can be taken advantage of. In this paper, we have gathered the historical demographic, bibliographic, dialectal, and administrative data to characterize a past language community.

References

- Ainsaar, M. 1997. Eesti rahvastik Taani hindamisraamatust tänapäevani = Estonian population from Liber Census Daniae up to nowadays. Tartu: Tartu Ülikooli Kirjastus

- Algee-Hewitt, M., Allison, S., Gemma, M., Heuser, R., Moretti, F. and Walser, H., 2016. Canon/archive: large-scale dynamics in the literary field. Literary Lab Pamphlet 11

- Andresen, L. 2002. Eesti rahvakooli ja pedagoogika ajalugu. / III, Koolireformid ja venestamine (1803–1918). Tallinn: Avita.

- Janda, R.D. & Joseph, B.D. 2003. On language, change, and language change – or, of history, linguistics, and historical linguistics. In: B.D. Joseph and R.D. Janda (Eds.), The Handbook of Historical Linguistics. Blackwell, Oxford. 3–180.

- Lahti, L., Marjanen, J., Roivainen, H. and Tolonen, M., 2019. Bibliographic Data Science and the History of the Book (c. 1500–1800). Cataloging & Classification Quarterly, 57(1), pp.5–23.

- Uiboaed, K. & Kyröläinen, A. Keeleteaduslike andmete ruumilisi visualiseerimisvõimalusi. Eesti Rakenduslingvistika Ühingu Aastaraamat 11, pp. 281–295

ID: 106
Long paper presentations
Topics: historical studies, cultural heritage collections, digital resources – publication and discovery, scholarly editing
Keywords: digital editions taxation wealth early modern England

Hearth Tax Digital: New Narratives on Restoration England

Andrew Wareham[1], Jakob Sonnberger[2], Theresa Dellinger[2], Georg Vogeler[2]

[1]Roehampton University, UK; [2]Graz University, Austria

The Restoration hearth tax was the first Parliamentary tax to impose a direct levy upon householders in Britain and Ireland, which did not unleash major political unrest and/or a regime change (e.g. Poll Taxes of the late 1370s). Because of its success at the political level, there are a remarkable number of extant records in national and local archives on tax payers, locations, numbers of hearths, and whether they were charged/uncharged (assessments) or paid/did not pay (returns). These data present economic historians with a substantial opportunity to provide a new understanding of social and economic conditions in the late 17th century Britain. The first part of the paper will discuss why it is useful to have hearth tax records in a digital format, and the second part will present some preliminary research findings from Hearth Tax Digital. This will not only use GIS to assess distributions of population and wealth iin diachronic and national contexts, but also draw upon extraneous data on occupations, rank and gender.

Since 2000 there have also been important developments in digital transcription and archiving. On ScotlandsPlaces (National Records of Scotland (NRS) website), all the hearth tax returns arising from the 1691 collection can be searched; and on British History Online (BHO) there is an Access database of the 1666 Lady Day return for London and Middlesex. Between 2020 and 2015 Hearth Tax Online (HTO) made PDFs, reprinted from hard-copy hearth tax editions, available electronically. Each of these methods has distinct advantages and disadvantages, dependent upon the aims of

BHO, HTO and NRS. BHO maximizes users' ability to manipulate the data, but does not enable users to read the 1666 return in its original order. ScotlandsPlaces is at the opposite end of the spectrum, taking careful note of manuscript marks, but with limited facility to manipulate the data; and HTO was best used in tandem with the hard-copy editions from which the printed transcripts were taken.

Hearth Tax Digital, arising from a partnership between the British Academy Hearth Tax Project/Centre for Hearth Tax Research at the University of Roehampton and the Centre for Information Modelling at the University of Graz uses the methods of an assertive digital edition to achieve 5 aims:

1. digital archiving and long-term preservation of hearth tax records

2. access to the digital transcripts in the original order in which they were written

3. manipulation of the statistical data synchronically/county based and diachronically/nationally

4. depiction and research enquiries on population/wealth distribution in GIS

5. searching based upon extraneous data on social conditions/rank/occupations etc. with standard data

The new website is hosted by the FEDORA-based, OAIS-compliant humanities digital archive infrastructure of Graz University (GAMS), a repository both for long-term archiving and publication of digital humanities resources. Hearth Tax Digital, essentially, is built upon two types of digital sources.

Firstly, for some regions we have been granted access to transcripts of the original records, which were produced for the print editions published by the British Academy Hearth Tax Project and the British Record Society. These transcripts are further encoded in XML, following the guidelines of the Text Encoding Initiative (TEI). Additionally, taking the 'assertive edition' approach, distinct semantic units are labeled using the ana-attribute. During the ingest process, a 'toRDF' stylesheet makes use of those labels, creating a graph database from the transcripts.

For other regions, lists of taxpayers are only available, lacking any contextual information or initial order given in the original documents (‘Returns in database Format’). In this case, the data – usually given in database files or spreadsheets – are directly transformed to RDF/XLM, and joined with the graph data arising from the transcripts in our triple store, forming one sole semantic database.

Notably, all these processes, once they have been set up for the project, automatically apply to all upcoming further data ingested to the repository following our schema, providing HTML and spreadsheet representations for both the transcripts and the ‘Returns in database format’, as well as adding the extracted semantic information to the database.

According to the aims of the project, it can be said that:

1. The GAMS repository, certified according to the criteria of the ‘Data Seal of Approval’ as a trusted repository, guarantees long-time preservation and archiving of all records in scope. Additionally, users may easily access and download the source data (TEI/XML, RDF) of all documents.

2. The visual representation of the digital transcripts is kept as close to the original transcripts as possible, maintaining the initial order and spelling, obtaining all conveyed information as well as trying to reconstruct the original layout (e.g. columns) of the documents. But, as the aim of a digital edition goes beyond the mere digital reproduction of the print edition, all additional information like regularizations, editorial notes, geographical hierarchies etc. have been marked up and visualized by optical highlighting and tooltips.

3. We are also able to deliver any kind of statistical information on our data just by formulating suitable database requests.

4. By adding the geographical information on county/parish boundaries (GML, Shapefiles) provided for the print editions to our database, we can visualize almost every statistic projected on various different background maps (e.g. Open Street Map). Ranges and parameters therefor can be manipulated by the users, offering a vast playground for research beyond the standard parameters.

5. The database provides both a full-text search for any terms occurring anywhere in the transcripts, as well as a structured search based on categories like number of hearths, personal names etc.

Currently (August 2019), Hearth Tax Digital holds more than 142,000 taxation entries, with further 46.000 in publication.

Hearth Tax Digital means that for the first time it is possible to study the hearth tax in a national context, moving across county boundaries and returns between the mid 1660s and early 1670s. This paper will set out both the methods which have been used in developing this digital resource, and some preliminary findings on social and economic conditions in the Restoration age.

References

Johannes Stigler & Elisabeth Steiner: GAMS – An infrastructure for the long-term preservation and publication of research data from the Humanities. In: Vereinigung Oesterreichischer Bibliothekarinnen und Bibliothekare. Mitteilungen. 71,1. 2018. 207-216. doi:10.31263/voebm.v71i1.1992.

Vogeler, Georg: The ‘assertive edition’ : On the consequences of digital methods in scholarly editing for historians. In: International Journal of Digital Humanities. 1,2. 2019. 309–322. doi:10.1007/s42803-019-00025-5

Wareham, Andrew ‘The unpopularity of the hearth tax and the social geography of London in 1666’. In: Economic History Review, 70 (2017) pp. 452–82.

ID: 205
Short paper presentations
Topics: historical studies, data mining / text mining, natural language processing
Keywords: aristocracy, newspapers, conceptual history, national identity

Foreignizing the Other: National Identity and the Concept of Aristocrat in Dutch Historical Newspapers

Leon Wessels

Utrecht University, The Netherlands

The Netherlands are commonly associated with a bourgeois culture. In a classic essay, the renowned Dutch historian Johan Huizinga emphasized that the single most important characteristic of the Dutch nation is its thoroughly bourgeois spirit [1]. Huizinga was not presenting a completely new view of Dutch culture. He was rather summarizing a common opinion of his time [2]. From the eighteenth century until at least the 1960s, the values, spirit and attitude of the middle class were widely seen as the pars pro toto of the Dutch nation [3]. As a result, historians have largely ignored the aristocratic elements in Dutch culture [4]. If they were noticed at all, they were disqualified as ‘un-Dutch’ [5]. Several studies have shown, however, that the elites ruling the Dutch Republic went through a process of ‘aristocratization’. They evolved into a closed oligarchy, especially in the eighteenth century [6], and adopted an aristocratic lifestyle exemplified by their luxurious mansions in the countryside [7].

How did language reflect the social and cultural presence of elites? In this paper, I will present some of the results of my ongoing PhD research into the broader conceptual history of the term ‘elite’ in the Netherlands. I will seek to understand how the word ‘aristocrat’ was conceptualized in Dutch newspapers between 1840 and 1994, examining in particular its spatial (in this case national) connotations. The corpus consists of articles (advertisements have been excluded) from over 30 different national and regional newspapers and contains almost 15 billion words. Newspapers are particularly interesting to study

the history of concepts, because their serial nature allows one to study change over time and because newspapers both produce and reflect public discourse [8].

Following the principles of Natural Language Processing suggested by Jurafsky and Martin [9], I have created a number of Python scripts to query the newspaper corpus. I started out by making a simple concordancer, similar to various openly available concordance tools [10]. Next, I wrote a script to generate frequency lists (per year) of words that occur close to the keyword “aristocrat”. This keyword was written as a Python list containing regular expressions that capture the Dutch words ‘aristocraat’, ‘aristocratie’, ‘aristocratisch(e)’ and compound words, in historical spelling variations. I applied this script to make frequency lists of words that occur within a window of three words of the keyword. For example, the sentence ‘De dwingelandij van de aristokratie van Spanje is alom bekend.’ (The tyranny of the aristocracy of Spain is widely known.) would add the following words to the frequency list: ‘van’ (2), ‘de’ (1), ‘dwingelandij’ (1), ‘is’ (1), ‘spanje’ (1).

The next step was to build a historical gazetteer suitable for extracting spatial information from the word frequency lists. A gazetteer is a geographical dictionary containing references to countries, regions, place names, et cetera. To avoid so-called ‘temporal dissonance’ I did not use an existing modern Dutch gazetteer, but created a historical Dutch gazetteer following the principles of McDonough et al. [11]. This gazetteer includes historical spelling variations and references to states that no longer exist. Using this gazetteer, I extracted references to nations from the word frequency list and saved the results as a tabularized set of data.

The resulting data were used to analyze how frequently references to various nations co-occurred with keywords related to the concept of aristocrat. Among other things, the analysis shows a clear tendency in Dutch newspapers to associate the concept of the aristocrat with foreign countries, in particular Great Britain. References to a domestic aristocracy on the other hand are only marginally present. My research thus shows that the concept of the aristocrat – as the

counterpart of the burgher – was effectively foreignized. This conclusion is in keeping with the generally held image of the Dutch as thoroughly bourgeois, in spite of the actual existence of an indigenous aristocracy.

In preparation for the DHN 2020 conference, two more steps will be taken to improve the methodology. So far, the research was based on absolute frequencies of co-occurences. The first step will be to use so-called ‘significant collocation’ to identify which words co-occur more often than would be expected based on statistics alone [12]. Secondly, in order to capture the relations with semantically similar words, such as synonymy and hyponymy, I will use synsets. Synsets are sets of cognitive synonyms that are interlinked based on semantic and lexical relations. This approach has been successfully applied also by other researchers to study historical and geographical concepts [13]. Using synsets the term ‘aristocrat’ can thus be analyzed at a more conceptual level [14].

References

1. Johan Huizinga, “Nederland's geestesmerk”, in: Geschiedwetenschap / hedendaagsche cultuur. Verzameld werk VII (Tjeenk Willink I\& Zoon N.V., Haarlem 1950) pp. 279-312. Originally published in 1935.

2. Henk te Velde, “How High did the Dutch Fly? Remarks on Stereotypes of Burger Mentality”, in: Annemieke Galema, Barbara Henkes and Henk te Velde eds., Images of the Nation. Different Meanings of Dutchness, 1870–1940 (Rodopi, Amsterdam/Atlanta 1993) pp. 59–80.

3. Remieg Aerts, “De erenaam van burger: geschiedenis van een teloorgang”, in: Joost Kloek and Karin Tilmans eds., Burger. Een geschiedenis van het begrip ‘burger’ in de Nederlanden van de Middeleeuwen tot de 21ste eeuw (Amsterdam University Press, Amsterdam 2002) pp. 313–345.

4. Conrad Gietman, “Adel tijdens Opstand en Republiek. Oude en nieuwe perspectieven”, Virtus. Journal of Nobility Studies 19 (2012) pp. 49–62.

5. Willem Frijhoff, “Verfransing? Franse taal en Nedderlandse cultuur tot in de revolutietijd”, BMGN – Low Countries Historical Review 104.4 (1989) pp. 592–609.

6. H. van Dijk and D.J. Roorda, “Sociale mobiliteit onder regenten van de Republiek”, Tijdschrift voor Geschiedenis 84 (1971) pp. 306–328; Yme Kuiper, “Adel in de achttiende eeuw: smaak en distinctie. Een verkenning van het veld”, Virtus. Journal of Nobility Studies 16 (2009) pp. 9–18.

7. Paul Brusse and Wijnand W. Mijnhardt, Towards a New Template for Dutch History. De-urbanization and the Balance Between City and Countryside (Waanders/Utrecht University, [Zwolle/Utrecht 2011]); Yme Kuiper and Rob van der Laarse eds., Beelden van de buitenplaats. Elitevorming en notabelencultuur in Nederland in de negentiende eeuw (Verloren, second revised edition, Hilversum 2014); Yme Kuiper and Ben Olde Meierink eds., Buitenplaatsen in de Gouden Eeuw. De rijkdom van het buitenleven in de Republiek (Verloren, Hilversum 2015).

8. Michael Schudson, The Power of News (Harvard University Press, Cambridge 1982) pp. 17-18; Dan Berkowitz ed., Social Meanings of News. A Text-Reader (Sage, Thousands Oaks/London/New Delhi 1997) pp. xi-xiv; Martin Conboy, The Language of the News (Routledge, London/New York 2007) pp 149-150.

9. Daniel Jurafsky and James H. Martin, Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Third Edition draft, 2018).

10. Geoffrey Rockwell and Stéfan Sinclair, Hermeneutica. Computer-Assisted Interpretation in the Humanities (MIT Press, Cambridge/London 2016) pp. 49–65.

11. Katherine McDonough, Ludovic Moncla and Matje van de Camp, “Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora”, International Journal of Geographical Information Science 33.12 (2019) pp. 2498–2522.

12. John Sinclair, Susan Jones and Robert Daley, English Collocation Studies: The OSTI Report (Continuum, London 2004) p. 10.

13. For example: Roberta Cimino, Tim Geelhaar, Silke Schwandt, “Digital Approaches to Historical Semantics: New Research Directions at Frankfurt University”, Storicamente 11 (2015) pp. 1–16; Francesca Frontini, Riccardo Del Gratta and Monica Monachini, “GeoDomainWordNet: Linking the Geonames Ontology to WordNet”, in: Zygmunt Vetulani, Hans Uszkoreit and Marek Kubis eds., Human Language Technology. Challenges for Computer Science and Linguistics. 6th Language and Technology Conference, LTC 2013, Poznań, Poland, December 7–9, 2013. Revised Selected Papers (Springer International Publishing, 2016) pp. 229–242.

14. For discussions on the relation between words and concepts, see (among others): Otto Brunner, Werner Conze and Reinhart Koselleck, Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland. Volume I (Klett-Cotta, Stuttgart 1972) pp. xxii-xxiv; Peter de Bolla, The Architecture of Concepts. The Historical Formation of Human Rights (Fordham University Press, New York 2013) pp. 19–26.

ID: 158
Long paper presentations
Topics: linguistics, interdisciplinary collaboration, sustainability and preservation, teaching / pedagogy / curriculum design, digital activism
Keywords: language endangerment, language technology, digital resources

From Research to Revitalization: Fighting Language Endangerment with Digital Humanities

Joshua Wilbur

Tartu University, Estonia

While the number of languages facing endangerment has increased at an alarming rate in recent decades, linguistics research on such languages has also seen a relative increase in attention and funding at the same time. Revitalization efforts are a common attempt to lessen language endangerment, and such ventures are in their essence practical undertakings; these are aimed at bringing about real-world change specifically concerning the status of the relevant endangered language, especially aimed at maintaining and ideally increasing the number of speakers. On the other hand, linguistics research is – nearly per definition – theoretical. Recent publications on endangered languages have sometimes highlighted the importance of collaboration between researchers and members of the language communities, and indeed the obligation on behalf of researchers to give something back to the communities they work with (e.g. cf. Austin & Sallabank 2011, Grenoble & Furbee 2010, Gippert et al. 2006). This awareness of the urgency of documenting endangered languages has arisen more or less simultaneously with the the digital age; in this, digital tools and computational approaches to linguistics have been and continue to be developed.

With the above state of affairs as a background, I have been working on documenting and researching linguistic structures in Pite Saami, a critically endangered Uralic language spoken in northern Sweden by currently approximately 35 speakers, for more than 10 years. Although a steady decline in the number

of native speakers has taken place over the last decades, interest in revitalizing the language has increased significantly over the last ten or fifteen years. In recent projects, I have adapted and developed language technology tools to help streamline my research, especially aimed at increasing efficiency in corpus annotation (partly in collaboration with colleagues with similar projects, as discussed in Blokland et al. 2015, Gerstenberger et al. 2016, Gerstenberger et al. 2017). Members of the language community have been active in their own revitalization projects as well, including a multi-year project to create a wordlist of Pite Saami words with Swedish translations (which ultimately led to Bengtsson et al. 2016), and the development of an orthographic standard which attained official status in 2019.

The question I pose now concerns the overlap between revitalization efforts and my linguistics resources in the Pite Saami context as a case study for how digital practices in humanities research can influence practical reality. Specifically, I look into what the expectations and needs of the language community in their revitalization efforts are, and how the digital resources I have developed mainly for more theoretical linguistics goals in mind can be successfully utilized in those efforts.

To begin with, the following are some of the needs and expectations on behalf of the language community which can be recognized:

1. Pedagogical materials, particularly for beginners, such as textbooks for use in language classes

2. Higher social status to improve the public sense of importance, relevance and value concerning the endangered language

3. Increased support for native speakers to be able to use their language in public, especially when dealing with the authorities

4. Lexical resources, both for native speakers (including monolingual definitions), and especially for non-native speakers (including translations into the local majority

language, and grammatical information to show how each lexical item behaves in the linguistic system)

5. An orthographic standard so that written communication is possible, and pedagogical materials write language in a consistent way

6. Keyboard layouts with the necessary characters for the operating systems used by speakers

7. An internet presence to promote the language online

8. Language recordings, especially for use in teaching activities

While these can be relevant for many communities facing language endangerment (some of these are discussed in more detail in Holton 2011), this list is based on my own familiarity with the situation for the Pite Saami community.

Thanks mainly to research projects on Pite Saami, the following digital resources currently exist for Pite Saami:

(A) An extensive transcribed and annotated digital corpus of spoken language recordings in international language archives (including audio and sometimes videos)

(B) Language technology tools (specifically Finite State Transducer and Constraint Grammar)

(C) A lexical database with grammatical information, as well as Swedish and English translations of nearly 6000 lexical items

(D) A web resource explaining the orthographic rules

(E) A few digitized transcriptions of written heritage materials

(F) A mobile phone app (for Android) with more than 4000 lexical items and translations into Norwegian, Swedish, English and German, along with example senteces

These resources were created as part of past and current research projects on Pite Saami by the author,[1] with the exception of item (F) (the mobile phone app).

[1] This was possible in part to generous grants from the Hans Rausing Endangered Langauges Project, Duoddara Ráfe Pite Saami center, and the German Research Foundation (DFG), as well as collabora- tive work with Giellatekno at the University of Tromsø.

The clearest example of how such outcomes can directly support the language community concerns the orthography standard (number 5 above), which only received official status in late August 2019. Specifically, the web resource documenting the orthographic rules (item (D))[2] provides an easy to access way to follow the writing standard, including audio samples for most example words. In addition, this adds to the web presence for Pite Saami, which support need number 7. However, although this resource is freely available, and most members of the language community have access to the internet, the site does not receive many visitors.

The corpus mentioned in item (A) also help promote the language's status by proving that the language is a valuable object of scientific research and funding at an international level,[3] thus support requirements 2, 7 and 8. On a related note, a limited number of heritage texts (item (E)) have been digitized, transcribed and will be published as part of the archived corpus in the future. Language technology tools (as presented below) are used to automatically annotate these texts for lemma, part of speech, morphological information and English glosses,[4] thereby significantly increasing the amount and consistency of annotations.

Language technology tools make it possible to generate and analyze Pite Saami word forms,[5] and include a derived spell-checker[6] for Libre-Office. In a similar way, the online lexical database (item (C))[7] is regularly updated, corrected and extended, and provides a searchable interface for lexical items, translations and some grammatical categories, most items also have audio files for users to listen to. While this version lacks inflected wordforms, another web version[8] is available which

[2] http://saami.uni-freiburg.de/psdp/stavningsregler
[3] Corpus materials can be found in the Endangered Language Archive (ELAR) in London and in the Lan- guage Archive (TLA) in Nljmegen/NL.
[4] Cf. Blokland et al. 2015, Gerstenberger et al. 2016, Gerstenberger et al. 2017
[5] This is done using Finite State Transducer and Two-level morphology.
[6] A beta version of a Pite Saami spell-checker is available from Divvun at http://divvun.org/proofing/proofing.html.
[7] http://saami.uni-freiburg.de/psdp/pite-lex/
[8] Available from Giellatekno at https://bahkogirrje.oahpa.no/

generates inflectional paradigms for most lemmas (but only includes minimal translations and other grammatical information). These aspects clearly support needs expressed in 4 and 7 concerning lexical resources and the internet presence, and can increase the chances of the main requirement about creating textbooks (number 1).

Four of those heritage texts (item (E)) are available online[9] as possible resources for language teaching as well. Hovering over a word in these texts with the cursor triggers a floating bubble with information about grammatical information and an English translation; the grammatical information is derived from the language tools mentioned above. Thus also increases the internet presence (number 7) and can be used for teaching (number 1).

Last but not least, a mobile phone app (item (F)) is under development[10] which clearly has significant potential to support the language community by making Pite Saami seem modern and valuable. This should be especially important for learners of Pite Saami, who are mainly from younger generations.

While a number of the requirements and wishes listed above are at least partially promoted by extant digital resources, it likely that any effects concerning the increased use of the language in public, especially when dealing with the authorities (number 3). As for need number 6 about the existence of keyboard layouts in operating systems (both computer and mobile) in the two majority languages Norwegian and Swedish will have to suffice for now, despite the fact that these frequently lack specific characters needed for writing Pite Saami (such as á, ŋ and đ), or at least users tend to lack knowledge of how to find such characters on the respective keyboards.

Considering the number of native speakers of Pite Saami, and the correspondingly small size of the ethnic population and group of interested second language learners, Pite Saami has

[9] http://saami.uni-freiburg.de/psdp/texter/

[10] The app is available for the Android operating system, and the beta version can be downloaded from the public facebook group BidumBágo.

a rather impressive collection of digital resources at its disposal, as outlined above. However, due to the arrival of these resources very late stage – at a point when the language is critically endangered and at significant risk of no longer being spoken actively in the foreseeable future – it has yet to be determined to what extent these digital resources will in fact contribute to revitalization efforts in a significant way. At the very least, digital opportunities are not being ignored, but being taken advantage of, to hopefully assist in warding off a disastrous fate. The coming years will show if this has been enough to help revitalize Pite Saami.

References

Austin, Peter K. & Julia Sallabank, eds. (2011). The Cambridge handbook of endan- gered languages. Cambridge handbooks in language and linguistics. Cambridge: Cambridge University Press.

Bengtsson, Nils-Henrik, Marianne Eriksson, Inger Fjällås, Eva-Karin Rosenberg, Gry Helen Sivertsen, Valborg Sjaggo, Dagny Skaile, Peter Steggo, & Joshua Wilbur (2016). “Pitesamisk ordbok”. In: Pitesamisk ordbok samt stavningsregler. Ed. by Joshua Wilbur. Samica 2. Freiburg: Albert-Ludwigs-Universität Freiburg, pp. 13–121.

Blokland, Rogier, Ciprian Gerstenberger, Marina Fedina, Niko Partanen, Michael Rießler, & Joshua Wilbur (2015). “Language documentation meets language technology”. In: First International Workshop on Computational Linguistics for Uralic Languages, 16th January, 2015, Tromsø, Norway. Proceedings of the workshop. Ed. by Tommi A. Pirinen, Francis M. Tyers, & Trond Trosterud. Septentrio Conference Series 2015:2. Tromsø: The University Library of Tromsø, pp. 8–18.

Gerstenberger, Ciprian, Niko Partanen, Michael Rießler, & Joshua Wilbur (2016). “Uti- lizing language technology in the documentation of endangered Uralic languages”. In: Northern European Journal of Language Technology 4, pp. 29–47.

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 (IWCLUL 2017). Ed. by Tommi A. Pirinen, Michael Rießler, Trond Trosterud, & Francis M. Tyers. St. Petersburg: Association for Computational Linguistics, pp. 25– 36.

Gippert, Jost, Ulrike Mosel, & Nikolaus Himmelmann, eds. (2006). Essentials of language documentation. Trends in Linguistics. Studies and Monographs 178. Berlin: Mouton de Gruyter.

Grenoble, Lenore A. & Louanna Furbee, eds. (2010). Language Documentation. Practice and values. Amsterdam: John Benjamins Publishing Co.

Holton, Gary (2011). “The role of information technology in supporting minority and endangered languages”. In: The Cambridge handbook of endangered languages. Ed. by Peter K. Austin & Julia Sallabank. Cambridge handbooks in language and linguistics. Cambridge: Cambridge University Press, pp. 371–399.

ID: 192
Long paper presentations
Topics: linguistics, medieval studies, philology, corpus linguistics, data modeling / knowledge representation, linking and annotation
Keywords: lexicography, Old Norse

Integrating TEI/XML Text with Semantic Lexicographic Data

Tarrin Wills, Ellert Thór Jóhannsson, Simonetta Battista

University of Copenhagen, Denmark

The Dictionary of Old Norse Prose (ONP – onp.ku.dk) is an extensive digital resource which links the semantic analysis of the lexicon of Old Norse with its material record (manuscripts and charters). Its citation index of around 800,000 words represents an estimated 7% of the entire corpus of Old Norse. Only a small proportion of the corpus, which is around 11 million words, has been prepared as digital editions using modern standards. The dictionary is not complete but nevertheless at this stage contains a semantic and/or grammatical analysis of around one in every twenty words of the Old Norse corpus. This analysis is spread fairly evenly across the corpus.

The methods and data structures for the dictionary, which began in 1939, were developed before digital corpus linguistics was possible. The dictionary's methods are based on manual excerption of words and surrounding text and are not in themselves compatible with corpus-based approaches. Other projects, however, have been developing manuscript-based Old Norse digital texts that belong to the same corpus that ONP covers and use a compatible manuscript-based approach. The most extensive of these is the Menota project (menota.org) which includes a catalogue of TEI/XML manuscript texts

encoded according to a specified subset of TEI. A previous paper by the authors (Wills, Jóhansson and Battista 2018) describes a fast and user-friendly workflow whereby Menota texts can be linked at the lexical level to dictionary headwords in ONP using a combination of automated and manual stages. This workflow is designed to achieve very high levels of

accuracy (close to 99.9%) for the automated stages. The workflow demonstrates an interoperable method whereby TEI/XML encoded texts can be integrated and linked into relational data models such as dictionaries. These methods are designed to maintain a link between the two external data sources at the level of the word so that they can be edited and maintained separately. (Documentation can be found at goo.gl/ncdWAC)

Linking lemmas at the lexical level means that users can access the dictionary directly by interacting digitally with the words in the text: clicking on a word, for example, can bring up a full dictionary entry regardless of homographs, and regardless of the normalisation or lemmatisation used in the particular text edition. It also means that full concordances for a particular lemma can be generated automatically.

The current research builds on these processes to link the words of the corpus deeper into the dictionary's semantic structure. A dictionary aims for not just a lexical but also a semantic overview of the corpus. This is done in traditional dictionaries such as ONP by excerpting relevant words from the corpus and analysing every citation excerpted, building a semantic tree of how the headword is used in the texts. Every node in that tree contains a sense and a definition of that sense, forming the structure of the dictionary entry. With such dense excerption of examples in a dictionary such as ONP, it is technically possible to link a high proportion of words in a given text to a particular semantic analysis as assigned by dictionary editors. That is, a high proportion of words can be potentially specifically linked to the individual senses and definitions of the structured dictionary entry. For the user this would mean that they can find a particular sense or usage of a word in a digital edition in the specific context they are reading, according to the dictionary's analysis.

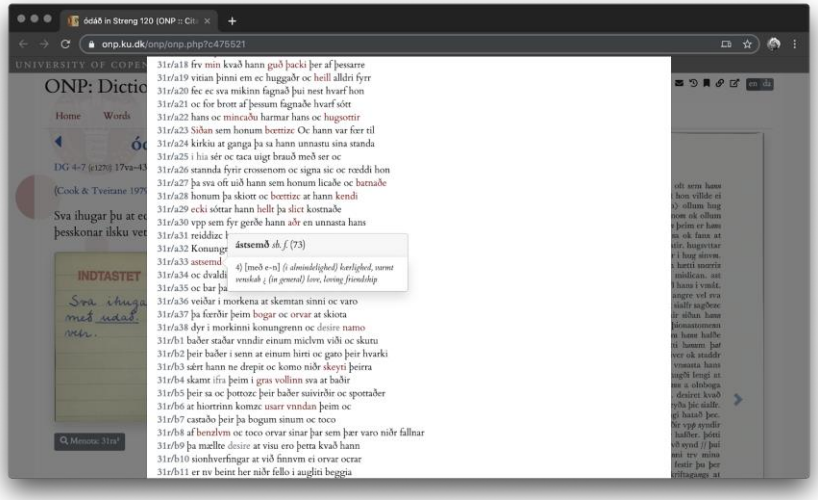
This process requires the digital linking of individual words in a text not only to the dictionary headword but to the particular citation in the semantic tree of the dictionary entry. This is a challenging task. The references in the dictionary for citations are in almost all cases to the physical page and line of the

published edition. For Menota-style TEI texts the words can normally be identified by the page and line of the manuscript version of the text. The two sets of references are in the same order but are not otherwise compatible. Not all words are excerpted by the dictionary, leaving no simple way of aligning and linking the two types of reference.

The methodology employed here has as its first stage to identify (by database queries) lemmas that appear only once in the TEI/XML text and which also appear only once in citations from the same text for that lemma in the dictionary. Because there is little chance for ambiguity that the word in such a case corresponds to the citation in the dictionary, the word and citation can be fairly reliably linked automatically. Accuracy is around 90% and so these links require manual checking that the citation is of the same word in the same context as the word in the text. The initial links between the words and the citation index provide a framework by which the same method can be applied to the smaller sections of text between the linked words. This again involves identifying lemmas unique to each section of text in both the manuscript-based edition and the print-based edition used by the dictionary. The process uses page and line references in each case to frame the extent of the section searched. Links are inserted in both data structures: in the dictionary to the word in the text, and in the text to the citation in the dictionary. This method is repeated, with decreasing gaps between the identified words, until no further automatic linking is possible. The remaining words tend to either be unlinked by the previous lemmatising processes, or are ambiguous in some way within this process and are therefore excluded.

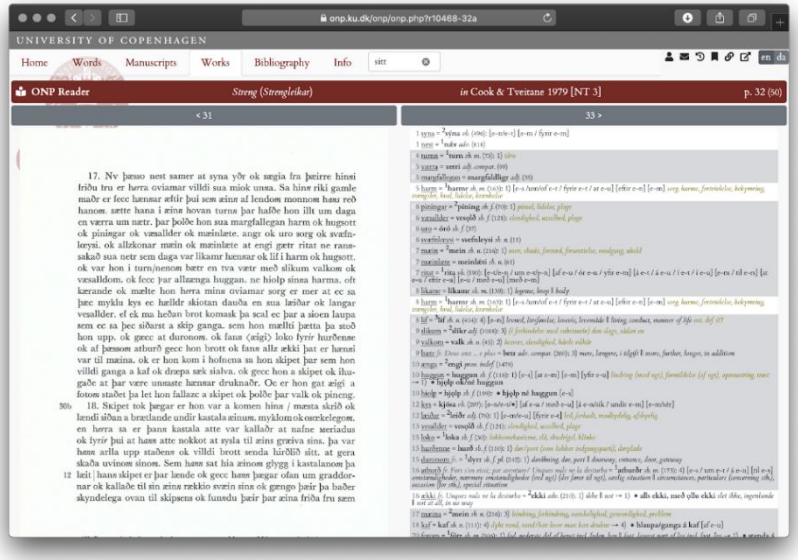
The result is that a very high proportion of the citations from a given Menota text can be quickly and accurately linked to and from the word in the text and dictionary. These links (as URIs and/or database keys) represent the minimal information needed to connect the words in each resource and are maintained even as the texts and the dictionary continue to be edited and developed separately.

For the user this linking means that when they access the text, the individual words are not only linked to the dictionary entry, but in a good proportion of instances they are linked to the individual definition and/or phrasal-grammatical context of the word as defined in the dictionary. A user – for example a student or researcher – can pull up a section of text and click on any word to get the dictionary entry, if available. Words linked at the citation level can be highlighted to indicate that further information is linked and when clicked will show the individual definition for the word, if available, and other information that the dictionary may record about that particular citation, such as the citation slip and edition information. Users of the dictionary can find specific examples for usages and can access the full text where that usage occurs, rather than the minimal surrounding text normally provided for each citation. (For an example see o np.ku.dk/c475521 and click on the Menota button. The red coloured words are linked to other citations in the dictionary, many of which have been defined.)



At this stage two texts have been extensively linked to ONP using this method, the first being Strengleikar (the Old Norse version of the Lais of Marie de France) in Uppsala manuscript

DG 4-7 (onp.ku.dk/r10468). The method described above automatically and accurately linked 3168 of the 4065 citations in ONP to the Menota edition, representing 8% of the whole text. A page from the edition is shown below with an overview of the excerpted words in the dictionary's database (onp.ku.dk/r10468-32a). This indicates which excerpted words were linked automatically (grey) and which could not be accurately linked by the automated process (white).



The advantages of the linking method described here include the automated generation of integrated glossaries for the text edition. Such glossaries can be used as a reading aid, for those less familiar with the language, and as a language learning tool. Glosses assist in language acquisition by improving text comprehension and aiding in vocabulary acquisition (Lomicka 1998). This applies also to digital glosses and 'authentic' texts (Abraham 2007), such as those used in this project.

The semantic analysis of significant portions of the text can be further developed if the dictionary at a later point, as is hoped, integrates a digital thesaurus. The thesaurus, when linked to particular senses in the dictionary, can be integrated into the text itself, potentially creating a semantic map of the text as a whole and helping users to find semantically similar material in the corpus. Lastly, the majority of words, those which are not analysed by the dictionary project, can be semantically analysed according to statistical or other digital methods so that the particular meanings of non-manually analysed words can potentially be predicted from those in similar contexts.

References

Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. Computer Assisted Language Learning, 21(3), pp. 199–226.

Lomicka, L. (1998). To gloss or not to gloss: An investigation of reading comprehension online. Language learning & technology, 1(2), pp. 41–50.

The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources. Gen. ed. Odd Einar Haugen. Version 2.0. Bergen: Medieval Nordic Text Archive, 2008. <http://www.menota.org/HB2_index.xml>

Wills, T., Jóhannsson, E., & Battista, S. (2018). Linking Corpus Data to an Excerpt-based Historical Dictionary. In J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.) Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 979-987.

ID: 229
Short paper presentations
Topics: philology, teaching / pedagogy / curriculum design
Keywords: distance learning, beginning language, pedagogics, teaching

Beginning Latvian and Lithuanian as University Level Distance Learning Courses ï Experiences and Reflections from the Past Two Years of Teaching

Lilita Zalkalns

Stockholm University, Sweden

The Baltic Section of the Department of Slavic and Baltic Studies, Finnish, Dutch and German at Stockholm University has offered beginning courses in Latvian and Lithuanian ever since the fall term of 2017. While it may seem unusual to teach a language over the internet with no physical contact at all, this teaching method has been shown to be especially well suited for the so-called “smaller” or “exotic” languages, that often lack sufficient student applicants for campus courses. As a point in case, both the Latvian and the Lithuanian courses have had an average of 20 registered students per term, a number which must be regarded as unusually high for these languages. Approximately 90% carry through to the end and take the final exam. About 10% of the students decline to participate in face-to-face contacts via Skype, Zoom or Adobe Connect, which could indicate any number of things, among them the possibility that the student is cheating, i.e. someone else is doing the work in the course modules, and that s/he does not want to reveal their lack of language knowledge, or it could simply be that the student is shy. These and other types of student/study observations and statistics will be presented and analyzed.

Over the past two years, the courses have successively changed based on student feedback, technological challenges and developments, and changes in teacher (my) attitudes. Among the issues I will discuss are technological problems, which for some students are a huge barrier to successful studies, and administrative issues, which can take up a major part of the teacher's allocated teaching time. Concerning

course design, the teacher must be prepared to create or find new content, as links to external study materials suddenly disappear or the materials themselves are changed. Also, the increasing student use of smartphones as their main learning platform, means that study materials must be continually redesigned with the small monitor in mind. These and other observations and reflections will be presented.

It can be concluded, that at least in Stockholm university, Latvian and Lithuanian will continue to be taught as distance learning courses, and that most likely, their scope and number will increase. In order to retain and augment student interest, the lessons learned and the experiences gained from the first two years of internet teaching should be gathered, systematized and implemented in the future language courses.

References

ECAR study of Faculy and Information Technology 2017 (https://www.educause.edu/ecar/research-publications/ecar-study-of-faculty-and-information-technology/2017/introduction-and-key-findings)

Darby, Flower: How to Be a Better Online Teacher in The Chronicle of Higher Education, April 17, 2019 (https://www.chronicle.com/interactives/advice-online-teaching)

ID: 241
Poster
Topics: historical studies, library & information science, cultural heritage collections, digitisation – theory and practice, GLAM: galleries / libraries / archives / museums, image processing, artificial intelligence
Keywords: digitized objects, authenticity of documents, historic documents, machine learning, OCR, quality enhancement

Limits of Authenticity of Digitized Objects

Alžbeta Zavřelová, Petr Žabička

Moravian Library in Brno, Czech Republic

The Moravian Library in Brno is the second largest library in Czech Republic - a legal deposit library that holds over 4 million volumes including valuable historical collections of old maps, incunabula, medieval manuscripts and old prints. It is also a research organisation whose main purpose is to carry out basic, applied and experimental research as well as software development, and to disseminate the results by means of education, publications and transfer of technologies.[1]

Our poster discusses the limits of authenticity of digital objects and the impact of currently available possibilities to modify or forge digital historical documents using methods of machine learning. The research results of the Moravian Library cooperation projects provide us with tools that make us reconsider the importance of digital interpretation and processing of the digitized objects.

In the last decades, mass digitization of library collections has become a necessity in memory institutions all around the world. During the digitisation process, many objects must be modified for higher accessibility or legibility. Such interventions include simple image enhancements (brightness, contrast or colour corrections, image cropping or stitching, etc.) or edits made using innovative digital tools. In our research practice we find multiple cases when we need to edit an image for better usability or legibility by advanced methods employing machine learning technologies. In special historical collections, basic tools are used to flatten curved pages, unfolding text lines, or scan text in narrow book binding. Many documents have been copied on microfilms and only these secondary copies could

have been digitized, although the quality of the microfilm copy was not very good. Likewise, digitisation of old audio recordings (e.g. shellac discs) extracts specific information thereby suppressing the authenticity of the original.

Noise reduction, quality enhancement and content reconstructions improve the comprehensibility and may also influence the quality of automatic conversion of the digitized object to text. Tools for automatic classification or full text indexing of the OCR often work with language models that may highly affect the content of the resulting text. Although such technologies do not misguide users intentionally, they may significantly influence relevance ranking or findability of a given document. Since the 90s, the scientific community has drawn attention to the possibilities of open digital images misuse [2], which has expanded significantly with technology development.

Critical Approach to the Digital Copy

While working with digital libraries we always have to think about the relationship between a digital object and its (analogue) source. Many people incorrectly attribute the same value to the original analogue and digital copy of an object. We should bear in mind that each collection is a subjective interpretation of the collection creator's point of view or ideological attitude. Moreover, until any given collection has been digitized if full, there are layers to its online presentation: Has the catalogue been converted into a searchable metadata database? Are all the catalogue records of the same or similar quality? What was the selection process when parts of the collection have been digitized? What is the quality of the conversion to text? The user of any collection should also bear in mind that any human-created metadata might reflect the bias of the cataloguer who created the record.

The intentional document forgery has always been around. Nowadays, as in the past, the main motivation for document forging is either financial profit, privilege, power or influence gain. With the ready availability of online content, the ties between the physical document and its digital surrogate are weakening. In some societies, outright censorship can readily

be applied. However, without tangible assets readily accessible, there is a growing opportunity to influence what has long been perceived as reliable information sources. Even a small modification in the text, e.g. a change in names or in linguistic and typographic phenomena, may produce the desired results for a certain group. Digitized historical collections are believed to be fully credible by its nature. Today, the general public learns how to critically assess the resources of social space, yet the level of critical approach to digitized historical documents is still limited.

How to Modify Historical Documents?

The digitized collection of the Moravian Library in Brno contains, among other documents, digitized microfilms of several newspaper titles. The microfilms were created in-house in the 1990s. At that time a lot of effort was put into finding all missing issues, which were borrowed from a number of other institutions. The microfilms themselves are therefore unique in their relative completeness. The quality of the pictures on the film, on the other hand, varies. When the films were scanned, it was impossible to have a usable OCR from the scans; and in some cases, the texts on the scans were very difficult to read even for humans. In some cases, the pages of the original newspapers have been damaged and parts of the text were missing [3].

To solve this problem, the library partnered with the Brno University of Technology. The “PERO – Advanced content extraction and recognition for printed and handwritten documents for better accessibility and usability” project aims to create technology and tools to improve accessibility of digitized historical documents based on state of the art methods of machine learning (convolutional neural networks), computer vision and language modelling.[4] The results of the project are available at GitHub[5] and will be integrated into Lindat/Clariah-CZ – Digital Research Infrastructure for Language Technologies, Arts and Humanities [6].

The main result of the project will be an automatic, machine-learning-based OCR tool for printed documents[7] and a semi-automatic handwriting recognition tool for current manuscripts.

These tools will be complemented by a system for the improvement of the quality or rather readability of images of scanned text. The tool is based on Generative Adversarial Networks (GANs) and its aim is to automatically propose replacements for missing or illegible text string by methods of text reconstruction and language modelling, and then to fix the image using the reconstructed text. At this moment there is a need for manual supervision but the benefits of this method for document legibility are obvious. Also obvious are its dangers: the operator can easily change the text in any way just by editing it and the software will then reconstruct the relevant part of the image so that the change will be indistinguishable from the surrounding unmodified text. Of course, the quality of the patch will be dependent on the tool having enough scans to learn from, which is not a problem when dealing with scanned periodicals. The moral implications are clear.

To sum up, the main purpose of the poster is to contribute to the discussion on the limits of authenticity of digitized objects and current possibilities of text manipulation. Our research demonstrates various forms of interventions to digitized documents are possible, as well as an extreme case of text manipulation using the tools developed to improve the accessibility of our digital library data. To combat this, some kind of visual cues or other functionality should be developed. To avoid Orwellian future, the users must have a way to check the authenticity of a digitized document and perhaps also be allowed to see the unmodified images as well and perhaps also have a visual cue with the modified parts of the image highlighted.

References

[1] Further information on the website of the Moravian Library in Brno: Research and Development lab <https://www.mzk.cz/en/about-library/research-and-development> or <https://www.mzk.cz/en/lab>

[2] This paper discusses the current possibilities of breaking the authenticity of documents, it does not open up any specific ethical issues related to technology development.

[3] Newspapers were printed on acidic paper which led to faster degradation and fragility of the physical objects.

[4] Further information about the project “PERO – Advanced content extraction and recognition for printed and handwritten documents for better accessibility and usability” on <https://pero.fit.vutbr.cz/>.

[5] You can easily download applications and tools from the PERO project at GitHub: <https://github.com/DCGM>.

[6] The aim of LINDAT/CLARIAH-CZ, large research infrastructure which acts as a distributed national node of the pan-European DARIAH-EU network, is to enhance the accessibility and usability of open digital data sets, resources and tools for Digital Humanities. <https://lindat.mff.cuni.cz/>

[7] An automatic technology for content extraction and an excellent OCR for early printed books (-1800/1860).

ID: 109
Short paper presentations
Topics: software design and development, computational science, artificial intelligence
Keywords: heritage monitoring, lidar, AI, 3D, software

3D and AI Technologies for the Development of Automated Monitoring of Urban Cultural Heritage*

Tadas Ziziunas, Darius Amilevicius

Vilnius University, Lithuania

Preservation of urban heritage is one of the main challenges for contemporary society. It's closely connected with several dimensions: global-local rhetoric, cultural tourism, armed conflicts, immigration, cultural changes, investment flows, infrastructures development and etc. Nowadays very often organizations responsible for heritage management constantly have to deal with lack of resources, which are crucial for proper heritage preservation, maintaining and protection. Particularly it is problematic for countries with low GDP or unstable political situation.

The possible solution of these problems could be automated heritage monitoring software system, based on the 3D data and AI technologies, which increase monitoring efficiency (financial, timewise, and data objectiveness factors). The system prototype was developed and tested by Vilnius University and Terra Modus Ltd. in frame of project "Creation of automated urban heritage monitoring software prototype" (2014). Next step is creation of full-capability software which is under development by Vilnius University on framework of project "Automated urban heritage monitoring implementing 3D and AI technologies". Project financed by Research Council of Lithuania (project time 2018–2022) . At this paper only general pipeline of the 1st stage of project is presented.

* This paper is published in DHN2020 Proceedings: http://ceur-ws.org/Vol-2612

Proposed digital monitoring technique is based on effective reality capture and comparisons of data in time. 3D laser scanners and digital photogrammetry are the most capable, accurate enough data collection methods. Collected information from different time period measurements could serve as data for artificial intelligence analysis, which can automatically identify needed valuable elements and its changes during the particular time period. Such monitoring can possibly be performed in a remote, non-destructive, and cost-effective way. Accordingly, main principles of suggested solution are listed below.

Digital monitoring is based on seven conditions. First: all objects in the monitoring process are tangible. Second: physical valuables could be expressed as simple geometrical forms or mathematical expression. Third: monitored objects could be fully scanned of photogrammetrically processed. Fourth: data from Lidar devices and data derived from photogrammetry are same quality (density, coverage, etc.). Fifth: detection of cultural heritage could be analysed by static and machine learning algorithms. Sixth: digitally processed results should be able to be checked. Seventh: digital monitoring is based on non-destructive and non-invasive 3D view technologies and analytical technologies.

Regarding of digital data there are two possible ways to perform detection and comparison of selected valuables. First case scenario mainly means lack of comparable data of the older status quo. This means that there are no earlier 3D data of selected cultural heritage. Newly collected data is compared with mathematical rules which can be written in coded form. These set of rules describes geometrical parameters of selected valuables of the cultural heritage. In the second case scenario there are two data sets from different time period. This data is compared with each other. In both cases comparison needs interpretation.

The first level of interpretation is in demonstrating some facts of geometrical change. The second level depends on the particular legal status and local legislation for managing cultural heritage, e.g. meaning of detected changes depends on

legislation). First level of interpretation could be evaluated by logical operators, for example alteration is described as “status quo unchanged”, “reduction in volume by 65%”, etc. Second level of interpretation could be legal analysis of first level results, for example, “reduction in volume = fact of illegal demolition works”.

According to the most frequent alteration of the Vilnius Old Town’s buildings’ valuables, a list could be stated: a) elements of the roof; b) shapes of the roof; c) cornices; d) doors; e) gates; f) the primary height and width of height buildings; g) the primary housing intensity of site; h) windows; i) chimneys. These are main valuables which can be traced in the manner of geometrical changes.

In order to perform the detection of valuables, we first need to train the AI algorithms to identify the desirable valuables from the data – 2D pictures or 3D point clouds. Google “Tensorflow” with DeepLab v. 3+ with default settings was used .

These are semantical segmentation procedures where some already annotated and trained data could be used. However, there are very little open data quality content for such topic. Hence, for performing the digital monitoring processes, a new database was established. Concerning future software’s usage for different oldtowns of Europe, only database with additional 2D pictures of elements or 3D scans are needed.

The newly established database consists of collected pictures from the main streets of the Vilnius’ Old Town. For data annotation, Labelbox is used. Currently there are 420 high-resolution photos (12 megapixels) where the first two classes (valuables) are created: windows and doors. All doors and windows are manually annotated in 420 photos. Annotations were performed so that an algorithm could identify the kind of pixels that denote windows as well as what pixels stand for doors. For performing the training task, the currently most powerful open data algorithms of Google’s Tensorflow were used. In this case, an XML file is the result of annotation. This means that the annotated information in the c++ language is described according to the standard of Pascal VOC. This standard is one of the most popular and widely used. To sum

up, two types of files are exported from Labelbox: XML and JPG. The further process could be described as follows:

1. JPG and XML are converted into RGB. The results are PNG files with segmentation masks – SegmentationClass;

2. Additionally, some PNG raw files with a semantical segmentation object contour are exported – SegmentationClassRaw;

3. JPG, PNG files (SegmentationClass) and PNG files (SegmentationClassRaw) are manually separated into two parts: “Train” (for training) and “Val” (for validation). The Train part is also automatically separated into tech and test parts in order to identify how accurate the training results are compared with human manual annotation. Hence, some extra Train, Val and Train/Val index are generated;

4. According to an index of JPEG, PNG, and PNG (Raw) files, we generated files special formats that were required by Tensorflow training – TFRecord (Train, Val, and TrainVal);

5. The system is trained using TFRecord files. In order to get the most accurate results, many hyper parameters should be optimized. This process is analysed in detail by J. Bergstra and Y. Bengio.

One of the biggest problems with hyper parameter optimization is overfitting. In the context of heritage monitoring, this would cause that newly presented valuables – windows, for example – could not be identified properly. In order to avoid overfitting, various techniques could be applied, e.g. early stopping. Once the progress shows that mistakes stopped reducing, all processes are then being stopped. That calculation of the quality of prediction is described as loss function. There are various methods on how to calculate the loss function, but in this experiment, a default “cross entropy” is used. The experiment results demonstrated that training progress was performed properly because the loss function was gradually decreasing and data were not overfitted. However, a powerful computer resources are needed for finalizing the whole experiment with all groups of valuables.

To sum up, presented project is still at an early stage; however, the results of the first laboratory experiments with the primary version of the pooled data resource achieving 80% accuracy in semantic segmentation of objects into two classes (windows and doors) suggest that the chosen technology solutions and developed methodology will be adapted successfully to achieve project objectives.

ID: 149
Long paper presentations
Topics: historical studies, corpus linguistics, data mining / text mining, natural language processing, big data
Keywords: newspapers, topic modeling, topic modelling, word embeddings, vector space

Disappearing Discourses: Avoiding Anachronisms and Teleology with Data-Driven Methods in Studying Digital Newspaper Collections

Elaine Zosa, Simon Hengchen, Jani Marjanen, Lidia Pivovarova, Mikko Tolonen

University of Helsinki, Finalnd

Newspapers have been a rich source of information for historians for the past hundred years or so. In the past twenty years, digitization of newspapers has made it possible to do simple tasks such as keyword searches or more elaborate text mining analyses. Advancements like this create unprecedented possibilities to the analysis of historical sources. While there is some truth to the promises of the future, the reality is such that the research on digitized newspapers remains underdeveloped with regard to reference corpora and reproducibility of the research. Digitized newspapers are particularly discussed with respect to the development of public discourse, but the idea of entering the realm of past discourse in toto through the digitized newspapers may in the end be harmful. In reality, historians are interested in the different layers of newspaper publicity, thus location and temporality always play a crucial role of any historical analysis of public discourse in newspapers.

With these aspects in mind, this paper takes advantage of digitized newspapers and data-driven approaches in identifying disappearing discourses in newspapers. In doing this, we want to revisit one of the key tensions in historiography, that is, the interplay between being relevant for the present and at the same time writing history in a way that is true to the experiences of past actors. History's presentism is sometimes discussed critically from the perspective of anachronism or teleology in

history (Koselleck 2010; Skinner 2002), or more appraisingly in terms of genealogies of the present or letting all be the history of the contemporary (Armitage forthcoming). Regardless of the historian's desire for contemporary relevance or for historical antiquarianism, the option to approach history without predefined questions from the present has not been possible. The advent of digitized sources that can be approached in a data-driven way opens up for a possibility of approaching history in a much more open-ended way. Hence, we propose to test the possibility of studying a historical case with as few presupposed categories as possible. To do this we study digitized newspaper collections (specifically, 19th century Finnish newspapers in Finnish and Swedish) through the perspective of discourses that fell out of fashion and disappeared from long-term diachronic newspaper data sets.

We believe there is more potential in the use of digitized newspapers when we are not pinpointing the words and concepts in our approach a priori. This may lead us to completely new avenues of research, challenge our take on history as some sort of progression and, hopefully, show the value of the data-driven approach for the humanities. To understand the boundaries and the development of the public sphere it is useful to identify those discourses that were important in a particular time and place, but have since disappeared while words and concepts of another discourse have replaced them and started to dominate the ecosystem of print publicity. It is a commonplace to note that religious discourse has lost much of its prominence or that technological advancements have brought with them new topics that have replaced old ones. Still, by turning the question around and asking which discourses disappeared, we get a broader picture. We then turn to the data again and zoom in on localities and languages in order to avoid a totalizing view and move on to looking at where and when discourse changed. Thus, while we produce an analysis of public discourse in Finland, we approach the topic by noting that this is not a unified whole, but composed of different entangled realms of public discourse (Tolonen et al 2019; Marjanen et al 2019a).

Using newspapers and periodicals data in Finnish and Swedish encompassing respectively 5.2B and 3.4B tokens (National Library of Finland 2011a, 2011b), we utilise two different methods: relative word frequencies as proxies for particular discourses enhanced with distributional semantics derived from diachronic word embeddings (Kim et al 2014, Dubossarsky et al 2019), and dynamic topic modeling that captures more general themes. The former method, i.e. the combination of frequency analysis and vector space similarity allows us to focus on specific themes and track their dynamics along a timeline to detect crucial events related to those themes. This has successfully been carried out by recent previous work on similar data (Martinez-Ortiz et al 2016; Hengchen et al 2019; Marjanen et al 2019b; van Eijnatten and Ros 2019). Training diachronic word embeddings on different time granularities (e.g. months, years, or decades) allows for different views on the evolution of semantic clusters – these themes are then given weight through frequency counts. The latter method allows us to paint a larger picture of the different dynamics taking place in the data, by harnessing the power of topic models designed to capture trends in time-series data such as Dynamic Topic Models (DTM, Blei and Lafferty 2006). In DTM, the data is divided into discrete time slices and the method infers topics across these time slices to capture topics evolving over time. This method models how a topic changes from one time step to the next. Unlike vanilla LDA topic modelling which does not take into account the evolution of a topic, DTM is more robust to topics that changes vocabulary over time to talk about the same issue. In LDA, topics like these would likely to be separated into separate topics since the words associated with them has changed but in DTM they would be treated as one topic that is developing over time. To address the additional training complexity of this model we subsample the data such that we have the same amount of data for each time slice of our corpus. This would also ensure that the topics inferred are representative of all the time slices in the corpora rather than favoring the latter years which have more articles and newspapers associated with them.

With thematically-labelled temporal representations of newspaper data, it becomes possible to quantify and qualify the evolution of certain themes that have been automatically inferred from the data – thus removing some bias in topic selection. We further use metadata to zoom in on changes in topics to see which towns, regions or types of newspapers to manually assess the driving locations of change and to produce a typology of disappearing discourses.

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye).

References

1. Armitage, D. (In Press). In Defense of Presentism. In D. M. McMahon (Ed.), History and Human Flourishing. Oxford: Oxford University Press.

2. Blei, D.M. and Lafferty, J.D. (2006). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning. ACM, pages 113–120

3. Dubossarsky, H., Hengchen, S., Tahmasebi, N. and Schlechtweg, D. (2019). Time Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

4. van Eijnatten, J. and Ros, R. (2019). The Eurocentric Fallacy. A Digital Approach to the Rise of Modernity, Civilization and Europe. International Journal for History, Culture and Modernity, 7.

5. Hengchen, S., Ros, R., and Marjanen, J. (2019). A data-driven approach to the changing vocabulary of the 'nation' in English, Dutch, Swedish and Finnish newspapers, 1750–1950. In Proceedings of the Digital Humanities (DH) conference 2019, Utrecht, The Netherlands

6. Kim, Y., Chiu, Y.I., Hanaki, K., Hegde, D. and Petrov, S. (2014). Temporal Analysis of Language through Neural Language Models. ACL 2014, p.61.

7. Koselleck, R. (2010). Vom Sinn und Unsinn der Geschichte: Aufsätze und Vorträge aus vier Jahrzehnten von Reinhart Koselleck – Suhrkamp Insel Bücher Buchdetail (C. Dutt, Ed.). Berlin: Suhrkamp.

8. Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L., & Tolonen, M. (2019a). A National Public Sphere? Analyzing the Language, Location, and Form of Newspapers in Finland, 1771–1917. Journal of European Periodical Studies, 4(1), 54–77. https://doi.org/10.21825/jeps.v4i1.10483

9. Marjanen, J., Pivovarova, L., Zosa, E. & Kurunmäki, J. (2019b). Clustering Ideological Terms in Historical Newspaper Data with Diachronic Word Embeddings. in Proceedings of the 5th International Workshop on Computational History. HistoInformatics2019 – the 5th International Workshop on Computational History, 12/09/2019.

10. Martinez-Ortiz, C., Kenter, T., Wevers, M., Huijnen, P., Verheul, J. and Van Eijnatten, J. (2016). Design and implementation of ShiCo: Visualising shifting concepts over time. In HistoInformatics 2016 (Vol. 1632, pp. 11–19).

11. National Library of Finland (2011a). The Finnish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2016050302.

12. National Library of Finland (2011b). The Swedish Sub-corpus of the Newspaper and Periodical Corpus of the National Library of Finland, Kielipankki Version [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2016050301.

13. Skinner, Q. (2002). Visions of politics. Vol. 1, Regarding method. Cambridge University Press.

14. Tolonen, M., Lahti, L., Roivainen, H., & Marjanen, J. (2019). A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 52(1), 57–78. https://doi.org/10.1080/01615440.2018.1526657

ID: 194
Poster
Topics: historical studies, digital resources – publication and discovery, infrastructure, open data, research data archiving
Keywords: research infrastructures, historical statistics data, data curation

Collecting and Storing the Historical Statistics Data on Baltic Countries in 1897ï 1939

Giedrius Žvaliauskas

Vilnius University, Lithuania

The third wave of democratization and collapse of communism in Eastern Europe in 1989–1991 opened the “windows of opportunity” to restore independent states for all three Baltic countries after 50 years of foreign Soviet and Nazi occupation. Last year all three Baltic nations celebrated 100 year anniversary of the since proclamation of state independence in 1918. This is proper occasion to take stock of their turbulent history in the long-run time perspective.

Since 2018 a team of Vilnius University researchers lead by prof. Zenonas Norkus implement the research project “Historical Sociology of Modern Restorations: A Cross-Time Comparative Study of Post-Communist Transformation in the Baltic States”. This research is funded by the European Social Fund according to the activity “Improvement of researchers’ qualification by implementing world-class R&D projects” of Measure No. 09.3.3-LMT-K-712 “Improvement of scientists', other researchers' and students' qualification by implementing practical scientific activities”.

The ultimate aim of the project is to construct and empirically apply a new sociological theory whose subject matter is social restorations in the modernising and modern societies. Baltic countries are interesting as most conspicuous cases of modern social restorations, with Latvia still living under constitution accepted in 1922. The fruitfulness of new theory is explored by the comparative study, comparing the economic, demographic, social, and political development of Baltic countries during post-communist and interwar periods. The main obstacle for

such comparison is the scarcity of cross-nationally and cross-time comparable data about the pre-WWI and interwar periods. They remain dispersed in the national statistical publications of Baltic States of interwar period or those on the pre-WWI time in the statistical publications and archives of Russian Empire.

During nearly 30 years no digital collections of historical statistical data (with partial exception for Lithuania) on the history of Baltic countries during first independence period were published. So interwar development of Baltic States still remains invisible for the broad international research community, accustomed to make research mainly using data, available in the electronic databases.

Therefore, one of the several activities of the researchers implementing research project is systematic collection data of historical statistics on the comparative development of Baltic States in 1897–1939. Most interesting historical statistical data are published in the Lithuanian Data Archive for Humanities and Social Sciences, LiDA (www.lidata.eu/en).

LiDA is a national social science and humanities research infrastructure that was developed in 2006–2008 by Policy and Public Administration Institute at Kaunas University of Technology in partnership with Vilnius University, Institute for Social Research, the Republic of Lithuania Ministry of Education and Sciences a project funded from EU Structural Funds. A second phase of the LIDA development took place between 2009 and 2012 and was also funded from EU Structural Funds. Currently, LiDA operation maintenance depends only on financial resources of the Kaunas University of Technology.

LiDA provides virtual digital infrastructure for acquisition, preservation and dissemination of digital social sciences and humanities data in Lithuania. LiDA is targeted the social sciences and humanities community, both institutional and individual researchers as well as students seeking to support their scientific and educational needs. It is also a useful collection of data for general public and governmental institutions.

LiDA have three data catalogues:

• Social survey data catalogue (QUANT) contains 314 data sets by October 2019;

• Catalogue of Data about Lithuanian political system (POLSYS) contains 217 data sets by October 2019;

• Historical statistics catalogue (HISTAT) contains 118 data sets by October 2019.

As a result of the project “Historical Sociology of Modern Restorations: A Cross-Time Comparative Study of Post-Communist Transformation in the Baltic States”, the first data sets were published in June, 2018, and 200 data sets of historical statistics data of three Baltic States will be published until the end of the project in late 2021. Circa 50 data sets published by October 2019 are divided into four thematic sections:

• Thematic collection “Economy: agriculture, forestry and fishing” contains data about land-tenure, farming manufacture, crop, harvest and fertility, number of birds and efficiency, wood area and wood industry, fishery and fishing, melioration, etc.

• Thematic collection “Population” contains data about population size, density, population size by place of residence (city/village), gender, confession, ethnic/national, social/estate cast; demographic historical data (birth rates, mortality, marriages, divorces, etc.); data about population migration; census data, etc.

• Thematic collection “Finance” contains data about central and local government revenue and expenditure, fiscal politics, financial institutions, money turnover, banks, deposits, credits, etc.

• Thematic collection “Prices” contains data about prices of goods and cost of living indexes, etc.

First data sets for new thematic section on education is in preparation.

Data archiving of historical statistics data in LiDA is based on the NESSTAR system and FEDORA repository. Data sets in the catalogues of LiDA are documented according to the Data

Documentation Initiative (DDI) metadata standard (1.2). Data description is documented bilingually, in English and Lithuanian.

NESSTAR data catalogue (containing all the archived data) can be accessed at www.lidata.eu/webview from the front page of the portal. The LiDA portal has clear terms and instructions of data use. They are available both in English and Lithuanian. All the information on how to search (see for example, http://www.lidata.eu/en/index.php?file=files/eng/data/help_search_data.html), use or analyse data online in Nesstar platform is available in English, too.

The whole IT infrastructure of LiDA was built with the basic aim to be highly interoperable. Therefore, each data set has its unique PID which is constructed to reflect the main attributes of the data set (see for example, www.lidata.eu/data/histat/LiDA_HISTDEM_0063). All the files of the data set can be accessed by registered LiDA users by following the standardized rules. For example, data in Excel format can be accessed by adding '/EXCEL.01.001' (for example www.lidata.eu/data/histat/LiDA_HISTDEM_0063/EXCEL.01.001) to the PID and DDI file – by adding '/DDI'. So external users or other infrastructures can easily access data and metadata stored in LiDA catalogues.

Historical statistics data stored in LiDA are available without restrictions for the registered users for non-profit purposes (such as research, self-education and training), except embargo period of data sets, which are published as a result of the project “Historical Sociology of Modern Restorations: A Cross-Time Comparative Study of Post-Communist Transformation in the Baltic States”. However, embargo will be lifted by the very end of the project.

Regardless of access restrictions to data files all the metadata are freely available without restrictions. LiDA IT infrastructure allows all the users around the world to access the data and metadata stored in the LiDA catalogues. The data are also documented in English, which makes the data sets potentially interesting for the international community.

TOPIC INDEX

3D modeling / virtual and augmented reality, 54, 57
anthropology, 80, 86
archaeology, 211
archiving, 20, 23, 64, 76, 78, 85, 86, 143, 205, 253, 254, 255, 291, 293
art history, 42, 50, 197, 198
artificial intelligence, 50, 83, 86, 88, 134, 165, 240, 243, 276, 281, 282
audio / video / multimedia, 54, 230
bibliographic studies, 59, 128, 164, 248
big data, 76, 80, 83, 85, 86, 120, 121, 146, 151, 153, 164, 193, 200, 232, 286
citizen humanities, 15, 120, 200
citizen science, 15, 120, 200
classical studies, 80
communication studies, 32, 86, 92, 123, 146, 164, 183, 189, 200, 235
computational science, 15, 46, 54, 59, 80, 83, 85, 92, 120, 121, 134, 140, 142, 146, 224, 281
Convolution Neural Networks, 167
copyright / licensing / Open Access, 23, 64
corpus linguistics, 18, 22, 27, 46, 55, 59, 70, 104, 120, 123, 128, 146, 156, 161, 164, 179, 201, 218, 223, 229, 232, 248, 268, 286
crowdsourcing, 98, 120, 140, 200, 226
cultural heritage collections, 15, 18, 23, 27, 32, 34, 36, 42, 55, 56, 59, 70, 76, 83, 98, 134, 140, 155, 197, 224, 227, 240, 248, 253, 276
cultural studies, 18, 27, 32, 34, 47, 59, 80, 86, 98, 115, 165, 193, 197, 205, 229
data journalism, 66
data mining / text mining, 15, 23, 55, 59, 76, 80, 115, 123, 128, 140, 146, 164, 165, 178, 179, 183, 187, 189, 206, 211, 218, 223, 224, 229, 231, 233, 243, 248, 257, 286
data modeling / knowledge representation, 64, 110, 120, 140, 200, 268
design, 16, 23, 32, 42, 46, 54, 55, 57, 58, 64, 65, 67, 80, 83, 92, 110, 111, 156, 164, 198, 199, 200, 240, 261, 274, 275, 281
digital activism, 66, 261
digital literacy, 42, 43, 77, 241
digital resources – publication and discovery, 15, 23, 32, 56, 64, 76, 80, 85, 110, 134, 146, 155, 164, 200, 205, 232, 236, 253, 291
digitisation – theory and practice, 18, 23, 32, 42, 64, 80, 134, 201, 240, 276

discourse analysis, 18, 90, 123, 183, 218, 219, 235
distant reading, 33, 196, 231
diversity and multilingual / multicultural approaches, 47, 64, 120, 211, 227
encoding – theory and practice, 70
ethnography, 29, 47, 59, 86, 197
film and media studies, 189, 226, 229, 230
folklore and oral history, 143, 227, 233
gender studies, 18, 80, 229
geography, 32, 59, 128, 197, 198, 256
geospatial analysis, 32, 197, 211
GLAM, 15, 16, 17, 23, 34, 42, 45, 76, 98, 143, 205, 276
historical studies, 23, 32, 50, 55, 56, 80, 104, 128, 140, 141, 155, 164, 178, 186, 189, 197, 205, 210, 211, 218, 248, 253, 257, 276, 286, 291
history and theory of digital humanities, 80, 141, 183
image processing, 50, 134, 165, 169, 171, 201, 276
information architectures, 32, 64
information retrieval, 32, 33, 55, 66, 187, 218
intellectual history, 178
interdisciplinary collaboration, 15, 59, 70, 80, 92, 142, 164, 211, 232, 261
interface & user experience design, 32
lexicography, 27, 54, 110, 132, 268
library & information science, 15, 23, 64, 65, 66, 76, 80, 98, 186, 193, 276
linguistics, 27, 46, 54, 55, 59, 70, 80, 104, 120, 123, 124, 127, 128, 134, 140, 146, 156, 158, 164, 200, 201, 211, 212, 214, 219, 223, 224, 248, 252, 261, 262, 266, 267, 268
linked data, 34, 83, 85, 155, 210, 248
linked data / semantic web / ontologies, 34, 38, 85, 155, 210, 248
linking and annotation, 34, 64, 92, 120, 200, 223, 230, 232, 268
literary studies, 33, 64, 65, 70, 73, 115, 119, 142, 151, 165, 179, 186, 187, 193, 197, 206, 230, 236, 244, 245, 246
medieval studies, 38, 50, 55, 110, 211, 268
museology, 42, 57, 205
named entities, 34, 35
named entity recognition, 35, 94, 97, 155, 224
natural language processing, 55, 56, 80, 81, 92, 115, 120, 140, 146, 156, 173, 178, 187, 201, 218, 224, 229, 231, 232, 257, 286
network analysis, 18, 20, 21, 22, 86, 105, 108, 153, 193, 194, 196
networks / relationships / graphs, 54, 104, 193
neural networks, 136, 137, 139

newspapers, 77, 124, 128, 130, 134, 141, 157, 192, 224, 257, 258, 278, 286, 287, 288, 289
OCR, 25, 56, 80, 82, 131, 134, 135, 136, 137, 138, 139, 141, 212, 224, 229, 276, 277, 278, 280
open access, 23, 24, 144, 145
open data, 38, 85, 134, 140, 143, 155, 164, 179, 232, 283, 291
open science, 59, 85, 164, 244
parliamentary, 122, 140, 157, 159, 218, 219, 221, 235
philosophy, 80, 206
political science, 80, 92, 140, 173, 189, 218, 243
project design / organization / management, 23, 57, 64, 92, 110, 164
religious studies, 18, 33, 80, 231
scholarly editing, 33, 64, 236, 253, 256
sentiment analysis, 72, 120, 174, 187, 200, 223, 227, 230, 231
social media, 39, 66, 86, 102, 103, 122, 147, 152, 173, 183, 184, 217, 231
sociology, 86, 128, 156, 164, 231, 240, 241, 248
software design and development, 32, 54, 80, 164, 281
standards and interoperability, 32, 64, 83, 164
stylistics and stylometry / authorship attribution, 55, 115, 186, 227
sustainability and preservation, 59, 64, 232, 261
teaching / pedagogy / curriculum design, 80
theology, 18, 20
topic modeling, 179, 182, 227, 286, 288
topic modelling, 115, 116, 117, 119, 233, 286, 288
user studies / user needs, 42, 57, 76, 80, 164, 200
visualisation, 32, 46, 47, 54, 92, 104, 108, 144, 145, 193, 210, 211, 231, 244
web research, 23, 76, 86, 205

AUTHOR INDEX

Abdulla, Aisha Al, 15
Agbozo, Ebenezer, 121
Agersnap, Anne, 18
Alén, Niklas Kristian, 23
Alstola, Tero, 104
Alumäe, Tanel, 243
Ames, Sarah, 15
Amilevicius, Darius, 281
Andronova, Everita, 27
Arthur, Paul, 32
Auziņa, Ilze, 232
Battista, Simonetta, 110, 268
Baunvig, Katrine F., 33
Bojārs, Uldis, 34
Borin, Lars, 223
Bray, Paula, 15
Brix, Antoine, 38
Burrows, Toby, 38
Camps, Martin, 39
Candela, Gustavo, 15
Carabarin, Lizeth Gonzalez, 50
Carvalho, Ricardo, 206
Chambers, Sally, 15, 76
Champion, Erik, 32
Chanjong, Im, 165
Charapan, Nadzeya, 42
Coats, Steven, 46
Cocq, Coppélie, 47
Costiner, Lisandra S., 50
Craig, Hugh, 32
Dagsson, Trausti, 54
Dahllöf, Mats, 55
Dannells, Dana, 56
Dave, Mohana, 173
Dellinger, Theresa, 253
Derven, Caleb, 15
Dignum, Virginia, 58
Dobreva, Milena, 15
Domeij, Rickard, 233
Einebrant, Erik, 57
Emery, Doug, 38
Engl, Isabella, 229
Ernštreits, Valts, 59
Evensen, Nina Marie, 64
Feola, Gabriela Di, 57
Fergencs, Tamás, 65
Fraas, Mitch, 38
Friberg von Sydow, Rikard Lars, 66
Gao, Jianbo, 187
Gasser, Katrine, 15
Gerassimenko, Olga, 70
Goldberga, Anita, 34
Grūzītis, Normunds, 232
Gu, Ning, 32
Gunnell, Terry, 74
Halbhuber, David, 230
Hämäläinen, Mika, 164
Harju, Auli, 92
Harvey, Mark, 32
Haskins, Victoria, 32
Helboe Johansen, Kirstine, 18
Hellström, Saara, 146
Hengchen, Simon, 286
Herzog, Juliane, 229
Hokkanen, Julius, 92
Holownia, Olga, 76
Houghton, Christopher Michael, 80
Hu, Qiyue, 187
Hyvönen, Eero, 38, 83, 85, 155, 210
Ikkala, Esko, 38, 210

Illés, Dominika, 65
Ivanova, Anastasia A., 86
Janicki, Maciej, 92
Jansson, Ina-Maria, 98
Jarvis, Oliver S., 33
Jauhiainen, Heidi, 104
Johannsson, Ellert Thór, 110, 268
Johnsen, Lars G., 134
Jokipii, Ilkka, 210
Jónsdóttir, Eva María, 54
Jónsson, Jón Hilmar, 54
Judisch, Lisa, 229
Juntunen, Hanne Emilia, 115
Kahusk, Neeme, 70
Kaindl, Florian, 231
Kaislaniemi, Samuli, 164
Kajava, Kaisla, 120
Kāle, Maija, 121
Kallioniemi, Noora, 226
Kanner, Antti, 92, 123, 128, 211
Karner, Stefan, 15
Karvo, Elina, 226
Kåsen, Andre, 134
Kettunen, Kimmo, 140, 224
Klaus, Barbara, 141
Kļava, Gunta, 59
Koho, Mikko, 38, 210
Koivunen, Anu, 92, 123
Kokegei, Kristy, 15
Kotkov, Denis, 142
Kozlovs, Normunds, 235
Kristensen-McLachlan, Ross Deans, 18
La Mela, Matti, 140
Laak, Marin, 70
Lagus, Krista, 164
Lahti, Leo, 164
Laime, Sandis, 143
Laine, Kimmo, 226
Laippala, Veronika, 146
Larsen, Birger, 186
Launis, Kati, 142, 151, 193
Laursen, Ditte, 15
Leskinen, Petri, 155
Levāne-Petrova, Kristīne, 156
Lewis, David, 38
Lindén, Krister, 104
Liu, Bin, 187
Loula, Angelo, 206
Mahey, Mahendra, 15
Mähler, Roger, 189
Majumder, Prasenjit, 173
Mäkelä, Eetu, 92, 123, 164
Mandl, Thomas, 165, 173
Marjanen, Jani, 128, 286
Martin, Benjamin G., 178
Martynenko, Antonina, 179
Maurer, Liina, 183
May, Andrew, 32
Meier, Florian, 65, 186
Modha, Sandip, 173
Myking, Synnøve, 38
Neovius, Mats, 142
Nevalainen, Terttu, 164
Nielbo, Kristoffer Laigaard, 18, 33, 187
Norén, Fredrik, 189
Nurmi, Olli, 193
O'Dochartaigh, Eavan Fiona, 197
Öhman, Emily, 120, 200
Page, Kevin, 38
Pascoe, Bill, 32
Persson, Simon, 56
Petzell, Erik M., 201
Piao, Hui, 120
Pilawka, Olga, 65
Piper, Alana, 32
Pivovarova, Lidia, 286
Potter, Abigail, 15

Povroznik, Nadezhda, 205
Pyysalo, Sampo, 146
Queiroz, Joao, 206
Ransom, Lynn, 38
Rantala, Heikki, 210
Rašmane, Anita, 34
Raunamaa, Jaakko, 211
Repo, Liina, 146
Roivainen, Hege, 128
Römpötti, Tommi, 226
Rönnqvist, Samuel, 146
Roozenbeek, Jon, 217
Ros, Ruben, 218
Rouces, Jacobo, 223
Ruokolainen, Teemu, 224
Ryan, Lyndall, 32
Sahala, Aleksi, 104
Säily, Tanja, 164
Salmela, Anna, 146
Salmi, Hannu, 226
Sarv, Mari, 227
Schjødt, Uffe, 18
Schmidt, Thomas, 229, 230, 231
Sebastian, Schmideler, 165
Seuri, Olli, 92
Sevänen, Erkki, 151, 193
Skadiņa, Inguna, 232
Skantsi, Valtteri, 146
Skeppstedt, Maria, 233
Šķilters, Jurģis, 234
Sköld, Olle, 98
Skott, Fredrik, 233
Skulte, Ilva, 235
Slavinska, Magdalena, 236
Smith, Rosalind, 32
Sonnberger, Jakob, 253
Spuriņa, Maija, 240
Stjernfelt, Frederik, 186
Straube, Armin, 15
Strode, Ieva, 241
Svärd, Saana, 104
Tahko, Tuuli, 128
Tahmasebi, Nina, 223
Talts, Siim, 243
Tereshchenko, Kseniia Alexandrovna, 244
Thomsen, Mads Rosendahl, 187
Thomson, Emma, 38
Tiedemann, Jörg, 120
Tinits, Peeter, 248
Tolonen, Mikko, 164, 286
Trella, Fredrik, 57
Tuominen, Jouni, 38, 210
Verhoeven, Deb, 32
Veskis, Kaarel, 70
Vider, Kadri, 70
Vogeler, Georg, 253
Wagner, Sophie-Carolin, 15
Wareham, Andrew, 253
Wessels, Leon, 257
Wiebke, Helm, 165
Wijsman, Hanno, 38
Wilbur, Joshua, 261
Willcox, Pip, 38
Wills, Tarrin, 110, 268
Wilms, Lotte, 15
Wolff, Christian, 231
Žabička, Petr, 276
Zalkalns, Lilita, 274
Zavřelová, Alžbeta, 276
Ziziunas, Tadas, 281
Znotiņš, Arturs, 232
Zosa, Elaine, 286
Žvaliauskas, Giedrius, 291