elte.dh

DH_BUDAPEST_2019

Centre for Digital Humanities – Eötvös Loránd University

GALE
A Cengage Company

cost
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

MINISTRY FOR
INNOVATION AND TECHNOLOGY

SPRINGER NATURE

qulto
for culture & knowledge

CONTENT

POSTERS 71

INDIVIDUAL PRESENTATIONS

Chris Houghton

Head of Digital Scholarship – International, Gale Primary Sources

email:
chris.houghton@cengage.com

Gale Digital Scholar Lab – distant reading 160 Million pages of digital archives.

As the leading global publisher of digital archives, Gale has been at the forefront of making digitised primary sources available to researchers for distant reading. The release of Gale Digital Scholar Lab has vastly increased the potential for analysing large corpora of unique or hard to access historical documents – now, anyone can text mine Gale archives, irrespective of technological infrastructure or coding skill.

Biography

As Head of Digital Scholarship, Chris Houghton leads Gale's efforts to support the diverse community of digital humanities internationally. Frequently asked to speak at conferences and academic events around the world, Chris is passionate about expanding the possibilities of Gale archives for all researchers, irrespective of nationality or level.

Zsolt Almási

Péter Pázmány Catholic University,Department of English Literatures and Cultures, Budapest, Hungary

email: almasi.zsolt@btk.ppke.hu

Data, Machine Reading and Literary Studies

Keywords: machine reading / data / Python / statistical analysis / word count.

Our use of language influences our method of dealing with the world. When one describes the route to knowledge, one normally starts with data, then out of the data information is created, out of the information we may arrive at knowledge. This hermeneutic gradual progress includes selection, contextualization and transformation, and it all starts with data. It is data that we deploy when reading distantly when the machine is made to “read” and analyse large corpora of texts. It is data that we gather, it is data we build our reasoning on, it is data that might reorient our research and making sense of the world around us. This foundational role of data is something that can hardly be denied. But what is data? Is it something given as the Latin origin of the word would dictate? Is it a givenness that is there for everybody in a sterile and objective manner? Is it not our language, the word choice (“gathering data, building reasoning on them”) that misleads our seeing data as they are? In this paper, I would like to argue that data are not something given, but that data is actually humanly constructed and also that we need a language, a conceptual framework in which the term can be used beneficially and constructively. This clarification and conceptualization seem indispensable, because the term is used more often in literary analyses, and is received with hesitation, occasionally with repulsion by the community. The hesitation most of the time results from the close association of the term with the STEM disciplines, i.e. the methods of natural sciences with which very few literary scholars would like to identify literary research with. I will analyse a short python script to show how algorithmically we can think about data, and show how data is constructed during the process of the analysis. Keeping the conclusions of these analyses in mind I will attempt to delineate a concept of data.

Silvie Cinková – Jan Rybicki

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Czechia
Jagiellonian University, Institute of English Studies, Poland

email:
cinkova@ufal.mff.cuni.cz
jan.rybicki@uj.edu.pl

Stylometry in Literary Translation via Universal Dependencies: Finally Breaking the Language Barrier?

Keywords: stylometry / Universal Dependencies / multilinguality / authorship attribution / translation / parsing / NLP

The method of stylometry by most frequent words does not allow direct comparison of original texts and their translations, i.e. across languages. We have tried to remove this language barrier for direct stylometric comparison of literary translation by replacing the text tokens with the cross-lingual morphological tags provided by the Universal Dependencies scheme (http://universaldependencies.org). We have prepared a collection of approximately 50 texts in Czech and German: it contained Czech originals and their German translations; German originals and their Czech translations; and works originally written in English, French or Russian, but only in Czech and German translations. We converted the texts into sequences of tags using the UD-Pipe parser. The output was a CONLL-u table (one token per line, different types of linguistic markup in columns: token, lemma, coarse part-of-speech tag, fine-grained morphological features, dependency relation to the governing node, ID of the governing node).

The coarse part-of-speech (POS) tags are called UPOS and the fine-grained morphological features are called FEATS. The dependency relations are called DEPREL. While the UPOS are truly universal, the inventory of relevant FEATS is language-dependent. For instance, a language that does not use noun definiteness does not use the \`Definite\` feature with its values \`Def\` and \`Indef\`. Since the FEATS inventories of Czech and German differ, we have stripped feature attributes and values specific to either language, keeping only features and values shared by both languages.

As a next step, we replaced the original lemmas with a ``FLEMMA", i.e. a "fake lemma". FLEMMA is a cross-lingual lemma for the German-Czech pair. We designed it as a numerical ID. We obtained a glossary of FLEMMAS drawing on a Czech-German glossary (Škrabal and Vavřín 2017) that was automatically generated from the multilingal parallel corpus InterCorp (Rosen 2016). We counted various combinations of tags obtained for the four parsing levels according the the usual Delta procedure (Burrows 2002). Each tagging string was treated as a single element (the counterpart of word-types in traditional stylometric analysis by most frequent words), and their frequencies were counted in the texts in the corpus and compared in text pairs to produce a matrix of distance measures; in this study, the distances were established by means of the modified Cosine Delta (Smith and Aldridge 2011), which is now seen as the most reliable version (Evert et al. 2017). The function classify() in the stylo package (Eder et al. 2016) for R (R Core Team 2016) was used to try to assess authorship attribution success when its reference set contained texts in one language and the test set contained texts in the other. Attribution success was counted whenever the Delta value for the pair of the translations of the same text was lowest.

The most successful combination for authorship attribution was FLEMMAS + UPOS (95.6%), while FLEMMAS alone achieved only (3.7%), and so did UPOS alone. This shows that even a very crude word-to-word translation (polysemy neglected), along with the coarse part of speech, helps bypass the language barrier. Another interesting finding is the seemingly modest 20.3% success rate for attribution by the combination UPOS + FEATS + DEPREL. This is, after all, much more than a coin toss for such a number of texts. More importantly, guessing was consistently successful for the same pairs of texts, suggesting that these particular pairs might be easier to guess. This, in turn, might indicate that the translations in these case applied strategies resulting in grammatical and syntactic structures somehow similar to the original, a phenomenon often observed in word-for-word translation, or, more generally, in translations aiming for the so-called “formal equivalence” (Nida 1964).

This work was funded by LTC18020.

Silvie Cinková – Tomaž Erjavec – Cláudia Freitas – Ioana Galleron – Péter Horváth – Christian-Emil Ore – Pavel Smrž – Balázs Indig

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Czechia
Department of Knowledge Technologies, Jožef Stefan Institute, Slovenia
Pontifícia Universidade Católica do Rio de Janeiro, Departamento de letras, Brazil
Université de Sorbonne Nouvelle - Paris 3, Littérature et Linguistique françaises et latines, France
Eötvös Loránd University, Hungary
University of Oslo, Norway
Brno University of Technology, Czechia
Eötvös Loránd University, Hungary

email:
cinkova@ufal.mff.cuni.cz
tomaz.erjavec@ijs.si
claudiafreitas@puc-rio.br
ioana.galleron@sorbonne-nouvelle.fr
horvathpeti99@gmail.com
c.e.s.ore@iln.uio.no
smrz@fit.vutbr.cz
indig.balazs@btk.elte.hu

Evaluation of taggers for 19th-century fiction

Keywords: NLP / morphological tagging / UD-Pipe / Universal Dependencies / multilinguality / language resources

We evaluate a selection of currently available taggers on fiction texts from the 19th century. The texts belong to the ELTeC collection created and maintained by the WG1 team of the COST project CA16204 "Distant Reading for European Literary History". It is a multilingual collection of 19th-century fiction in the following languages: Czech, English, French, German, Hungarian, Italian, Norwegian (both varieties), Portuguese, Serbian, and Slovene (with new languages being added). Each language is represented by at least 20 novels in the TEI format.

We have evaluated the taggers for the following languages:

Czech, English, French, German, Hungarian, Norwegian-Nynorsk, Portuguese, and Slovene. We have tagged them all with the UD-Pipe tagger. On top of that, Hungarian was also tagged with e-magyar. We have selected UD-Pipe because of the appealing cross-lingual markup and uniform model building for all languages. It was known before that the Hungarian model is very small and that the Hungarian tagging is poor. Therefore we have also evaluated e-magyar, the best tagger currently available for Hungarian.

The annotation scheme of UD-Pipe

The UD-Pipe tagger and parser draws on Universal Dependencies (universaldependencies.org), henceforth UD. UD is a cross-lingual annotation scheme for morphology and syntax. At the moment, more than 70 languages have their own UD treebank to train a parser on. The categories that UD-Pipe annotates are: lemma, coarse part of speech (e.g. VERB), universal features (e.g. Plural, Past Tense), and syntactic dependency relations (e.g. Subject, Predicate).

The procedure

We have selected a random sample for a manual tagging correction in each language. Each sample comprised approximately 5,000 tokens in entire sentences. The sentences had been extracted from ELTeC across all documents in each language and tagged in the CONLL-u format: one token per line, columns: token, lemma, universal POS, universal features. The annotators performed the annotation in a spreadsheet, in four additional columns for tokenization, lemma, POS, and features, respectively. In the first three columns, they were to indicate the corresponding error type(s) with an "F". In the fourth column, the annotators were to type the number of incorrectly recognized linguistic features.

Hence, the annotation captures four different error types for each token. The number of feature errors measures only the precision, not the recall, of the tagger. This is because the morphological annotation is rather complex and it would have been too time- and resource-consuming to train a fully competent morphological annotator for each language, who would have been able to correct the automatical tagger output to a new gold standard.

We have produced descriptive and explorative statistics and visualizations for each language to compare the performance of UD-Pipe among languages and to analyze the errors. As a second step, we have compared the vocabulary of the individual samples to the vocabulary of the referential treebanks of the respective languages to find out how much the domain of the 19th-century fiction differs from the domain of the referential treebank. We have manually classified the tokens specific to the 19-century samples into the following categories: Foreign word/heavy dialect, archaic spelling/morphology, morphological ambiguity, archaic tokenization, typo, proper noun. The annotator was allowed to leave a cell blank when none of the labels matched. The category labels could also be deliberately combined.

Based on both annotations, we are able to tell, for each language, where the tagger has most problems (POS, lemmatization? Tokenization? Features?) and with which words (how much is archaic spelling associated with lemmatization performance?). The ability of the tagger to "guess" unknown words differs in each language. With this double annotation, we are able to start a collaboration with UD-Pipe developers on a possible domain adaptation for selected languages and build reliable models to tag 19-th century fiction.

This work was funded by LTC18020.

Marek Debnár

Faculty of Arts, Constantin the Philosopher University in Nitra, Slovakia

e-mail:
mdebnar@ukf.sk

Quantitative Research of Essays in Slovakia: Past and Present

Keywords: digital humanities / quantitative formalism / essayistic genre / quantitative formalism / distant reading

The paper compares the methods of quantitative corpus analysis applied to essays in Slovak computational linguistics in the past and the present. First, following the example of compiling the first electronic “Frequency Dictionary of Slovak Language” by J. Mistrík (ed.) (published in 1962) and mainly based on the study: “Dictionary of Essayistic Style” from 1972, we clarify the quantitative research of language in Slovakia. The article also presents the methodology of compiling the dictionary (notion ‘corpus’ was not used in Slovak linguistics back then) and results the author achieved in the quantitative analysis of essays.

In the second part, we focus on the process of creating and marking up/annotating Corpus of essays in the Slovak National Corpus created in 2016. The comparison of these methodologies and construction of these corpora, shows that the current technological word processing in the field of linguistics gradually brings us closer to the interdisciplinary field of distant reading.

Amelie Dorn – Barbara Piringer – Yalemisew Abgaz – Jose Luis Preza Diaz – Eveline Wandl-Vogt

Austrian Centre for Digital Humanities (ACDH-OeAW), Austrian Academy of Sciences, Austria
Adapt Centre, Dublin City University

email:
amelie.dorn@oeaw.ac.at
barbara.piringer@oeaw.ac.at
yalemisew.abgaz@adaptcentre.ie
joseluis.prezadiaz@oeaw.ac.at
eveline.wandl-vogt@oeaw.ac.at

Enrichment of Legacy Language Data: Linking Lexical Concepts in Data Collection Questionnaires on the example of exploreAT!

Keywords: legacy language data / concepts / DBpedia / Linked Open Data / wikification

The project exploreAT! (Wandl-Vogt et al, 2015) runs as a module within exploration space @ ACDH-OeAW, an Open Innovation Research Infrastructure (OI-RI) for the Humanities.

It aims at revealing and making accessible cultural knowledge contained in a non-standard German language resource (DBÖ [Database of Bavarian Dialects in Austria]), exploiting it from the perspectives of semantic technologies, visual analysis tools and cultural lexicography. Cultural information captured in legacy language collections often remains unaccessed, or requires specific data processing efforts. The DBÖ collection is a large and rich heterogeneous resource (~3.5 mio. entries) from the time of the former Austro-Hungarian monarchy, comprising of digitized data collection questionnaires, answers and excerpts from vernacular dictionaries and folklore literature. With 109 systematic data collection questionnaires, linguistic but also a wealth of cultural information (customs, festivities, food, etc) was captured. By opening up the questionnaires through lexical concepts and their linking to other resources, we aim to enable and foster to understand the “rural world” in the early 20th century as captured in the data (cf. Arbeitsplan, 1912).

In this paper, we describe and demonstrate our approach to linking individual questionnaire topics/lexical concepts to DBpedia concepts and we outline the processes and challenges.

First, individual questionnaire topics, originally in German, were automatically extracted from the questionnaire title which defines the topic of a questionnaire, e.g.

Farben (DE)/ colours (EN), and all questions contained therein. Next, titles were translated into their English equivalent or nearest best fit. To link the questionnaire titles/topics to external resources, e.g. DBPedia, dbpedia spotlight service was employed. For each questionnaire, English DBPedia concepts were identified with a certain degree of confidence, and the corresponding concept generated. Then, experts evaluated the accuracy of the results in a csv file by comparing the DBpedia concepts definitions with the topic of the questionnaire. This knowledge was tapped from experienced experts familiar with the detailed contents of the questionnaires and questions. Where necessary, additional DBpedia concepts were added manually. Once an agreement was reached, these concepts are added permanently to the database, and used as an authoritative source to link the questionnaires with DBpedia concepts.

A number of challenges arose in this process: linking topics to matching German DBpedia concepts was not always straightforward. At times, concepts were only available in English, and sometimes there was no equivalent in DBpedia at all. In these cases, experts opted for the nearest fitting equivalent and noted this uncertainty in the evaluation file. Further, nuanced differences in meaning between German and English words and their corresponding DBpedia concepts (e.g. German: Schuster; English: shoemaker; cordwainer) also proved challenging. As a result, levels of detail in assigned concepts across questionnaires vary.

Next steps involve the improvement of the concepts assignment as well as adding further concepts to provide a more detailed picture of the conceptualisation of these legacy language data questionnaires, including visualisations of relations and networks (e.g. Apache Superset).

Ghazal Faraj – András Micsik

Eötvös Loránd University, Budapest, Hungary
Hungarian Academy of Sciences Institute for Computer Science and Control, Budapest, Hungary

email:
ghazalgf@hotmail.com
micsik@sztaki.hu

Enriching and linking Wikidata and COURAGE registry

Keywords: Linked Data / Cultural Heritage / Wikidata / Individual Discovery / Link Discovery

The COURAGE (Cultural Opposition: Understanding the CultuRal HeritAGE of Dissent in the Former Socialist Countries) project explored the methods for cultural opposition in the socialist era (cc. 1950-1990). It has a database of historic collections, people, groups, events and sample collection items stored in an RDF triple store. The registry is browseable online, and has been used to create virtual and real exhibitions, learning material, and it is also planned to serve as a basis for further narratives and digital humanities (DH) research.

Wikidata is the main storage for structured data which is related to Wikipedia, Wikisource, and others. This data can be edited by volunteers worldwide. The most important benefit of using Wikidata is linking datasets using relationships that can be understandable by humans and machines. Currently, Wikidata is one of the best repositories for Linked Data and it had a campaign for collecting cultural heritage data. In order to provide better connectedness between the two datasets, we tried to link people, organizations and groups existing in Wikidata with their representatives in COURAGE.

The production of Linked Data is growing up rapidly these days, but the linking between these datasets is weak. Different entities describe the same individual in different datasets, similar statements are described differently in different datasets**.** This is a huge objective for users of any linked dataset, as they access a subset of statements on individuals, so they have to work with a partial view of known facts.

As a first step, we tried to find the individuals present in both Wikidata and COURAGE datasets. Although persons in COURAGE dataset are described with historical preciseness, there were persons who did not allow to publish their birth year (and thus age) and birthplace. On the other side, persons and groups in Wikidata often have very short and void descriptions**.**

There was no human capacity for research of matching individuals one-by-one, and a fully automatic matching also proved to be unfeasible. Therefore, we aimed at detecting the cases where human decision is needed and also at minimising their number.

We present our experiments with various approaches for automated matching of COURAGE entities with Wikidata entities. A classification is introduced for matched pairs: correct match, incorrect match, undecidable match and cases where an expert decision is needed. Statistics and examples are provided for each class of matches.

Finally, the result of statistics showed that more than 78% of people, organization and group data can be safely matched automatically. As a next step, a list of transferable properties has been collected and triples to extend Wikidata have been compiled. As a result, both Wikidata and COURAGE datasets will be enriched with cultural heritage data which, in turn, allows more insights to be obtained by researchers.

Maryam Foradi – Johannes Pein – Jan Kaßel

University of Leipzig, Germany

email:
maryam.foradi@uni-leipzig.de
johannes.pein@studserv.uni-leipzig.de
jan.kassel@studserv.uni-leipzig.de

Phonetic Transcription of Classical Literature: A Learnersourcing-Based Approach

Keywords: smarttexts / learnersourcing / computer assisted language learning (CALL) / phonetic transcription / corpus building

Distant reading for high-resource languages like English, German, and French is well-researched both in the domain of philology and the digital humanities. Further research is facilitated by a number of software implementations and a wide array of bilingual corpora. However, under-researched languages like Persian are held back by a lack software solutions and bilingual corpora alike.

Bilingual corpora allow interested readers without knowledge of the source language to unrestrictedly access texts, compared to the limited number of native readers and language learners who are able to proficiently interact with source texts. The lack of open-access software solutions leads to challenges in making Persian corpora machine-actionable. Hence, the problem of accessibility arises for the scholars and readers with no sufficient knowledge in that language. A possible solution that offers enhancements to text readability for the non-native speakers of Persian is the phonetic transcription of these corpora. Furthermore, the automatic annotation of phonetic transcribed corpora requires less effort, as rule-based algorithms can be used. In this context, phonetic transcription simplifies and accelerates other types of annotation like, e.g., lemmatization, morpho-syntactical annotation.

To research the possibilities of disambiguation and correction of transcribed Persian corpora, we crafted a tool which is subject to user studies and aims to provide twofold results: First, study participants acting as learnersourcers may gain further knowledge in Persian, while second, participants' submissions are incorporated into a data-driven workflow to improve existing computational methods for acting on Persian language.

Although there are some studies available focusing on algorithmic transcription of Persian, this study with the focus on learnersourcing adds a new mode of annotation, i.e., audio annotations, with proven efficiency on language learning, to the extensively annotated corpus of poetry by the Persian poet Hafiz. The tasks proposed to the learnersourcers have three different difficulty levels, which enables us to estimate users' ability in order to rate their annotation in further stages of the project, where no ground truth is available. The results of the pre-experimental study indicate that annotators with no knowledge of Persian are able to add annotations to the Persian source with 67.8% accuracy.

This experiment shall also be combined with a second experiment, in which the performance of "regular" language learners on solving the defined task will be compared with the performance of the users with no background knowledge in that particular language in terms of data accuracy. The experiment still requires an appropriate scaffolding, consisting of a tutorial in combination with the protocol provided for familiarizing participants with the structure of the study. In this regard, we believe that familiarity of users with the International Phonetic Alphabet (IPA) and also adding some out-of-test exercises (e.g., a tutorial) will enable the users to understand the task concretely, so that the results are not negatively affected by the deficit of conceptual formulation of the task.

Arjun Ghosh

Indian Institute of Technology Delhi, India

email: arjunghosh@hss.iitd.ac.in

Ideological battles on Wikipedia: A computational study of the linkages between academic and popular histories

Keywords: wikipedia / ngrams / history / hindutva / computation

Wikipedia is an established first source for information for internet users across the globe. In the global south, where internet penetration is fast growing, the Wikipedia has the power to shape public discourse. The Wikipedia is an important platform for recording and redrafting public memory.

In India a sharp ideological polarisation between the Hindu Right and those who stand for a secular polity has been growing. The Hindu Right attempts to consolidate its hold over the public discourse by excluding, including and appropriating historical figures through rewriting of history text books, producing movies and other popular texts and inventing new practices of community memory (Thapar 2009, Mahajan 2018). These practices have also been complemented with banning books and films and attacking writers and artists who question a desired view of history.

In a post-truth world the digital media has played and increasingly important role in shaping the trajectory of the public discourse on each instance of conflict. More often than not these conflicts in the virtual world are subject to inventions by organised groups of volunteers and heavy instances of trolling. I shall use a combination of distant reading and close reading to unpack the role played by Wikipedia in this battle of ideas.

Using established practices of studying edit wars on wikipedia (Yasseri et. al. 2014) I analyse the edit activity on a set of Wikipedia pages corresponding to the most contested histories and biographies in Indian history – a set which I derive from established scholarship on the writing of history. While Yasseri et al. concentrate primarily on establishing the levels and patterns of dispute in Wikipedia edit histories, I use N-gram analysis to identify specific issues of contention within each page. In fact, in most cases these disputes are long standing. They existed hitherto either as disputes between different ideological groups or between academic research and the popular imagination. I study these specific disputes in the light of established scholarship around each of these issues. I also compare the treatment of disputed issues on Wikipedia with those in academic scholarship, school textbooks, propaganda material and popular forms like movies and graphic novels. My study reveals that the non-proprietorial and open structure of the Wikipedia enables academic research converse with popular discourse.

Further, I undertake a network analysis of the editors of these pages to reveal that while there exists a stronger network correspondence among the Wikipedia editors arguing from the position of the Hindu Right, the Wikipedia editors opposed to them are more loosely structured. The organised structures of Rightwing editors are similar to those found to be working in social media platforms. Very often instances of ideological confrontation on social media leads to heavy trolling (Chakrabarti et al. 2018; Chaturvedi 2016). However, the presence of Wiki-ettiquettes and an open administrative structure keeps Wikipedia largely a troll-free space.

My paper also demonstrates that a combination of distant and close reading – each guiding the other – ensures a productive conversation between methods of cultural and literary analysis and computational methods.

Gregory H Gilles

King's College London, United Kingdom

email: gilles.greg@gmail.com

Female Agency in the Late Roman Republican: A Social Network Approach

Keywords: Female agency / Social networking / Roman Republic

In this paper I employ social network analysis to study female agency in the late Roman republican period. My project uses female centred networks to connect women, and men, during this period as visualisation enables an easier identification of different patterns of connectedness, whether they be social, familial and/or political. With the use of the Digital Prosopography of the Roman Republic created at KCL (http://romanrepublic.ac.uk/), as well as various ancient sources, such as Cicero's letters, Plutarch's biographies and the histories of Livy, Polybius, Appian, Suetonius and Cassius Dio, I have created familial data sets so as to identify connections within four generations (one above and two below) of the various central female nodes. Through trial and error, focussing on four generations not only enables the identification of possible repeated familial connections, but also pin points new connections forged with powerful men or families in subsequent generations. There are, in total, 12 different female centred networks which include over 150 elite and equestrian women from Rome with dates ranging from c.250 B.C. to c.10 B.C.

The numerous female centred networks provide data to help answer four key questions: Were marriages mainly used to cement, or initiate, political alliances between powerful men and/or families? Findings indicate that certain elite families appeared to remarry into each other every second or third generation. Was the, often, great age disparity between spouses intentional and the norm, or was it simply due to the military and/or political careers that Roman men had to undertake before they could marry? The networks cannot definitively answer this, but analysis of first marriages correlate with findings by Richard

Saller (Classical Philology 82 [1987] 21-34) and would indicate that men’s public careers took precedence over marrying at a young age. Was a rich widow or divorcée an attraction for politically aspiring new man/impoverished noblemen? The majority of networks demonstrate that this is not the case. Did stepmothers play an active role in the upbringing of their husband’s other children? The networks highlight that most stepmothers were of similar ages to their new stepchildren and so an active role would often not have been required.

This paper, therefore, showcases the networks that have been created from the analysis of literary materials and demonstrates how social networks can be used to answer these, or similar, historical questions. The issues with the data, and their impact on the creation of these networks, as well as their analyses, will also be discussed.

Andrej Gogora

Constantine the Philosopher University in Nitra, Slovakia

email: gogora@gmail.com

The issue of the term 'digital humanities': translation and language diversity

Keywords: digital humanities / terminology / translation strategies / language diversity

In general, the paper focuses on the problem of language translation of the english term 'digital humanities'. In practice, we encounter the fact that in many non-English languages the term 'digital humanities' is commonly used in the original English version or the abbreviation DH. In a number of countries, there is a debate on how to translate (and whether to translate or not) this new-coined and problematic term. In this case, two factors are colliding here: the need to build relevant domestic terminology, and the difficulty of translating the term 'digital humanities'. The aim is to map how the term 'digital humanities' is translated into selected non-English languages, while also highlighting the related translatological issues and discussions within each language. We will focus on three major European languages - German, French, Italian - and then briefly mention the translations of other languages, including Central European ones. Moreover, we will present and summarize the main arguments used to support translatological strategies for this problematic term, which may help in choosing the appropriate translation equivalent. Finally, we will explain how this seemingly marginal problem is approached from the position of cultural criticism, i.e. from the perspective of cultural and linguistic diversity.

Andrea Gogova

FMC TBU Zlin, multimedia and design, Slovakia

email:
andgogo@nextra.sk

From Grid to Rhizome: a Rethinking of a Layout arrangement of the Post-digital Text

Keywords: grid / post-digital text layout / process / algorithm / Rhizome

In the nineteen-seventies, typographers started to think about new categories of digital typefaces. Approximately half a century later digital designers were able to deploy parametric and generative fonts. These algorithmic fonts can be shaped variedly with sets of attributes defined by code independently of the typographic design software. Although typographers are now able to solve a problem of differing screen sizes and resolutions by responsive and adaptive layout, often these fonts, whilst based on a principle of dynamic changes, remain closed into a mainly static grid arrangement of layout. In the light of the emergence of digital technologies and intellectual approaches to postdigital media, the grid arrangement of a post-digital text layout as a residuum of Modern thinking. The cultural pattern made by grid which is applied in layout of a text could be changed by actual nature of the process.

A layout based on rhizome , is increasingly recognised as a model of how to transform the view of fixed relations of a closed system. In typography, the pattern represented and reinforced by the grid might, if such a rhizomatic logic were to be deployed, .

new post-digital text around the opportunities afforded the posthuman self-organised system.

Emese Ilyefalvi

MTA-ELTE Lendület Historical Folcloristics Research Group, Hungary

email: mseilyefalvi@gmail.com

Distant reading of Hungarian verbal charms

Keywords: distant reading / Voyant Tools / computational folkloristics / digitized folklore texts / verbal charms / digital database

Based on Éva Pócs manual charm index an online database was created for Hungarian verbal charms within the East-West Research Group at the Institute of Ethnology, Hungarian Academy of Sciences (Budapest), between 2013 and 2018. (See: http://eastwest.btk.mta.hu/.) The main goal was to create a multidimensional digital database.

Digital text preparation would open the gates to new interpretations and analyses, which would bring us closer to understanding the compound and complex phenomena of charms. In the digital database of Hungarian verbal charms users can search by various metadata, like date and place of collection/recording, name of collector/scribe, informant, type of source, function of the charm, rites/gestures, language of the text, keywords etc. Free word search is also possible within the entire corpus. (See: http://raolvasasok.boszorkanykorok.hu/)

The trial version of the database contains 1712 texts, but more than 6000 verbal charms were digitized during the project. In my paper, I will analyze the entire corpus with the help of Voyant Tools, which is web-based text reading and analysis environment for digital texts, developed by Geoffrey Rockwell and Stéfan Sinclair. (See: https://voyant-tools.org/). Using Voyant Tools, I will show how different new arrangements and distant reading of the corpus can reshape our knowledge about the Hungarian verbal charms.

Margit Kiss

Hungarian Academy of Sciences, Hungary

email:
kiss.margit@btk.mta.hu

The potentials of stylometry for analyzing Hungarian historical texts

Keywords: stylometry / Hungarian historical texts / authorship attribution / style classification

For the analysis of literary texts, stylometry have played an important role in solving authorship problems for a long time. It can also help the various interpretative and analytical tasks, e.g. periodization of authors' works, comparison of different periods of an oeuvre etc. Computational analysis of Hungarian historical texts has many difficulties, e.g. it is characterized by inconsistent morphology, orthography, punctuation, foreign language inserts etc. I investigated how stylometry analysis could cope with all these problems. For this study I analyzed the works of Mihály Csokonai Vitéz (1773-1805) (123 letters) and Kelemen Mikes (1690-1761) (a large corpus of 1.5 million words). My aim was to investigate whether stylometry is adequate for analyzing these historical texts, and to determine what kind of measure distances, style markers, multidimensional methods are adequate for authorship attribution and style classification.

For authorship attribution I analyzed the debated letters of Csokonai. After a few iterations Eder’s Simple Distance was the most adequate for this purpose. The cluster analysis confirmed that the debated letters are written by the same person. Another group of experiments was focused on the works of Mikes. First I tested whether the stylometry is capable differentiate the own works and the translations. The Eder’s Delta was the most suitable for this purpose. Then I investigated the genre-thematical classification of the oeuvre. The stylometry analysis was able to clearly distinguish the spoken language texts from the religious and moralistic ones. During this the Canberra Distance gave the best results.

Based on these experiments, it is clear that stylometry is capable to handle Hungarian historical texts, however certain difficulties need to be managed. These texts contain many various morphological and orthographic forms of the same word. The frequency-based methods used in stylometry treat these as if they were separate words. The results can be improved if we could eliminate these differences. One possible way to achieve this is the normalization of historical word forms, e.g. replacing “e kepen”, “e képen”, “ekepen”, “eképen” with “ekképpen”. For the normalization task using an author’s dictionary could be an optimal solution because every word occurrences is attached to its modern equivalent. For the Mikes corpus I performed this normalization using the Mikes Dictionary. After the normalization the clustering results have been improved further.

From these experiments it is clear that a properly configured stylometry method can provide effective help in analyzing Hungarian historical texts. It is an especially valuable tool in cases when the corpus is too large to manually analyze the authorial language and style.

Tamás Kiss

Central European University, Hungary

email:
tamas.mail@gmail.com

Identifying Authors, Themes, and Authorial Intent Computationally in Early Modern Ottoman Chronicles

Keywords: Rumi 1.0 / stylometry / topic modeling / text analysis / text mining / Ottoman / Early Modern / Selaniki / Tevarih-i Ali Osman

At last year's DH_ELTE conference I presented my software named Rumi 1.0, the only software to date optimized to computationally analyze Ottoman Turkish documents. Since then, besides fine-tuning the software, I have been using Rumi 1.0 to explore the feasibility of three quantitative approaches to Osmanli narrative sources (i.e. fifteenth- and sixteenth-century chronicles such as the various versions of *Tevarih-i ʻAl-i ʻOsman* and *Tarih-i Selaniki*), namely (1) authorship attribution with distance measurement methods, (2) theme identification within and across sources, and (3) word distribution patterns across narrative time. In my paper at DH_ELTE_2019, I seek to present my research results and offer my text analytical methods to debate.

Authorship attribution: Well-tried statistical methods (Burrow's, Manhattan, Euclidean, Eder's, etc. delta or distance) are commonly used to measure the distance between texts with high success rates – although almost exclusively in modern western languages. However, there is not a method that is universally applicable to all languages: The combination of effective analytical parameters and applicable delta analyses are language specific, and as they have never been tested on the Ottoman Turkish language, so far we have not had a blueprint for authorship attribution in the Ottoman context. In my paper I will present my research results, which indisputably point to Eder's Delta and a specific combination of analytical parameters that work best for the Osmanli language.

Theme identification: Drawing on two basic assumptions of topic modelling, namely that (1) topics constitute texts, and (2) specific lexicons define specific topics, I have been experimenting with "projecting" relative word frequencies of (relatively) self-contained themes within Ottoman chronicles (such as the foundation legends of Constantinople and stories about the building of the Hagia Sophia or Ayasofya) on other chronicles to explore whether the same themes are featured in them. While looking for individual terms in search for specific topics in large corpora or applying topic modelling methods such as Latent Dirichlet Allocation clearly will not yield the required effect, my approach to identify the presence and location of topics within a corpus produces promising results. As my theme identification technique is not language specific, after discussing its benefits and limitations, in the Q&A I would find the educated comments of fellow text mining experts useful to be able to develop this method further.

Word distribution patterns across narrative time: Inspired by similar projects pursued by Ben Schmidt, Matt Jockers, and David McClure, I have been working to find the answer to a theoretical question of mine regarding authorial intent in the context of Ottoman chronicles. Namely, I have been interested in exploring whether there is an overarching narrative (a meta narrative of some sort) holding an Ottoman chronicle "together," or whether a chronicle is only a collection of several smaller-scale narratives each perpetuating an individual historical event or a notable political or social phenomenon. The obvious answer is “Yes, there always is an underlying narrative at play,” but as it is not clear how this meta narrative – or the mere existence thereof – could be discerned with the help of quantitative methods, so far Ottomanists have needed to rely on their critical skills to guess the authorial intent that encompasses the individual narratives contained in the sources’ chapters and sub-chapters. My approach to explore the distribution of certain words across a text’s narrative time reveals tendencies that may bring us closer to reveal such meta narratives. As my experiment is not specific to the Osmanli language, here, again, I would be grateful for any comments coming from fellow researchers in the Q&A.

Jiří Kocián – Jakub Mlynář – Pavel Obdržálek

Charles University, Czechia

email:
kocian@ufal.mff.cuni.cz
mlynar@knih.mff.cuni.cz
pavel.obdrzalek@mff.cuni.cz

Challenges of integrating multiple divergent audiovisual oral history collections: The case of the Malach Center for Visual History in Prague

Keywords: Oral history / Audio-visual data / Digital archives

Digital collections of audiovisual oral history interviews (OHIs) can be conceived as a specific category of large text corpora that require implementation of distant reading methods. Following the digital revolution in oral history research (e.g. Thomson, 2007; Boyd & Larson, 2014), many institutions now provide access to divergent digital collections at once, which provides the users and researchers with an opportunity to combine and compare, for example, several OHIs conducted with the same person in varied situational and socio-historical framework as part of various "process-generated" (Freund, 2009) oral history projects. This constitutes a pertinent issue for such access institutions: How can we make it as easy as possible for users to work with several separate digital archives of OHIs? Which computational methods can facilitate distant reading and efficient use of large collection of audiovisual OHI materials? And what are the persistent problems -- technical, methodological, ethical, etc. -- that have to be solved by the institutions of multiple access?

In outlining answers to these questions, in our presentation, we will discuss our experience and technological solution from the Malach Center for Visual History (CVHM) at the Charles University in Prague (see Mlynář, 2015). Over the last decade, CVHM has been providing access for students, researchers and general public to several established collections of OHIs. Since 2009, CVHM is an access point to the USC Shoah Foundation's Visual History Archive (VHA), which is an ever-growing collection of interviews with witnesses and survivors of genocides, especially the Holocaust. At the present moment, the VHA contains nearly 56,000 audiovisual recordings of OHIs in more than 40 languages (see Gustman et al., 2002). Since 2018, the Fortunoff Video Archive for Holocaust Testimonies of the Yale University Library with more than 4,400 audiovisual recordings of OHIs is also available at CVHM. In addition, users in CVHM can work with smaller collections lacking an integrated user interface such as the Refugee Voices archive (150 English interviews), and a small portion of interviews from the Jewish Holocaust Center in Melbourne (15 interviews with people of Czechoslovak origin).

The present situation of hosting several disparate collections at once poses several challenges on the level of improving effectivity of search and textual corpora analysis methods. On one hand, users-researchers are in need of a complete metadata overview, in order to generate relevant datasets, discover duplicate cases and analyze OHI collections' profiles. On the other hand, the question of performing content-based queries across the whole corpus is imminent. For these purposes, CVHM developed an in-house interface integrating several collection at once providing solutions to both of these issues. Besides providing general access to metadata, automated speech recognition based transcripts (generated by AMALACH algorithm, post-edited) serve simultaneously as textual data for the multilingual cross-corpus search in English, Czech and Slovak and searchable automatically generated keywords dataset (KER - Keyword Extractor provided by LINDAT/CLARIN). Owing to this approach, users are able to easily acquire relevant results on several levels without the necessity of separately accessing the individual collections or OHIs.

Thomas Koentges

University of Leipzig, Germany

email:
thomas.koentges@uni-leipzig.de

Measuring Philosophy in the First 1,000 Years of Greek Literature

Keywords: Corpus Building / Historical Language Processing / LDA Topic Modelling / CTS/CITE / Framework / Epidoc XML

With the increased availability of texts, knowing exactly which passages of the many now available apply to any given research question is a challenge. Recently, the international Open Greek and Latin project (OGL) has made available close to 30 million words of literary Ancient Greek text under an open license (the First1KGreek corpus). This workflow involves Optical Character Recognition (OCR), XML editing, and the curation of library metadata. The OGL team is transforming a book- and page-based citation format to a work- and section-based citation format, which means that a machine can identify not only the start of a work, but also the start of subsections in that work, or even sub-sections of that sub-section. This provides granularity and makes the OGL Corpus the largest open machine-actionable collection of Ancient Greek. It ranges from works by philosophers, such as Plato, and Ancient commentary, such as the Scholia in Pindarum, to Ancient medical texts. If philology is the art of reading slowly, this paper shows a method of identifying passages to read slowly when faced with a 30-million-word corpus.

The author used LDA topic modelling—that is, a statistical method to find recurring patterns of co-occurring words—to produce a topic model for all Greek CTS-compatible text in OGL’s First1KGreek corpus. The goal was to see whether a machine can detect philosophical texts after only a few hours of modelling. While the most common topics for Plato and Aristotle seemed to differ at first glance, this is simply because the most common topic in the Corpus Platonicum is a structural one associated with the form of a dialogue. Once that topic was removed, I could distil two characteristic topics for philosophy: one topic related to scientific inquiry and one topic related to the nature of virtue. Correlation analysis showed that the two topics are independent. Not only can these now be used to express a numeric measure for philosophical text in Ancient Greek, we can also trace ‘philosophicalness’ through the whole corpus. If we use the ‘scientific inquiry’ topic our search will lead us to Aristoxenus of Tarentum, Nicomachus of Gerasa, Theon of Smyrna, and so on. In short, philosophical texts with mathematical or astronomical connections. If we search for the ‘virtue’ topic, we find Plotinus and Maximus of Tyre among many other philosophers. In this pilot study it seems that we can express some kind of philosophical thinking as numeric values of the dimension ‘virtue’ and the dimension ‘scientific inquiry’. While it is not surprising that text by philosophers is philosophical, it is now also possible for a machine to extract passages from works by non-philosophers that contain philosophical arguments based on only a few hours of modelling and human analysis. This paper will describe the corpus generation, the modelling and testing, and the application of the model as a finding aid for those seeking to read Greek slowly.

Leonard Konle

Universität Würzburg, Germany

email:
leonard.konle@uni-wuerzburg.de

Semantic Zeta: Distinctive word cluster in genre

Keywords: word embeddings / metric learning / genre / authorship attribution

Introduction

Burrow’s Zeta (Burrows 2007) is a common method in the field of Digital Humanities to distinguish between two groups of texts by measuring word frequencies in equal sized chunks of data. While its origins are in stylometric and authorship research it is also used to give insights on differences in various other categories like genre or literary epochs. While raw word frequencies seem suitable for authorship attribution, they have shortcomings in identifying words for genres as pointed out in the following example:

Given 100 novels from the genre of western where 80 are written by an author favoring the word “colt” and the rest by a colleague using “revolver” instead. Both words would be good markers to classify authorship, but if the topic of interest is to find distinctive words for western in contrast to love genre there is a good chance of missing the word “revolver” due to its low frequency in the overall western group. If Zeta results are used for a clustering, texts from the “revolver”-author will show as less generic for western.

To resolve this issue a more abstract semantic class of distinctive words is needed. In the example something like “old guns” would be a solution. In the following a method combining Zeta, Word Embeddings and Metric Learning is proposed to meet this requirement.

Method

Texts from both groups are split into test and trainingdata. Than Zeta is applied to determine distinctive words for both groups in the trainingset. Schöch 2018 evaluated different variants of zeta coming up with sd2-zeta as the best choice. These words are transformed into vector representation using a pre-trained embedding. After that unsupervised clustering (Zhang et al. 1996) shrinks the embedded zeta words to more abstract cluster centers. These centers are utilized to bend the embedding space with supervised metric learning. This is performed with Uniform Manifold Approximation (McInnes et al. 2018) and aims at altering the vector space in a way, that cluster centers from both groups are near to other centers of their own class forming a cluster of clusters, while extending the distance between the two classes to a maximum. For the final classification the distance of all token to their nearest cluster center from both groups is calculated. These distances get subtracted from another resulting in a final value for semantic distinctiveness. To determine the quality of distinctiveness scores, classification with linear regression takes these values as input and predicts its group.

Resources

The corpus holds a collection of low brow novels from the genres Love, Medical, Regional, Family, Crime, Horror and Science Fiction (200 each) from various authors. The word embedding is trained on larger corpus containing 20.000 novels with FastText (Bojanowski et al. 2017).

Results

20-fold cross validation over four pairings of genre shows a slight advantage of semantic zeta (.90 f1-score) over sd2-zeta (.85 f1-score). This effect is reversed in authorship attribution leaving sd2-zeta as the prefered choice.

Cvetana Krstev – Jelena Jaćimović – Branislava Šandrih – Ranka Stanković

University of Belgrade, Faculty of Philology, Serbia
University of Belgrade, School of Dental Medicine, Serbia
University of Belgrade, Faculty of Philology, Serbia
University of Belgrade, Faculty of Mining and Geology, Serbia

email:
cvetana@matf.bg.ac.rs
jelena.jacimovic@stomf.bg.ac.rs
branislava.sandrih@fil.bg.ac.rs
ranka@rgf.bg.ac.rs

Analysis of the first Serbian literature corpus of the late 19th and early 20th century with the TXM platform

Keywords: textometry / Serbian literature / specificity / Correspondence Factor Analysis

The SrpELTeC corpus is a Serbian part of the ELTeC corpus that is being compiled in the scope of the COST Action 16204 – Distant Reading for European Literary History which would consist of 100 novels per language that were first published 1850-1920. Contrary to a number of other European languages involved in this action, the Serbian corpus is being produced from scratch, because the vast majority of novels from this period were not digitized before, they were not digitized in the proper manner or were not available. This involved several steps: selection of novels, retrieval of hard copies, scanning, OCR, automatic correction of OCR errors for which a specialized tool based on the Serbian morphological dictionaries (Krstev, 2008) was produced, human correction of remaining errors and basic structural annotation for which a number of volunteers was recruited, and for metadata preparation as well.

Current version of the srpELTeC corpus is available from:

https://distantreading.github.io/ELTeC/srp/index.html

Texts of the srpELTeC corpus are stored in the XML-TEI format. TEI headers contain bibliographic information about electronic and original version of the novel, as well as information about persons responsible for preparation of the electronic version. Texts are structurally annotated, containing information on the logical structure of the text. In addition, metadata are also used in the TXM environment (Heiden, 2010ab) as a new structural text element, represented by the following attributes: author, title, date, author’s gender and type. The srpELTeC corpus was imported into the TXM environment using XML-TEI Zero + CSV module, while lemmatization and PoS tagging was done using the TreeTagger software (Schmid, 1994; Utvić, 2011).

The srpELTeC corpus consists of texts containing a total of 935.902 concrete token realizations. The results say that in this corpus, which has 78.542 token, 32.604 different lemmas were used. Based on the morphological tags, a list of occurrences of certain grammatical categories is generated. In addition to nouns and verbs that dominate the texts of the srpELTeC corpus, pronouns are also highly represented.

Frequency distribution of nouns, adjectives, verbs and adverbs in the srpELTeC corpus divided into partitions based on authorship is calculated. The results clearly show parts in which adjectives are extremely specific, as well as authors who use nouns much more often than others. Elements which are far less used compared to their degree of use throughout the corpus is reflected in the high negative value of the specificity score (Lafon, 1984).

Based on the specificity scores, Correspondence Factor Analysis (CFA) (Benzécri , 1973) is conducted. Each calculated specificity score is represented by a specific position on the factor map. It is noticeable that some PoS, more commonly used in certain partitions, are located on the positive sides of the horizontal or vertical axis, while on the opposite sides are partitions and PoS with negative specificity scores.

Cluster analysis of the matrix obtained by the previously carried out CFA was also made. Resulting dendrogram shows a hierarchical grouping of texts based on the authors and used lemmas, providing better understanding and simpler interpretation of the results of the CFA.

Davor Lauc – Darko Vitek

University of Zagreb, Croatia

email:
dlauc@ffzg.hr
dvitek@hrstud.hr

Developing Logic of Inexact Concept Comparison for Historical Entity Linking

Keywords: Historical Entity Linking / Temporal logic / Inexact concept reasoning

Named entity linking and related procedures like entity resolution, database deduplication and Wikification, is a crucial step in many endeavours such as semantic search (Derczynski, 2015), improving performance of digital libraries (Smalheiser, (2009), distant reading (Lauscher, 2016) and many others. Due to the non-existence of a comprehensive entity database and standard challenges of natural language processing, the state-of-the-art entity linking systems are still facing many difficulties (Kolitsas, 2018). Those problems are even more challenging when named entity linking is engaged in the context of digital humanities, especially for resolution of historical entities (Won, 2018) (Cura, 2018). In these contexts, standard disambiguation challenges are even more robust because of multiple language usage, temporal changes of person, organisation and geographical names, the noise created by sources digitalisation and many others. In that context, decisive predictors for successful entity resolution are often buried in inexact concepts like born at the beginning of the 18th century, occurred in the Austro-Hungarian Empire or near the city of Venice.

In this research, we are developing logic for comparison of inexact spatial-temporal concepts that can be used to improve entity linking — extending the previous research on reasoning with inexact dates (Lauc, 2019) we are representing inexact dates as a discrete probability distribution over the timeline and the inexact geo-location as a multivariate probability distribution over locations. For concept comparison, the joint probability is calculated using strong naive Bayes assumption. To achieve usability in semantic search and named entity linking systems, we have developed deep learning neural model to reduce the dimensionality of massive sparse probability distributions to standard dense vectors embeddings. The evaluation was performed on entity linking of historical person’s to Wikidata entities, and preliminary results demonstrate significant improvement of F1-score when dense vectors representing inexact concepts are added to the resolution system.

Anna Moskvina

Universität Hildesheim, Germany

email:
moskvina@uni-hildesheim.de

Analysing Online Book Reviews - A Pattern Based Approach

Keywords: Online Book Reviews / Manual Annotation / Machine Annotation / Corpus search / Queries / Pattern Search

The following research is a part of a bigger project aimed at a thorough analysis of online book reviews, investigated from different angles and by different disciplines. The focus of this paper is on utilising methods of computational linguistics to the needs of literary studies and to what extent these methods can be helpful in qualitative analysis of book reviews.

The cornerstone of our main project is to look into the modern reviewing process, analysing the characteristics of reviews, their cross-platform similarities and differences.

As a preprocessing step we devised a hierarchical annotation scheme in the spirit of Grounded Theory methodology that grasps as many various aspects of a review as possible: from narrating a book plot to the use of metadata, such as an overall structure, headlines, use of keywords, number of likes and shares. The scheme was later approbated during the manual annotation (sentence level) of 430 reviews, chosen from 27,857 Amazon book reviews (McAuley, J., Leskovec, J. 2013).

For this paper we restrict ourselves only to the aspects that are of primal interest to the field of literary criticism. Some of the chosen categories and the number of annotated sentences are listed below in the following format:

annotation label - annotation title - number of annotated sentences.

A1.4.4 - Emotions present in the reviewed work - 69

1.4.5 - Logic/Consistency of the reviewed work - 64

A1.8 - Form/Writing Style of the reviewed author - 398

A3 - Consequences of the perception - 192

A4 - Emotions of the reviewer - 316

A6 - Motivation for the review and or /reading the book - 78

A7 - Expectations before reading - 100

As the traditional machine learning techniques usually yield less than moderate results on the scarce data (after the first training attempt the average F-score for the categories never exceeded 0.16), we have decided to enlarge our dataset by analysing the structural patterns that distinguish the annotated sentences as belonging to a certain category.

Such patterns were manually extracted, designed according to the CQP query language (Evert, S. 2005) and, afterwards used in queries applied to the full corpus.

The extraction of indicators cannot be limited to the search for single words or word combinations. Instead, morphosyntactic variations (constituent order, passivization, etc.) and lexical variation (synonyms, related items, nominalizations, etc.) need to be accounted for; in addition, the extraction must be aware of the context, due to the polysemy and to the broad range of contexts in which some of the indicator expressions can appear. For example, the main difference between the categories A4 and A1.4.4 is who experiences the emotions: A1.4.4 «Natürlich wird Maik neidisch» («Of course Maik gets jealous»), vs. A4 «Mich hat es zu Tränen gerührt!»(«It brought me to tears»). Moreover, out of 117 occurrences of the word «deswegen» («therefore»), only 12 were related to the motivation of reading/buying the reviewed book.

The number of resulting patterns for each category with regard to the number of annotated sentences is listed below (Annotation label - number of sentences - number of patterns)

A1.4.4 - 69 - 8

A1.4.5 - 64 - 10

A1.8 - 398 -11

A3 - 192- 9

A4 - 316 - 16

A6 - 78 - 10

A7 - 100 - 11

Some patterns were combined with external sources like Germanet (Hamp, B., Feldweg H. 1997) in order to find more synonyms of a particular word, without losing the connection to the semantic field of the indicator, for example: «enttäuscht» - «trauervoll» - «bitter» - «unglücklich» («disappointed» - «mournful» - «bitter» - «unhappy»).

In the next phase of our research, we plan to investigate if the established patterns are suited for search in corpora, aggregated from other platforms, such as literary blogs or book fora and to establish cross-platform similarities and differences.

Minako Nakamura – Kohji Shibano

Ochanomizu University, Tokyo, Japan
Tokyo University of Foreign Studies, Japan

email:
nakamura.minako@ocha.ac.jp
shibano@aa.tufs.ac.jp

Mining formulaic sequences from a huge corpus of Japanese TV closed caption

Keywords: Formulaic Sequence / Spoken Corpus / Closed Caption

As a result of quantitative studies of large corpora, including the British National Corpus (BNC 1993), such insights as usage-based linguistics (Barlow et al 2000), formulaic sequences (Schmitt ed. 2000), and formulaic linguistics (Wray ed. 2008) have gained attention.

The CEFR (Council of Europe 2001) was widely accepted as the standard for language education. However, the abovementioned tidal changes in linguistic studies have not significantly influenced language education because there are few corpora of languages that cover spoken interactions.

The BNC's most significant characteristic is its inclusion of a spoken corpus. Google (2006) released the Web 1T corpus of n-grams as well as a large Japanese n-gram corpus (2007). A corpus of spoken Japanese, known as the Corpus of Spontaneous Japanese (CSJ) (Maekawa 2006), was also developed. Similar to the BNC spoken corpus, the CSJ is a corpus of spoken productions that consists primarily of academic presentations. The CSJ and the BNC spoken corpus consist of 7 million words and 10 million words, respectively. Compared with web or written corpora, spoken corpora are still small.

To overcome this limitation, we first developed a system to record Japanese TV programming from all seven nationwide networks; the system has been operational since December, 2012. Approximately half of the TV programs have closed captions. From 125 thousand programs that totaled 75 thousand hours, we developed a spoken corpus of 500 million words from the closed captioning texts.

For an analysis of formulaic sequences, we generated all of the n-grams, a computation that requires a large data processing capacity. Thus, we applied Google's MapReduce algorithm

(Dean et al 2008) and obtained formulaic sequences of spoken Japanese. 214,265 formulaic sequences with more than 100 occurrences have been found.

Details of the recording system, MapReduce algorithm, and analysis of the formulaic sequences will be provided.

Ovio Olaru

Babeș-Bolyai University, Romania

email:
olaru.ovio@gmail.com

The Swedish crime fiction boom in numbers. Quantitative approaches to Scandinavian Noir.

Keywords: metadata mining / Scandinavian Noir / Book market analysis

This individual presentation will test whether the recent international literary phenomenon originating in the Nordic countries, "Scandinavian Noir", better known under the name of "Schwedenkrimi", can really be regarded as synonymous with the Swedish production of crime fiction, as its title suggests. We will be putting this assumption to the test by looking at how contemporary Swedish crime fiction authors performed internationally. By employing Zotero and Palladio and mining metadata from the online catalogue of Sweden's National Library, from the online catalogue for translated Swedish Literature, "Suecana Extranea", as well as from the online catalogues of several influential agents of the international book market (Such as Germany or France), we attempt to draw a translational network spanning 2004 and 2017 starting from several renown authors of Swedish crime fiction featured on the Swedish bestseller list during that time and that ultimately enjoyed international success.

A simple inquiry shows that a few names dominate the Swedish bestseller lists during this period. By comparing two significant European book markets, the French and the German one, we attempt to test whether these authors owe their success entirely to their German renderings and, implicitly, if the title "Schwedenkrimi" has been correctly coined.

As for the theoretical background, we will be relying both on important works in the field of "Scandinavian Noir", as well as on seminal works in the field of Digital Humanities, of computational analysis and of distant readingam Main (1971).

Róbert Péter

University of Szeged, Hungary

email: robert.peter@ieas-szeged.hu

Distant Reading of 18th-century English Newspapers: Challenges and Opportunities

Keywords: distant reading / bibliographic data / 18th-century newspapers

The aim of this paper is to discuss the opportunities and challenges that one encounters with the distant reading of 18th-century English newspapers and periodicals when using digital humanities methods and tools. It demonstrates how the investigation of bibliographic records of thematic full-text databases has the potential to identify hitherto overlooked trends and themes in 18th-century studies. The introduced case studies demonstrate that posing simple statistical questions about the well-selected and cleaned metadata of thematic databases can highlight ignored patterns, processes and relations in cultural studies and trade relations. The first case study explores the changing public perception of English Freemasonry by investigating press accounts of the fraternity. It sheds new light on the considerable impact of a theatrical performance and a leading masonic newspaper editor on the perception of the contemporary fraternity. The second case study provides a distant reading of the largely unmapped English press material about Hungary, which, for instance, signposts new areas for research into Anglo-Hungarian economic relations of the period. The analysis has been carried out by the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) research tool, which is being developed as part of an ongoing digital humanities project at the University of Szeged. Its functions include (i) the analysis and interactive visualisation of the networks of authors, publishers and booksellers, (ii) the automatic identification and visualization of the sex of the authors in several languages, (iii) the interactive modeling of the relations between the different metadata types, (iv) the N-Gram analysis of metadata and texts, (v) named entity recognition.

Orsolya Putz – Zoltán Varjú

Eötvös Loránd University, Budapest, Hungary
Montana Knowledge Management

Email:
putz.orsi@gmail.com
zoltan.varju@gmail.com

Distant reading of US presidential speeches. Diachronic changes of the concept of American nation

Keywords: close and distant reading / word embeddings / distributional semantics / American national identity / construction

Cognitive linguistic researches on the concept of American nation in political discourse (Putz 2018, Putz manuscript) suggest that the concept of American nation is dominantly constructed by the conceptual domains (Barcelona et al. 2011, Croft et al. 2004) of person, community and object. This result is partly based on a qualitative analysis of the speeches of seven influential American presidents out of forty-five. Consequently, due to the limitations of close reading, no conclusion can be drawn from such an analysis about the diachronic changes of the concept of American nation.

As a contribution to the aforementioned projects, this paper intends to trace the evolution of the concept of American nation in American presidential speeches from the very beginning of US history until nowadays. Furthermore, this paper intends to demonstrate that the most successful and fruitful research projects adopt both close and distant reading. It is proposed that the two approaches jointly provide a complex view of the studied phenomenon. The following research questions are answered: 1) What concepts did the particular presidents associate with the concept of American nation? 2) What features distinguish the way they construct American national identity?

As for the research method, distant reading is considered the most suitable method to process all the presidential speeches of the US from 1789 to 2019 and to reveal words that are most frequently associated with the term of nation. The following steps were executed.

Firstly, an extensive corpus of the US presidential speeches was compiled, which contains all the presidential speeches of the US. Data was automatically collected by Data Collector, which is an intelligent data harvesting software of Precognox Ltd.

Secondly, eleven sub-corpora were generated, in line with the periods of American history (Kutler 2003). The sub-corpora comprise the presidential speeches of the following periods: 1776–1789, 1789–1849, 1849–1865, 1865–1918, 1918–1945, 1945–1964, 1964–1980, 1980–1991, 1991–2008, 2008–present.

Then, after data pre-processing, a model was trained to find the closest words in the vector space to the key word of nation in each sub-corpus. Although neural word embeddings (like Glove and word2vec) are treated as dominant approaches in Machine Learning, their priority is often questioned (e.g. Levy et al 2015) and they are replaced by alternative methods. Inspired by this idea, a traditional pointwise mutual information-based word embedding was trained and reduced by singular-value decomposition. The diachrone analysis of the words was motivated by Hamilton et al's (2016) work and was based on Procrustes analysis.

Finally, the output of the models were carefully studied, namely the top 20 closest words to the nation per sub-corpus and the distance between certain selected words (e.g. Americans) and the nation.

It was found that 1) the concepts associated with the American nation are related to the person, community and object source domains, in harmony with Putz's (forthcoming) previous findings. 2) The results suggest that the meaning change of the concept of American nation is motivated by the social, economic and political atmosphere of the period in which the speeches were formulated.

Mark Ravina

University of Texas, Austin, United States

email:
mark.ravina@austin.utexas.edu

Distant Reading Political Discourse in Nineteenth-Century Japan

Keywords: distant reading / Japan / text mining / Meiji Restoration / political discourse / topic modelling

In 1867, after toppling the Tokugawa shogunate, the new Meiji government of Japan declared its openness to public discourse. It would henceforth receive petitions from all imperial subjects, regardless of rank or status: ordinary commoners as well as court nobles could now address the central government. Between 1867 and 1889 thousands of Japanese used that new avenue to voice their complaints and concerns. The petitioners represented a striking cross-section of Japanese society, including commoners, Buddhist monks, Shintō priests, former samurai, and former daimyo lords. The topics are equally diverse, including the need for a constitutional convention, mining subsidies, international treaties, conscription, education reform, and the government's regulation of marriage ceremonies.

Historians have long been fascinated by these texts and in the 1980s Gabe Takao and Irokawa Daikichi painstakingly collected, transcribed, and published nearly 3000 petitions culled from archives across Japan. (Irokawa and Gabe 1986-2000) This corpus is widely available, but its size overwhelms conventional close reading. Historians have examined small groups of petitions, but neglected the broader corpus. Sadly, the petition corpus has become what Margaret Cohen describes as a "great unread." (Cohen 1999, 23; Khadem 2012, 410)

My project employs "distant reading" to explore these neglected texts. My presentation will focus on two discoveries. First the language of petitions varied with class. Commoners wrote primarily about practical considerations, such as the details of local economic development programs, tax collection, and school administration. Former samurai, by contrast, were concerned more with foreign policy, in particular the renegotiation of unequal treaties, and with abstract questions of sovereignty, such as the emperor 's relationship to elected assemblies.

Second, petitions employed a syncretic language, reworking older Chinese-style terms while creating neologisms for Western ideas. When Japanese activists advanced ideas such as government with the consent of the governed, they freely mixed neologisms, such as the term minken for the foreign concept of “civil rights,” with older Asian political terms. Meiji-era discourse seamlessly combined references to the Confucian classics, invocations of samurai loyalty, and ideas adapted from J.S. Mill and Rousseau. Using distant reading, I show how classical Chinese and modern Western expressions were invoked across thousands of texts to create a new political language. Using topic modelling, I trace how classical Chinese terms changed valence as part of an emerging modern political vocabulary.

Jake Reeder

Marist College, New York, United States

email:
jreed03@gmail.com

Returntocinder.com: A Concordance-Style Research Index Inspired by Jacques Derrida

Keywords: deconstruction / derrida / digital pedagogy

I am proposing to introduce and discuss the website returntocinder.com, a concordance-style electronic database inspired by the work of Jacques Derrida. When first created, the database divided over 14,000 entries from over 100 works by Derrida into 841 motifs. The motifs include philosophers' names (Hegel, Heidegger, Descartes, etc.), famous ideas attributed to Derrida (pharmakon, supplement, différance, dissemination, etc.), and many classical concepts like justice, law, knowledge, and love. Each entry is what I call a quasi-quotation: it is a condensed paraphrasing or citation that maps out the page or pages to which it refers. The site allows you to search by motif, by source, or to search any words or phrases in the entries.

In my presentation, I want to discuss the way in which the organization of the site provides a unique efficiency and a unique pedagogical instruction. Regarding efficiency, each entry includes an initial that refers the user to the original book or essay, a page number, and a fragmented sentence or two. The entries or sentences never use capitalizations, giving the reader a clear sense that the passage has been cut out from somewhere else. This notion of "having been cut out" forces the user to face the prosthetic value of these entries. The reader is encouraged to follow more threads, whether they be thoughts, new entries, or the retrieval of the book in question. Regarding pedagogy, I believe it is important to realize that "people don't always know what to search for." By distributing entries into key terms or motifs, returntocinder.com not only helps you find information but it also provides suggestions that will help you organize your own research. This final point leads directly to a spinoff project of returntocinder.com, a research note app called Databyss, which is currently under construction. Databyss will allow other thinkers, annotators and researchers to create similar style databases of their own. Far in the future, I imagine that people will be able to search in and out of each other's databases.

The website has already expanding beyond Derrida. It now includes entries from Nietzsche, Sharpe, Marx, Moten, Augustine, Plato, Kierkegaard and more. This, as one can imagine, exponentiates the possible usages of the site, and speaks of a rich dissemination of the cornerstone ideas of the Western canon. During the workshop, I would like to walk people through the many features of the site, explaining how to find the bibliography for each text, how to utilize the search engine, and how to move back and forth between each author. Finally, I will introduce wireframes of the new app (Databyss), which I hope will eventually encourage an enormous crowdsourcing of researchers' margin notes.

I should add that I do not think of this site as a data visualization project. All my notes were manually entered, and I treat this project as an attempt to explore and perform what Derrida calls the contaminated border between thought and machine.

Gabriele Salciute Civiliene

King's College London, United Kingdom

email: gabriele.salciute-civiliene@kcl.ac.uk

Distant, Deep, Thick Reading: Towards the Embodied Computation of Texts across Languages

Keywords: translation / psychodramatic response / word repetitions / cross-linguistic computing / distant reading / thick reading

Word frequencies have a wide currency in textual scholarship. While using them to study texts in one language is rather straightforward, we run into a whole range of complexities and complications when in attempt to trace what becomes of them in translation. Not only words become estranged due to their linguistic transformation in another language, but they also shift in how they are interpreted by translators and their editors, banned and censored by ideologies of the time as well as in how they evoke and provoke personal commentaries on social, ethical or existential issues either featured in or mediated by the original text. A translation cannot be denied its authenticity in what it becomes and what actual role it plays in cultural dissemination despite the original-derivative ontology that underlies the convention of how we position it vis-à-vis its original text. As implied by Venuti (2009), translations and their originals are better defined as intertexts within a large network of other texts.

Although we are not short of versatile cases of what becomes of the original text in translation, the slippery area is where our methods produce thin descriptions and reductive theories. The mainstream theory of how repetitions are translated is a case in point. It argues that translators tend to avoid repetitions for aesthetic reasons regardless of the contexts in which they appear to be working (Toury 1991; Ben-Ari 1998; Malmkjær 1998; Jääskeläinen 2012). A closer look at how evidence is gathered and interpreted to support this claim reveals a deep inter-semiotic problem: selective word counts of omission are taken to embody aesthetic intolerance. Yet the epistemological impasse is imminent because, first, the founding question whether we lose repetitions is empty. The phenomenological remedy here would be to ask how we lose those repetitions. Second, word counts *per se* do not represent what they might have summoned in a translator's memory and imagination. Instead, they need to be modelled to 'resemble' complexities inherent in human response.

In this paper, I will overview my work done on the modelling of translatorial response as an experiential construct. The major task has been to develop the methodological and practical framework for a distant cross-linguistic reading based on aligning Faulkner's *The Sound and the Fury* with its translations by word frequencies. Unlike mainstream cases that explored repetition in an unsystematic, and discrete manner, the model here relies on the premise that repetition may have a resounding psychodramatic effect (Leech and Short 2007) on translators and that their responses to repetition would take certain shapes. Hence, Faulkner's repetitions and translatorial choices have been modelled as wave-like patterns of fluctuation. Computed by shapes rather than by disembodied counts, the translations emerge as unique constructs. I'll also revisit the notion of distant reading to argue for deep and thick reading or computing (notion akin to 'thick description' in Geertz 1973) that would aim at recovering experiential diversity and multiplicity behind a linguistic sign rather than sameness as argued in the reductive theories of repetition in translation.

Diana Santos – Alberto Simões

University of Oslo, Norway
2Ai Lab - IPCA, Portugal

email:
d.s.m.santos@ilos.uio.no
asimoes@ipca.pt

Towards a computational environment for studying literature in Portuguese

Keywords: literary analysis / annotated corpora / emotions

We present here Literateca, a literary and linguistic environment to study texts written in Portuguese, which currently includes more than 800 literary works, and which allows one not only to query the individual data but also create visualizations of sets of literary works.

In addition to traditional morphosyntactic annotation and named entity annotation, performed by PALAVRAS (Bick, 2000), we also make available annotation of such semantic domains as colour, body parts, clothing, health, family and emotions.

While for many functionalities we use separately R modules and/or additional Perl programs, which should later be better integrated in the environment, we believe the setup already allows several interesting research alleys.

In this paper we will present the following:

a) Emotions across time, and their relationship with literary school: we will first explain how emotional annotation was performed, how classification of literary school was achieved, and then provide data on the total of emotions and also per emotion

b) Speech description profiles of different authors (reported speech, presence of specific speech verbs, emotional speech verbs): we will explain how we measure reported speech (mainly direct and indirect), and the issue of emotional speech verbs, very relevant in Portuguese

c) Feelings associated to clothing: we measure the co-occurrence of mentions of clothing together with emotions, colours, and body parts

d) The presence of the medical profession in lusophone literature: we investigate the presence of doctors and nurses in the literature, and explain why

We will also name briefly related work in progress being done on topic modelling and character networks, using the same underlying material.

Thomas Schmidt

University of Regensburg, Germany

email:
thomas.schmidt@ur.de

Distant Reading Sentiments and Emotions in Historic German Plays

Keywords: Sentiment Analysis / Emotion Analysis / Distant Reading / Visualization / Lessing / Computational Literary Studies

Sentiments and emotions are important parts of the interpretation of literary texts (Winko, 2003; Mellmann, 2015) and are of special interest for literary scholars interested in plays. Furthermore, many famous playwrights focused on the role of emotions to construct a drama theory for their plays e.g. the *Katharsis* model by Aristotle (335 B.C./1994) or the *Affektlehre* by Lessing (cf. Fick, 2016). Therefore, the computational method to analyze sentiments and emotions in written text, sentiment analysis, has found its way into computational literary studies and quantitative drama analysis. Sentiment analysis is used to analyze fairy tales (Alm & Sproat, 2005), novels (Jannidis et al., 2016) and historic plays (Mohammad, 2011; Nalisnick & Baird, 2013) oftentimes with a focus on annotation and evaluation of various methods. However, only few studies explore possibilities of integrating Distant Reading (Moretti, 2013) and visualizations into sentiment and emotion analysis (Kakkonen & Kakkonen, 2011; Mohammad, 2011; Nalisnick & Baird, 2013).

I present a web-based Distant Reading tool to explore sentiments and emotions in all 12 of the historic German plays by *Gotthold Ephraim Lessing* (1729-1781). To calculate metrics I employ a rule-based approach using lexicons with sentiment annotations for words. This is a well-known approach in sentiment analysis and in a previous study, various lexicons and NLP techniques were evaluated on a sub corpus of speeches to identify the best performing method (including among other methods lemmatization and the extension of the lexicons with historic variants; Schmidt & Burghardt, 2018). The number of sentiment words in a textual unit are calculated to get an overall value. Values for the polarity (positive/negative) and eight emotion categories are calculated and users can regard either absolute values or values normalized by the number of all words of a text unit.

I distinguish between three main-concepts of analysis: structural analysis (analysis among acts, scenes and speeches), character analysis, and analysis of character relationships. The tool offers various visualization types to explore polarity and emotion distributions among those concepts like bar, column and line charts. There are several analyses I performed to examine the possibilities of Distant Reading emotions and polarities in the plays. Two use cases are briefly described in the following:

For the first use case, I analyzed the polarity progression throughout all the acts of Lessing's plays and were able to identify a constant progression to larger amounts of negativity leading up to the fifth act. This result illustrates how the plot becomes more and more negative since disputes and conflicts become more apparent towards the end.

For a more character-specific use case, I focus on the character *Marinelli*, who is the villain in the play *Emilia Galotti*. Analyzing his speeches in the entire play indeed shows a tendency towards negativity. Comparing Marinelli's speeches to other characters, he is ranked as the most negative character in absolute values. Analyzing the character relationships, the results for Marinelli are in line with the plot. To calculate relationships I follow Nalisnick and Baird (2013) and regard heuristically every speech a character expresses as directed towards the previous speaker and take all those character-to-character speeches as input for the sentiment analysis of relationships. For many of the main characters the relationships prove to be rather negative. However, it is telling that the relationships towards *Angelo, Pirro* and *Battista* are positive since those are *Marinelli's* allies. A positive relationship between Marinelli and Emilia shows the limitations of sentiment analysis. Although Marinelli plans to damage Emilia throughout the entire play and his behavior is instrumental in her committing suicide, the relationship is shown as positive. The reason for this is that when both characters meet in the play, Marinelli acts very nice and polite disguising his true intentions, which cannot be noticed on a solely textual level.

I will present more examples illustrating Distant Reading possibilities of sentiments but also emotions in plays and discuss limitations and future work in more detail during the presentation.

Sentiments and emotions are important parts of the interpretation of literary texts (Winko, 2003; Mellmann, 2015) and are of special interest for literary scholars interested in plays. Furthermore, many famous playwrights focused on the role of emotions to construct a drama theory for their plays e.g. the Katharsis model by Aristotle (335 B.C./1994) or the Affektlehre by Lessing (cf. Fick, 2016). Therefore, the computational method to analyze sentiments and emotions in written text, sentiment analysis, has found its way into computational literary studies and quantitative drama analysis. Sentiment analysis is used to analyze fairy tales (Alm & Sproat, 2005), novels (Jannidis et al., 2016) and historic plays (Mohammad, 2011; Nalisnick & Baird, 2013) oftentimes with a focus on annotation and evaluation of various methods. However, only few studies explore possibilities of integrating Distant Reading (Moretti, 2013) and visualizations into sentiment and emotion analysis (Kakkonen & Kakkonen, 2011; Mohammad, 2011; Nalisnick & Baird, 2013).

I present a web-based Distant Reading tool to explore sentiments and emotions in all 12 of the historic German plays by Gotthold Ephraim Lessing (1729-1781). To calculate metrics I employ a rule-based approach using lexicons with sentiment annotations for words. This is a well-known approach in sentiment analysis and in a previous study, various lexicons and NLP techniques were evaluated on a sub corpus of speeches to identify the best performing method (including among other methods lemmatization and the extension of the lexicons with historic variants; Schmidt & Burghardt, 2018). The number of sentiment words in a textual unit are calculated to get an overall value. Values for the polarity (positive/negative) and eight emotion categories are calculated and users can regard either absolute values or values normalized by the number of all words of a text unit.

I distinguish between three main-concepts of analysis: structural analysis (analysis among acts, scenes and speeches), character analysis, and analysis of character relationships. The tool offers various visualization types to explore polarity and emotion distributions among those concepts like bar, column and line charts. There are several analyses I performed to examine the possibilities of Distant Reading emotions and polarities in the plays. Two use cases are briefly described in the following:

For the first use case, I analyzed the polarity progression throughout all the acts of Lessing's plays and were able to identify a constant progression to larger amounts of negativity leading up to the fifth act. This result illustrates how the plot becomes more and more negative since disputes and conflicts become more apparent towards the end.

For a more character-specific use case, I focus on the character Marinelli, who is the villain in the play Emilia Galotti. Analyzing his speeches in the entire play indeed shows a tendency towards negativity. Comparing Marinelli's speeches to other characters, he is ranked as the most negative character in absolute values. Analyzing the character relationships, the results for Marinelli are in line with the plot. To calculate relationships I follow Nalisnick and Baird (2013) and regard heuristically every speech a character expresses as directed towards the previous speaker and take all those character-to-character speeches as input for the sentiment analysis of relationships. For many of the main characters the relationships prove to be rather negative. However, it is telling that the relationships towards Angelo, Pirro and Battista are positive since those are Marinelli's allies. A positive relationship between Marinelli and Emilia shows the limitations of sentiment analysis. Although Marinelli plans to damage Emilia throughout the entire play and his behavior is instrumental in her committing suicide, the relationship is shown as positive. The reason for this is that when both characters meet in the play, Marinelli acts very nice and polite disguising his true intentions, which cannot be noticed on a solely textual level.

I will present more examples illustrating Distant Reading possibilities of sentiments but also emotions in plays and discuss limitations and future work in more detail during the presentation.

Alexander Schütze

Ludwig Maximilian University of Munich, Germany

email: alexander.schuetze@lmu.de

A 'distant reading' of officials' statues from Ancient Egypt’s Late Period

Keywords: Ancient Egypt / statuary / titles / prosopography / typology

High officials of Ancient Egypt’s 26th dynasty (664–526 BCE) are a well-documented social formation due to hundreds of inscribed monuments like sarcophagi, statues, shabtis, stelae, seal impressions etc. being dispersed over many museums and private collections in Egypt and the Western world. Among these monuments, more than five hundred statues of these officials are of particular interest as they are complex three-dimensional objects combining iconographic and stylistic features with religious and biographic texts.

Surprisingly, only a limited number of these statues is frequently discussed in scholarly literature: Monuments of better-known officials are still overrepresented in studies of Late Period statuary. There are several reasons for this selective perception of Late Period monuments. On the one hand, the distribution of the statues over numerous collections complicates a proper overview over the material. On the other hand, our view on Late Period statuary is still highly influenced by Bernhard V. Bothmer’s Egyptian Sculpture of the Late Period although the selection of statues in this book is far from being representative for the whole material.

The situation seems to be structurally analogous to literature studies: Scholars of Western literature of the 19th century were used to confine themselves to a canon of selected works considered as classical and high quality literature although millions of other literary works of the very same period are known. Franco Moretti criticizing this scholarly practice proposed a distant reading of the whole literary production of the 19th century (instead of close reading a limited set of canonical literary works) to gain a more nuanced picture of this phenomenon of cultural production. Similar ‘distant viewing’ approaches play an increasing role in art history.

Inspired by Moretti’s thoughts, I would like to present a data science informed approach to the study of the statuary of officials of the 26th dynasty that considers its potential for both prosopography and art history of this very period. By applying tools of the statistical programming language R, I will illustrate how I analyse chains of administrative and/or priestly titles in the biographic texts as well as the iconographic features of the statues themselves in order to trace trends in the administration of the period in question and to solve problems of establishing a typology of these monuments. Finally, a precise dating these monuments will be discussed by applying a probabilistic approach.

Gábor Simon – Tímea Borbála Bajzát – Júlia Ballagó – Kitti Hauber – Zsuzsanna Havasi – Emese Kovács – Eszter Szlávich

Eötvös Loránd University, Budapest, Hungary

email:
simon.gabor@btk.elte.hu
bajzat.timi9696@gmail.com
julia.ballago@gmail.com
hauber.kitty@gmail.com
havasizsuzsi4@gmail.com
mesii@hotmail.com
szlavicheszter@gmail.com

Metaphor identification in different text types

Keywords: metaphor / annotation / structure / text types

One of the most actual questions in contemporary corpus-based and corpus-driven metaphor research is whether metaphoric expressions show text type specific variations or not (see Koller 2006, Partington 2006, Lederer 2016, Semino 2017). A further important aspect of this question is what kind of subcorpora are to develop in order to investigate the metaphorical structures of a certain language, exhaustively relying on corpus data.

Our paper investigates the tendencies of metaphorization in Hungarian texts belonging to different text types (literary and informative texts). To identify metaphorical expressions, the so-called MetaID-method is used (see Simon et al. in press): it is the adaptation of the MIPVU-process (Steen et al. 2011) for Hungarian. The method has been implemented on the WebAnno online annotation software (Eckart de Castilho–Mújdricza-Maydt–Yimam et al. 2016). In the present study, a 2000-word reference corpus serves as the basis for comparing different text types; this small-scale corpus was established in the course of methodological adaptation. The amount, the structural types, as well as the proportion of the metaphorical expressions, are scrutinized in research corpora of the same size, being built from representative texts of the chosen text types (literary texts in prose and informative literature). The aspects of structural analysis are the following: (i) the components of metaphorical structures (e.g. lemmas, inflections, arguments of a verb) and (ii) the semantic relationship between these components (e.g. elaboration of the primary or secondary figure of the verb, see Langacker 2008; possessive relation, other elaborating relations).

After demonstrating the use and the validity of the MetaID-method, the paper details the structure of the investigated corpora. Then it discusses the quantitative and qualitative results

obtained from the process of annotation. The main achievement of the research is a method for metaphor annotation that makes it possible to explore and compare the metaphorical potential of different text types in Hungarian.

Ranka Stankovic – Diana Santos – Francesca Frontini – Tomaž Erjavec – Carmen Brando

University of Belgrade, Serbia
University of Oslo, Norway
Linguateca & Praxiling UMR 5267 CNRS - UPVM3, Montpellier, France
Jožef Stefan Institute, Ljubljana, Slovenia
EHESS Paris, France

email:
ranka@rgf.bg.ac.rs
d.s.m.santos@ilos.uio.no
francescafrontini@gmail.com
tomaz.erjavec@ijs.si
carmen.brando@gmail.com

Named Entity Recognition for Distant Reading in Several European Literatures

Keywords: corpus annotation / named entity recognition / distant reading novel / annotation infrastructure

In this paper we discuss NER needs in the context of the COST action "Distant Reading for European Literary History", which encompasses European novels in the period 1840-1920. The first task was to choose the type of entities, having in mind cultural and language differences and from a literary analysis point of view. This paper presents the result of manual annotation of text samples for seven languages: Czech, English, French, Norwegian, Portuguese, Serbian and Slovene. The annotation included tags for people, roles (including professions), places, facilities, demonyms, organizations, and works with additional attributes assigned. We describe the setup, show some basic quantitative data and note interesting issues and annotation choices. Preliminary conclusions from the annotation experiment, a selection of tools for automatic annotation and the supporting software infrastructure will be presented in detail.

Charalampos Stergiopoulos – Georgios Markopoulos – Aggelos Malisovas – Dionysios Benetos

National & Kapodistrian University of Athens, School of Philosophy, Faculty of Philology Department of Linguistics, Greece

email:
cstergio@gmail.com
gmarkop@phil.uoa.gr
aggelosmalisovas@yahoo.com
dmpen@phil.uoa.gr

«Boethii, De institutione musica: A case study for the collation, editio, and annotation of the complete corpus through computer-assisted collation analysis»

Keywords: distant reading / digital humanities / auto-punctuation process / critical edition / computational Linguistics / computer-assisted collation analysis

A text witnessed in more than two manuscripts (witnesses) raises challenging issues that primary concern the collation of all the extant manuscripts; actions related to repeated norms are usually defined by systematic rules (Greetham, 2013). The comparison of witnesses during the stage of “collation” (Haentjens, 2014; Andrews, 2014) is a particularly laborious task and often leads to the “elliminatio”, the only feasible practice that allows reading and transcribing a limited number of witnesses. In the case of texts witnessed by a corpus of many manuscripts, the critical edition is a particularly time-consuming, not entirely defined process (Haentjens, 2014; Andrews, 2014; Schmidt, 2009).

Computational Linguistics, which involve natural language processing technologies, develops machine learning tools capable of training on specific datasets, thus improving their performance and help researchers to reduce arduous repetitive work (Schmidt, 2009). Several attempts have been made to create such tools for supporting critical editions and the creation of “apparatus criticus” (Haentjens, 2014; 2011; Jänicke 2017; Schmidt, 2009).

Our research is focused on the thorough analysis and publishing comparison of text correlation, which constitute the corpus of “De Institutione musica”, a written testimony by Boethius (Bower, 1978; André, 1990) that, played a decisive role in the development and evolution of the theory of music, in terms of technical rules and philosophical method. Aim of this collaborative research is the text processing of a vast volume that exceeds the amount of 8,000 transcribed words from a compilation of 178 manuscripts.

This ongoing project includes: a) the systematic use of existing tools, b) the development of new tools for the task at hand, and c) a computational processing methodology that will be able to produce quantitative and qualitative results. The suggested approach will focus on the process of automated “collatio” and the creation of “apparatus criticus” by considering the case of proper text punctuation and using as source a complete corpus of witnesses without numerical limitation.

In its preliminary phase, our research involves text preprocessing tasks in terms of tokenization, lemmatization, and part-of-speech tagging. Our training set consists of the two major codices of the corpus “De Institutione musica”. This stage consists the first step in designing a computational methodology for the representation of textual information and the detection of similarities and differences, while performing batch tasks that include a greater number of codices.

Katerina Tiktopoulou – Konstantinos Theodoridis – Vasilis Vasiliadis – Eleni Petridou – Anna Saggou

Aristotle University of Thessaloniki, Greece
Centre for the Greek Language, Greece
Aristotle University of Thessaloniki, Greece

email:
atiktopo@lit.auth.gr
ktheodo@gmail.com
vasvasilid@gmail.com
elpetrid@sch.gr
saggou_anna@hotmail.com

Building distant reading tools for handling variations/ polytype in spelling: the case of the "Digital Solomos" project.

Keywords: distant reading / corpus building / spelling variation / normalization / data transformation

This paper focuses on distant reading tools to be included in the "Digital Solomos" project currently being implemented at the Aristotle University of Thessaloniki. The project involves the digitization of the entire corpus of Dionysios Solomos' manuscripts and aims at providing access to digital photos and topographic transcriptions of almost all of the poet's autographs. Dionysios Solomos, not only the major 19th century romanticism Greek poet but also the national poet of Greece, has left most of his major works unfinished and unpublished, accessible only through untidy and fragmentary manuscripts. As a consequence, his manuscripts – and the textual material included in them – are considered the only "legitimate text" scholars might study to approach his works and poetics. The edition follows a tradition of documentary and genetic editing formed within the Modern Greek literary studies paradigm through the very study of Solomos' work and aspires to provide scholars of the field with useful tools for accessing and analyzing the corpus of his manuscripts.

The presentation mostly focuses on the problems the project team is facing in building a corpus that is not only suitable for close reading, but can also support distant reading approaches. Although the application of OCR/HTR technologies (namely Transkribus) on the printed transcription of the existing diplomatic edition (Politis 1964) results in a fully digital text that through automatic and manual annotation can be visualized in ways that facilitate close reading and interpretation, distant reading tools (advanced searches, indexing and concordances) are more cumbersome to build because of poet's additive bilingualism (Greek & Italian) and orthographic habits. Solomos, an ardent supporter of "dimotiki" (modern Greek vernacular) uses a very idiosyncratic, inconsistent and entirely personal orthographic system. His almost phonetic writing of the Greek language indifferent to the orthographic conventions of his time with alternation of Greek and Latin alphabet at word level create numerous versions of words thus complicating the quantitative analysis the corpus. In the presentation we describe how the project team was led to overcome the orthographic inconsistencies and peculiarities of the corpus through the process of transliteration, that is through an automatic transcription of the text in Latin that unifies orthographic allographs and permits the detection of all occurrences of a word throughout the corpus regardless of how it is written down originally. The application of the technologies proposed not only facilitates the study of a very important corpus of the Greek 19th century allowing its inclusion and analysis within larger corpora of the same period, but may also appear useful in other cases of Greek corpora or corpora of other languages where spelling inconsistencies and variations prevent the quantitative analysis of texts.

Chahan Vidal-Gorène – Aliénor Decours – Thomas Riccioli - Baptiste Queuche

École Nationale des Chartes-PSL / Calfa, France
Calfa, France
Calfa, France

email:
chahan.vidal-gorene@chartes.psl.eu

Crowdsourcing and Machine Learning: Case-study on Classical Armenian

Keywords: Crowdsourcing / Handwritten Text Recognition / Computer Vision / Databases Web Technologies / Classical Armenian

Calfa is a natural language processing (NLP) project for Classical Armenian with numerous purposes: providing resources for the study and the preservation of the language, and developing a Handwritten Text Recognition engine (HTR) combining the latest technologies in computer vision and NLP. For this research, the machine learning has become essential, especially neural networks (Deep Learning). The outcomes associated are promising, particularly concerning their implementation to OCR/HTR which are constantly improving. However, the processes are high costs in terms of resources and data. Massive annotated databases are needed to make these systems operational. Building such databases and ensuring their reliability is a crucial time-consuming step for which research teams don't always have time. Faced with that problem for the Calfa project, we developed crowdsourcing playful and advanced solutions for the users in order to build and annotate Classical Armenian characters/lines databases, such as tools for character/line labelling, text lemmatization, etc.

Through an introduction to the project and its tools, the presentation will be an experience feedback on crowdsourcing for automatic and manual corpus annotation, its needs, its limits, and on the challenge of reconciling the researchers' needs with simple and useful crowdsourcing tools.

Posters

Anikó Ádám

Péter Pázmány Catholic University, Budapest, Hungary

email:
adam.aniko@btk.ppke.hu

Distant Reading – Rich Reading: Reading and writing literary texts in the age of digital humanities in Europe

The communication aims to present the Erasmus+ Strategic Partnership project entitled: Reading and writing literary texts in the age of digital humanities.

During the three-year project coordinated by Pázmány Péter Catholic University, professors and researchers (members of the LEA (Lire en Europe aujourd'hui) group) from 13 European universities throughout France, Portugal, Spain, Luxembourg, Belgium and Hungary, and the Hungarian Academy of Sciences (MTA SZTAKI) are working on responding to the question what new reading strategies have developed recently, and what new and innovative methods can be applied in order to popularize reading in the age of communication flooded with images.

Based on the methodological and scientific results of the workshops and seminars of the project, a Rich Annotator System (RAS) will be developed to link the text of the commentary literature to some major literary texts, and form (a) direct hyperlinks from the comments to the quoted text and (b) by generating inverse links, enable a new form of reading of the main text where each commentary is immediately visible. With the methods of Rich Reading (RR) based on digital applications (RAS) and e-publication, we believe that we will be able to transform young people into strategic readers. The core idea is to offer and to suggest different kinds of reading, doing a genetic reading or / and a critical and augmented reading threw linked data, information and knowledge about the historical period, the literary or artistic context, the writer's role, work and life, or even though doing a linguistic reading, based on the return of words, images, rhetoric patterns (with visual, graphic and verbal semiotic aspects) and notifying significance networks.

The digital tool will be based on a library of French literary texts (1800-1945), of the main French authors such as Proust, Chateaubriand, etc. The reader could edit two different or much more versions of the same literary text, underlying the upper text and the many under texts contained in the previous one, on the same screen, in a very visible, concrete and significant way. The visual and interactive aspects, designed with a user experience concern, will bring a new approach of reading, taking into account the new process of digital media and developing a more interactive thinking from the reader's point of view.

The natural place for rich annotated texts is the browser. The future website will work on a data base of literary texts, referenced and pined by a taxonomy of fields, with different kinds of libraries (a library of texts, a musical library and a library of pictures), interacting with each other depending on the edited text and the reading chosen by the user-reader of the website.

Tamas Biro

Eötvös Loránd University, Hungary

email:
biro.tamas@btk.elte.hu

PeregrMXL: Building and annotating a 17th century corpus of Medieval Christian Hebrew

Keywords: corpus building / Hebrew / poems / second-language production / theology

Our ongoing project “Hebrew Carmina Gratulatoria of the Hungarian Peregrines in the 17th Century” (K 125486, the National Research, Development and Innovation Fund of Hungary [NKFIH], 2017–2021) focuses on gratulatory poems written in Hebrew by Hungarian protestant students studying theology at Dutch and German universities (“peregrines”) in the seventeenth century. These poems – together with similar poems written in Latin, Greek, Syriac and other languages – were recited at public events, such as theses defences, and published in print subsequently.

Our project consists of collecting, analysing and publishing these poems. First, we bring together photocopies from 350-year-old rare books dispersed around the globe. Second, we transcribe them, normalize them and translate them. Third, we pose and attempt to answer research questions pertaining to the linguistic, literary, cultural and religious aspects of these compositions. The project aims at better understanding the linguistic skills and cultural backgrounds of the peregrines, their social network, the historical layers of the Hebrew language, the contemporaneous educational systems and theological debates. Moreover, we also look at examples relevant to general theories in literary, linguistic and cultural studies (cases of intertextuality, cases of L1 and L2 language transfer in foreign language production and so forth).

Consequently, building a serviceable corpus lies at the centre of the project, which could also serve as a sample for an unusual variety of the Hebrew language, viz. “Medieval Christian Hebrew”. The poster focuses on the challenges posed both by the input-side and by the output-side.

On the input-side, the transcription of the photocopies raises difficulties. The quality of the picture and the quality of the original edition might both be suboptimal, typos occur not infrequently, and the imperfect Hebrew skills of the authors might also have introduced textual problems. Since these poems were published only once – and with low circulation, at that – we have no access to second, corrected editions. Most texts appear in Hebrew script, with or without vocalization, but a few ones in transliteration with Latin characters. Hence, normalization is indispensable. Yet, normalization introduces arbitrary decisions, and it is unclear to what extent we should compensate for the authors’ lack of linguistic competence. It all depends on the research questions to be posed. At the same time, non-linguistic aspects of the texts – the typography of the poetic structures – are very clear. Therefore, we shall introduce PeregrXML, a markup language that allows for linguistic uncertainty on several levels (the original text, the normalized text and the interpreted text), while maintaining the poetic structures.

Finally, on the output-side, we would like our corpus to help answer an extremely broad and open range of research questions, within the current project, but also by ourselves and others in the future. The corpus will be annotated for a number of features, while other features will be introduced only on-demand, should a specific research require them. Finding the balance between excessive and insufficient investments into corpus building is a major topic for the last two years of our ongoing project.

Andrea Götz

Károli Gáspár University, Budapest, Hungary

email:
gotz.andrea@kre.hu

A corpus-based, individualised investigation of Hungarian interpreted discourse

Keywords: corpus-based interpreting studies / interpreted Hungarian discourse / discourse markers and connectives / filled and unfilled pauses / delivery speed

This poster presentation introduces a corpus-based analysis of the discourse characteristics of English to Hungarian simultaneously interpreted speech with a special reference to individual variation. Despite a wide range of research topics – frequency of connectives (DMCs) (Defrancq, Plevoets and Magnifico, 2015), ear-voice-span (Collard and Defrancq, 2019), hesitation (Plevoets and Defrancq, 2018), delivery speeds (Russo, 2018) – being investigated by the recent field of corpus-based interpreting studies, these properties are customarily studied in isolation and tend to be linked to sociocultural, rather than other linguistic phenomena, with the exception of a few complex studies dedicated to the interrelationships of multiple phenomena (e.g. Plevoets and Defrancq 2018). This means that the relationship between certain properties of interpreted discourse are not well known.

The corpus is drawn from the Hungarian Intermodal Corpus (Götz, 2017) which sources European Parliamentary speeches. The entire corpus is over three hours long and contains the speech production of 22 speakers, that is 11 female and 11 male interpreters. The discourse of the interpreters is investigated from three aspects: (1) delivery speed (words/minute), (2) filled and unfilled pauses (frequency and duration), (3) the frequency of discourses marker and connective items (DMCs). Delivery speed is known to differ in interpreted speech and vary with particular language pairs and has a profound impact. DMCs play a crucial role in maintaining and re-creating cohesion and can be used strategically by interpreters (Defrancq, Plevoets and Magnifico, 2015; Defrancq 2016). While hesitation phenomena, such as filled and unfilled pauses, and DMCs are closely linked components of speech (dis)fluency (Crible, 2018). Therefore it stands to reason that they should be studied in their wider discourse context. This

is especially relevant since previous investigations found an increase in the frequency of connectives in interpreting (Defrancq, Plevoets, Magnifico, 2015) but also a higher presence of hesitation and lower delivery speeds in the speech interpreters (Götz, 2018, 2019 in terms of Hungarian). How these phenomena interrelate, on the other hand, is not well understood despite possible connections between them.

This study therefore investigates how the aforementioned three phenomena impacts each other in the speech of Hungarian interpreters, and how individual variation influences this relationship. With regard to the frequency of DMCs, it can be hypothesised that (1) its frequency increases with delivery speed. This hypothesis would be supported by the observation that the number of connective items increases in interpreting in multiple language pairs (Defrancq, Plevoets, Magnifico, 2015). It can be further hypothesised the frequency of DMCs also responds to the amount and duration of filled and unfilled pauses. This relationship, however, has not been probed. Therefore it can be both proposed that (2) DMC frequency increases in response to a high level of hesitation in order to compensate for this, but equally, it can be expected that a (3) high level of hesitation in terms of pauses decreases DMC frequency due to the reduction in speaking time. This study explores these three possibilities through the interrelationships of DMC frequency, delivery speed, as well as filled and unfilled pauses.

Luca Guariento

University of Glasgow, United Kingdom

email: luca.guariento@gla.ac.uk

Curious Editors and Diners: two successful use-cases of collaborative Digital Humanities projects

Keyword: XML / digital humanities projects collaboration / digital editions

My poster illustrates how I successfully applied and deployed XML and related technologies to real-world digital editions. Specifically, I describe my work with the Curious Travellers project (Universities of Wales and Glasgow), and the Dined project (University of Oxford).

In both cases the end-result was highly satisfactory not only from a technical point of view, but also from an academic and collaborative perspective. A balanced use of the above-mentioned technologies and tools has proven to be the successful key to overcoming issues such as geographically-dispersed teams, people with different technical skills, accessibility, and different tagging needs.

Péter Horváth

Centre for Digital Humanities, Eötvös Loránd University, Budapest, Hungary

email:
horvathpeti99@gmail.com

The ELTE Poetry Corpus project

keywords: poetry corpus / automatic annotation / emtsv / TEI XML

The poster presents the project of ELTE Poetry Corpus. The main aim of the project is to create a database which contains the complete poems of numerous Hungarian poets until the first half of the 20th century. Besides the text of the poems, the corpus contains the metadata and the annotations of poems. We annotate the structural units, the grammatical properties of words and some poetic features. The grammatical properties are annotated by the NLP tool emtsv developed in the Hungarian Academy of Sciences (Váradi et al. 2018; Indig et al. 2019). The tool is used for the annotation of lemma, part of speech and morphological features. For the project, we have developed a Python program which annotates the rhyme patterns, the rhyme pairs, the rhythm, the alliterations and the main phonological features of words. The input of the annotation process is the document files from the Hungarian Electronic Library containing a great number of digitized Hungarian literary texts. The output is TEI XML containing the tags and attributes of the structural, grammatical and poetic features. We have run the whole annotation process on the complete poems of 44 Hungarian poets. Currently, we are checking the output to discover mistakes. We are also working on a query application of the corpus.

Philipp Hartl – Dominik Ramsauer – Thomas Fischer – Andreas Hilzenthaler – Thomas Schmidt – Christian Wolff

University of Regensburg, Germany

email:
Philipp1.Hartl@stud.uni-regensburg.de
Dominik.Ramsauer@stud.uni-regensburg.de
Thomas1.Fischer@stud.uni-regensburg.de
Andreas.Hilzenthaler@stud.uni-regensburg.de
thomas.schmidt@ur.de
christian.wolff@ur.de

Studying Web Culture: Distant Reading of Online Memes

Keywords: Memes / Internet Culture / Text Mining / Sentiment Analysis / Topic Modeling / Online Memes / Internet Memes

Memes are a popular part of today's online culture reflecting current developments in pop-culture, politics or sports. We present first results of an ongoing project about the study of online-memes via computational Distant Reading methods. We focus on the meme type of image macros. Image macros memes consist of a reusable image template with a top and/or bottom text and are the most common and popular meme types. We differentiate between the meme template, which is basically just the image of a meme and the meme derivatives, which are the multiple manifestations of a meme template differing in regards of the text of the meme.

Although memes are distributed and shared in large quantities, the majority of current research is of qualitative research, e.g. analyzing patterns and stylistic rules of a small number of memes (Shifman, 2012; 2014; Osterroth, 2015). Since memes typically have a textual component, we want to use computational methods of Distant Reading (Moretti, 2013) to analyze memes in a large-scale approach to gain insights about the language and the content of this specific internet phenomenon.

We gather a corpus for 16 of the most popular image macros memes by scraping the platform knowyourmeme.com thus creating a corpus consisting of 7840 memes derivatives and their corresponding metadata. Furthermore, we gather the text of the memes via OCR (*Google Cloud OCR*). We explore the application of various text mining methods like Topic Modeling and Sentiment and Emotion Analysis to analyze the language, the topics and the moods expressed via online memes. For all approaches, we have implemented various preprocessing steps commonly used in text mining (e.g. lemmatization). For topic modeling, we use *Latent Dirichlet*

Allocation (LDA, Blei et al., 2003) to calculate 16 LDA topics. For the sentiment and emotion analysis, we use lexicon-based approaches (Liu, 2012; Mohammad & Turney, 2013).

Some of our first findings are that most of the topics are expressions of a single meme template, which shows that some memes consist of homogenous and reoccurring word patterns. However, there are some topics with overlaps, expressing words common in multiple memes. One preliminary interesting finding concerning emotion analysis is that the “Ancient Alien” meme has the highest values for disgust and fear, which is a fitting result since those memes are often used in the context of conspiracy theories.

In future work, we want to continue our analysis by increasing our corpus, filtering out noise during the acquisition and using other Distant Reading visualization techniques.