6th Estonian Digital Humanities Conference

Data, humanities & language: tools & applications

September 26–28, 2018

Conference Centre of University of Tartu Library (W. Struve 1, Tartu)

PRESENTATION ABSTRACTS

Wednesday, 26.09

Keynote lecture: **Versification and Authorship Recognition**

Petr Plecháč (Institute of Czech Literature, Czech Academy of Sciences / Institute of Czech National Corpus, Charles University in Prague)

Contemporary stylometry has developed extremely accurate and sophisticated methods of authorship recognition. The logic behind them is to tell the author by measuring the degree of stylistic similarity between the text in question and particular texts written by candidate authors. Various style markers are being taken into account for this purpose: frequencies of words, frequencies of parts-of-speech, frequencies of character n-grams, frequencies of collocations… One important aspect of style (of one important form of literature) however seems to be completely disregarded – versification.

The talk will present the ongoing project focusing on whether characeristics such as frequencies of stress patterns, frequencies of rhyme types etc. may be useful in the process of authorship recognition. Some pilot experiments comparing various classification methods (Delta family, SVM, Random forest) and their evaluation with Czech, German, Spanish, and English poetry will be presented.

Collections of the National Library of Estonia as a source for analysis and mining of (textual) data

Jane Makke (National Library of Estonia)

National Library of Estonia is celebrating its 100th anniversary. During this period, it has produced and collected a significant volume of data consisting both of the metadata describing its collections as well as digital texts. There are two strong pillars on which the library collections rely on – physical and digital collections. The latter gets its input from three different channels – publishers' deposit of the born digital materials (including the web-domain .ee) and the pre-print files – all subject to the Legal Deposit Act enforced since 2017, and digitization of National Library's physical collections. The large variety and extent of the digital data enables the library to conceptualise itself as a resource for researchers.
This presentation will focus on the library data, and the perspectives and problems arising from the use of the library data as the source of analysis and mining of the textual data whether it is in the interest of the researcher or in the interest of the library itself to improve the production and use of the metadata.

Introducing 'Elias Lönnrot Letters Online'

Tarja-Liisa Luukkanen & Maria Niku (Finnish Literature Society)

Elias Lönnrot Letters Online (http://lonnrot.finlit.fi/omeka/) is the conlusion of several decades of work on the correspondence of Elias Lönnrot (1802–1884), doctor, philologist and creator of the national epic Kalevala. The correspondence consists of 2500 letters or drafts written by Lönnrot, 3500 letters received. So far, we have published over 2200 private letters sent by Lönnrot. The official letters, such as medical reports, will be added by the end of 2018. The final stage will be publishing the letters that Lönnrot received.

The letters are mainly in Finnish and Swedish, with some handfuls in German, Russian and Latin. The rich correspondence offers source material for research in biography, folklore studies and literary studies; for general history, medical history and the history of ideas; for the study of ego documents and networks; and for corpus linguistics and history of language. At the same time, the transcription work of the letters involves challenges. The documents themselves range from
carefully written official letters to barely readable, faded pencil drafts, and involve issues requiring special knowledge such as medical terms and symbols. In this way, the publishing project combines old-fashioned archival research with modern digital methods.

Elias Lönnrot Letters is built on the open-source publishing platform Omeka. Each letter is published with facsimile images and an XML/TEI5 file with metadata and transcription. A light form of TEI encoding is used. Lönnrot's own markings and unclear and indecipherable parts of text are encoded, place and personal names are not. One of the reasons is that the primary search tool offers easy access to the same information. This is a faceted search powered by Apache's Solr, which covers the metadata and transcriptions and allows limiting the results by five different categories.

The guiding principle of the Lönnrot edition is openness of data. All the data contained in the
edition is made openly available under CC-BY 4.0.

The XML files of all the letters are available for download, and the download feature is also integrated in the faceted search. A researcher can obtain XML files to study them with existing
linguistic tools such as those provided by the Language Bank of Finland. The raw data is available
for processing and modifying by researchers who develop digital humanities tools and methods to solve research questions.

For users who need data suitable for qualitative analysis tools like Atlas or do not need XML files, the transcriptions are made available for download as plain text.

Users can export the statistical data contained in the search results for processing and visualization with tools like Excel. The feature is useful in handling large masses of data, as statistical data can reveal aspects that would remain hidden when examining individual documents. For example, the statistical data readily reveals when and with

whom Lönnrot primarily discussed a given theme, while compiling such statistics manually would be a time-consuming process.

Hydra: Integrated Tagger-Lemmatiser with Deep Learning and Parallel Computing

Łukasz Gągała (University of Göttingen)

Rapid growth of electronic resources in Digital Humanities we observe today requires also a constant endeavour of tools development for members of the research community. In the field of historical computer linguistics we invariably face two diffculties related to reliable (pre-)processing, more precisely tagging and lemmatisation of premodern texts. The first problem is the nature of premodern spelling, both in Latin and in the vernacular languages of Europe, exposing a high degree of variance (between regions and even between particular scribes, as far as scribal culture is considered) that inevitably impedes the automatic text processing on scale. The second concern is the question of easily deployable algorithmic approach that would be able to deal with that abundance of historical dialects and scribal traditions exceeding the contemporary linguistic variety of Europe.
To address that matter we propose an integrated tagger-lemmatiser based on recurrent neural networks with parallel computation. Deep Learning as a fourishing domain of research has proofed that deep neural networks can effectively explore complexity of language (e.g. by tagging and lemmatising). However a character-based approach that would be very desired for non-standard spelling of premodern texts, both in Latin and in the vernaculars, requires a heavy model of neural networks. This in turn results in a very long computation time making the enterprise highly unpractical and unattractive. To improve a general neural-net approach for tagging and lemmatising we introduce a parallel computing technique (gossip protocol) to the training that aggregates multiple GPUs and shortens the total computation time. Additionally, we implement a novel recurrent-neural-net model (SRU) being much faster than state-of-theart solutions (LSTM, RNN).
The gain of time (by parallelisation and the SRU-algorithm) is used to improve upon tagging and lemmatising multi-label tokens and multi-token labels (Figure 1 and 2) that characterise premodern vernacular languages without consistent norms of orthography and punctuation. Moreover, due to an extended context window our approach is meant to catch words dependencies in languages with a more free word order (like Latin and Old Church Slavonic/Old Russian). Finally, our deep-learning approach can be benefciary for preprocessing of any language, as far as an annotated training corpus is provided (even for a pipeline process by building up a corpus, where new token forms may be mapped into already existing labels).

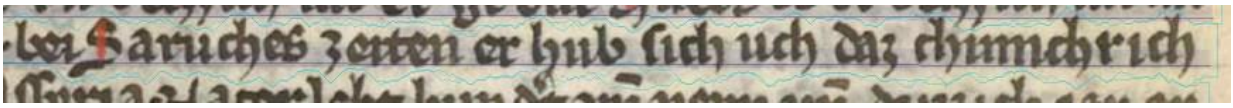

Figure 1. Multi-token single label problem. In this Middle High German sentence bei Saruches
zeiten er hub sich uch das chunichrich [in Assyria] [‘in the time of Serug there arose also the kingdom [of Assyria]’] the refective verb er hub sich [arose, emerged] is presented by two tokens in the diplomatic transcription widely used in medieval studies.

Figure 2. Multi-label single token problem. In this Middle Low German sentence dat hauestu ghedan [‘you did it/have done it’] the token hauestu is actually the enclitic form of hauest tu and needs two labels (part-of-speech tags and lemmas for both the verb hauest ‘(you) have’ and the personal pronoun tu ‘you’).

Selected Bibliography

Blot Michael, David Picard, Matthieu Cord, and Nicolas Thome. ‘Gossip training for deep learning’. NIPS 2016 workshop, Barcelona, Spain, December 2016
Kestemont, M., De Pauw, G., Van Nie, R. & Daelemans, W., ‘Lemmatisation for Variation-Rich Languages Using Deep Learning’. DSH – Digital Scholarship in the Humanities 32:4 (2017), 797815. DOI: https://doi.org/10.1093/llc/fqw034
Kestemont, M. & J. de Gussem, ‘Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning’, Journal of Data Mining & Digital Humanities (2017), pp. 17. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages, ed. M. Buechler and L. Mellerin.
Lei Tao, Yu Zhang, Yoav Artzi, ‘Training RNNs as Fast as CNNs’, arXiv:1709.02755
Piotrowski, Michael, Natural Language Processing for Historical Texts, San Rafael: Morgan & Claypool Publishers, 2012

Keynote lecture: **Something about the weather. Daily forecasts and the Dutch image of Europe.**

Joris van Eijnatten (Utrecht University)

Historical newspapers offer insight into 'collective mentalities'. Since these emerge through the iterative nature of information exchange, frequency counts gleaned from media involving a high degree of periodicity, such as newspapers, are one important means of outlining the contours of such a mentality. For domain experts (in my case historians), simple counts produce results that are as intuitively convincing as the complicated 'shock and awe' algorithms often vaunted in digital humanities projects. The research for this lecture focuses on twentieth-century Dutch newspapers, employing weather reports to trace popular conceptions of Europe as part of a collective mentality.

Contemporary tools for analyzing archaic variation: creating a corpus of 19th century Estonian communal court minute books

Gerth Jaanimäe[1], Liina Lindström[1], Kersti Lust[2], Kadri Muischnek[1], Siim Orasmaa[1], Maarja-Liisa Pilvik[1]

[1]University of Tartu, [2]The National Archives of Estonia

In this paper, we present a project dealing with the digitization and analysis of Estonian communal court minute books from the period between 1866 and 1890. The communal courts at that time were judicial institutions that tried peasants for their minor offenses and solved their civil disputes, claims, and family matters. The minute books, held in the National Archives of Estonia, are therefore a massive and rich historical resource shedding light on the everyday lives of the peasantry, and provide fascinating material also for linguists – the books contain regional varieties tied to specific genre and early time period, while also reflecting the writing traditions of the old spelling system.
We discuss the workflow and challenges of creating a digital resource from this kind of data with
the focus on automatic processing and analysis of the texts. We have structured the digitized texts using XML-markup, where body text is separated with tags referring to information about the titles, dates, indexes, participants, content and topical keywords. In order to provide input for network analysis, enhance the events described in the minute books with geospatial information, and enable queries with different degrees of specificity in the corpus, we have further analyzed the body text using tools for named entity recognition and morphological analysis from the EstNLTK library (Orasmaa et

al. 2016) in Python, which is developed for processing contemporary written standard Estonian.
The twofold (spelling-related and dialectal) variation in the communal court minute books yielded diverging results between the parishes when using contemporary tools for linguistic analyses. As Standard Estonian is based on the Northern dialects, the automatic morphological analysis and lemmatization also gave better results for texts from the Northern parishes where the analyser failed to recognize (i.e. provide any lemma for) 20% of the word forms, on average. The corresponding percentage for the Southern parishes was 33,5%. In comparison, the percentage of unknown words in present-day Standard Estonian is only 2%. At least one text file was manually checked from each parish in order to evaluate the accuracy of automatic morphological analysis. Human annotators provided analyses for unknown forms, corrected false analyses, and disambiguated forms with multiple analyses. Using custom dictionaries created based on these files, the analyzer was run again, and the results improved significantly. In the presentation, we introduce the main steps in creating and analyzing the collection of the communal court minute books and provide some scenarios on how to improve the quality of automatic text analysis.

References

Orasmaa, Siim, Timo Petmanson, Alexander Tkachenko, Sven Laur, Heiki-Jaan Kaalep (2016).
ESTNLTK - NLP Toolkit for Estonian. – Proceedings of LREC 2016, pp 2460–2466.

Integrating databases to study language history

Peeter Tinits (Tallinn University)

In the domain of linguistics, historical sociolinguistics has come to suggest that linguistic changes, and hence the shape of modern languages, can be better understood by investigating the structure of the past linguistic communities: who said what where to whom and when. A detailed look at the distribution of the variants can help us understand the mechanisms responsible some changes, and the historical pathways taken.
For example in the decades around 1900, multiple spelling variants became consolidated in their use in written Estonian. Historically, this may have been for example due to influential guidelines of central authority figures, due to gradual accommodation within the population, or simply a new generation picking one of the options. A detailed look at the data may help assess the comparative value for these different explanations.
In this presentation, I will describe a project that combines biographic, bibliographic and corpus data to study the diffusion of spelling norms in written Estonian around the

turn of the 19th century. I will describe the reasoning behind combining the different datasets, the workflow and effort that goes into linking them together, and some of the results from a quantitative analysis based on that.
The presentation aims to show how in combining different types of data, new insights can be made, and encourages to keep this in mind when compiling new digital humanities datasets. Easy integration and linking between different datasets can greatly facilitate the use of datasets both for their intended purposes as well as for ideas from other perspectives that may be unexpected at a time.

Suitable methods in network science for dealing with archaeological data

Maarja Olli (University of Tartu)

The aim of this paper is to discuss the suitable methods in network science for dealing with archaeological data. Namely the artefacts found from the cemeteries in southeast Estonia and east Latvia of Roman Iron Age (1st to 5th centuries AD). Were smaller groups present that interacted lively and shared similar material culture? Similarity between the cemeteries is studied using network science. Range of similarity measures are used (such as co-presence, Brainerd-Robinson similarity and $\chi 2$ distance) to study the similarity of grave furnishings, together with statistical sensitivity analyses in order to assess the impact of incomplete data because the nature of archaeological material is that it is fragmented. To see whether there is a pattern of the grouping of items, different methods are compared and on the bases of those results conclusions about the relationships are being made. This study gives a new insight into the relationships within southeast Estonia and east Latvia in the Roman Iron Age based on the similarities of certain grave goods. Also the most suitable measures for dealing with such data are outlined.

Parallel sessions: Conference hall

The impossible imitation: the reproduction of poetic texts via recurrent neural networks and the question of style

Artjom Šela (University of Tartu), **Boris Orekhov** (Higher School of Economics – Moscow)

Recurrent neural networks (RNNs) now are heavily deployed in text generation tasks in different environments (from poetry generation, machine translation to chat-bots and AI, on the principles see field-setting paper: Sutskever et al. 2011). It is often argued, that RNNs generative loop could somewhat faithfully reproduce features of the original

corpus, reveal hidden patterns or amplify certain underlying structures. For example, it was shown that character-level model trained on the corpus of Russian hexameter texts give in the output precise hexameter lines (Orekhov 2018).

Our paper will explore how poetic texts generated by the recurrent neural networks (RNNs) built with LSTM relate to their original corpuses in distribution of lexical features. The measure of model's "temperature" here will be the main variable that changes the type of RNNs output. Lower temperature texts exhibit "conservative" and "reliable" reproduction of patterns of original corpus and it could be compared to a human-like imitator: model amplifies the more visible features and present condensed and exaggerated design of original corpus (e.g. 2000 original word-types became 100 in the output). The higher temperatures (more "free" and "inventive") make output completely different: from the perspective of a reader it stops being coherent and has high amount of neologisms, the observed relatedness to original corpus fades. Yet, as we will show, if we use multivariate distance measurements based on 100 (200, 300...) most frequent words, the higher temperature output became computationally indistinguishable from the original. The high-temperature model reproduces the distribution of word-frequencies on the level of no human cognition is capable to imitate. The output of high-temperature RNNs could be seen as absolute forgery (our most precise authorship attribution tools cannot distinguish between texts), but at the same time it wouldn't never trick human reader as it is totally incomprehensible lexical chaos.

The results of stylometric inspection of RNNs' output could be useful to underline several aspects of the style and imitation problems. The high-temperature output highlights what long distance lies between our positivistic understanding of "style" of text and the "meaning" of it. As RNNs are not (yet) capable of reliable reproduction of long-term overarching connections to create "meaningful" texts, we still can use them as "machines of style", the pure imitation devices that take meaningful texts as input and return the different level of "style" carcasses in the output.

References

Orekhov, Boris (2018). Iskustvennye neironnye seti kak osobyi tip distant reading [Artificial neural networks as special type of distant reading]. In: Vestnik Priamurskogo gosudarstvennogo universiteta. 2. P. 32-43.

Sutskever Ilya, Martens James, Hinton Geoffrey (2011). Generating text with recurrent neural networks. In: ICML'11 Proceedings of the 28th International Conference on International Conference on Machine Learning. P. 1017-1024.

Analysing Reflection of Estonian holidays on Twitter

Siim Orasmaa & Dage Särg (University of Tartu)

People often use Twitter for communicating noteworthy events from their daily lives and for expressing opinions and emotions related to these events. Common holidays, such as national holidays and folk holidays give occasion to celebrations and elevated experiences. Several interesting studies have been done on the reflection of those in social media, most of which use tweets in English. For example, Wei Hu (2013) has analysed the sentiment of tweets published on Thanksgiving and Christmas, and Joshua H. West et al. (2012) have explored the temporal variability of tweets about drinking, identifying New Year's holiday as a hotspot. There are also works on other languages, e.g. Florian Kunneman and Antal van den Bosch (2015) have proposed a method for identifying periodic social events (including holidays) from Dutch tweets. However, to our knowledge, there is no previous research about holidays in Estonian tweets, nor are we aware of any natural language processing related research on Estonian tweets in general.

In this study, we collect a corpus of Estonian tweets and investigate how holidays are reflected in them. We use TweetCat (Ljubešić et al. 2014) to collect a corpus, and then process the corpus with EstNLTK (Orasmaa et al. 2016), adding linguistic annotations, and enabling lemma-based search and counting in the corpus.

To extract the holiday mentions, we compile a list of Estonian holiday names, covering the most common national holidays (such as Eesti Vabariigi aastapäev `Independence day') and folk holidays (such as Mardipäev, `Saint Martin's day') along with their respective dates. We investigate which holidays are mentioned most often, and how early or late the holidays are mentioned with respect to their actual dates. In addition, we analyse tweets published on holiday dates: which holidays are most popular for tweeting and does this popularity correlate with the popularity of holiday mentions. We also explore the trending keywords on the holidays and how the trending vocabulary in tweets published on a holiday compares to the vocabulary used when mentioning the holiday before or after the actual date.

References:

Hu, W. (2013). Real-Time Twitter Sentiment toward Thanksgiving and Christmas Holidays. In Social Networking, Vol. 2, pp. 77-86

Kunneman, F., & Van den Bosch, A. (2015). Automatically identifying periodic social events from Twitter. In Proceedings of the International Conference Recent Advances in Natural Language Processing (pp. 320-328).

Ljubešic, N., Fišer, D., & Erjavec, T. (2014). TweetCaT: a tool for building Twitter corpora of smaller languages. In Proceedings of LREC.

Orasmaa, S., Petmanson, T., Tkachenko, A., Laur, S., & Kaalep, H. J. (2016). EstNLTK - NLP Toolkit for Estonian. In Proceedings of LREC.

West, J. H., Hall, P. C., Prier, K., Hanson, C. L., Giraud-Carrier, C., Neeley, E. S., Barnes, M. D. (2012). Temporal variability of problem drinking on Twitter. In Open Journal of Preventive Medicine, Vol. 2 No. 1, pp. 43-48

Quantitative Formalism and Philosophical Research

Marek Debnár (Constantine the Philosopher University in Nitra, Slovakia)

The interactions between Natural Language Processing (NLP) and Digital Humanities (DH) have major effect on contemporary text analysis. The methods used in humanities (Linguistics, Literary Theory, Literary History, Philosophy etc.) that are used for the text analysis, undergo essential transformation caused by the application of digital technologies within all areas of humanities research. The paper aims to describe and analyze the methodology of quantitative formalism, better known under the umbrella term "distant reading" (a term coined by Franco Moretti currently used for various methods in IT processing of texts and bigger corpora) with regard to contemporary inquiries, and to define specific philosophical features of quantitative formalism in comparison to literary science and linguistics, from which these methods originated and where they had been applied. Nowadays, this approach, generally called DR2 - distant reading and data-driven research - is being explored in various projects at several leading centers for philosophical research. The paper seeks to answer question, whether it is possible for DR2 research to discover hidden or overseen premises and arguments in philosophical texts, what are the requirements and specific features of philosophical text corpora with regard to research objectives (level of annotation, etc.), what are the challenges with respect to development and usage of already accessible IT tools in philosophical research. The paper also shows usage of this theory on two case studies, first: interpretation of values in the epitaphs on the tombstones of an old Jewish cemetery in Bratislava and second: interpretation of the term freedom based on research of two corpora - corpus of religious texts and corpus of essays.

Parallel sessions: Tõstamaa seminar room

Gamification of history: explaining Soviet culture through digital formats

Alexandra Milyakina (University of Tartu)

Holistic approach to education implies that the knowledge is acquired through multiple languages and sources, which are becoming increasingly digital. However, system of school education is still reluctant to acknowledge anything but original texts and their official interpretation. According to Juri Lotman, the culture itself could be understood as a system of education (2000: 417). In line with the ideas of cultural semiotics, the project "History in screen" argues for a multi-faceted narrative, allowing a variety of perspectives and voices, in which the national culture is not portrayed as single and singular story.

“History on Screen” is a digital learning platform based on Leelo Tungal’s novel “Comrade Child” and its filmic adaptation. While literary adaptations have been for a long time used in literature class, history teachers lack a coherent framework for employing the format. Unlike more conservative history textbooks, our project aims not to give a solid version of the story but to map the versions and develop literacies for navigating between them. At first, film is seen as an illustration to a certain chapter of history — Stalinist era. On the next level, film becomes a starting point for discussing deeper questions, such as mediation of the past, the role of the perspective, and the construction of cultural memory.

Platform consists of two parts: interactive map of “Comrade Child” and three lessons. Interactive map includes 8 objects related to the Stalinist era: cult of personality, grassroot resistance, “vene värk”, etc. Each point on the map is illustrated with the brief video lecture by an expert, interactive tasks or games. For instance, students are asked to design their own propagandist poster or translate words from the Soviet newspeak. All materials are available in Estonian and Russian. By using the map, students are getting a deeper understanding of the Stalinist era and learn to work with different historical sources. Three lessons are focused on developing literacies needed for analysing memory and history. First lesson discusses the role of different texts in mediating the past, with special attention paid to artistic languages. Second introduces the notions of the perspective – how the history differs based on who and when is telling it. Third lesson focuses on the problem of the cultural memory and memory conflicts.

"History on Screen" is developed by a team of semioticians from the University of Tartu with the support of Õpiveski. Platform will be launched in September 2018. The presentation will be focus on the development of interactive tasks and using digital sources in history education.

Retrospective visualization of key landmarks of St. Petersburg based on a big corpora analysis

Antonina Puchkovskaia (ITMO University)

The project is aimed at developing interdisciplinary research involving history, librarian studies, cultural studies and information technologies, by creating an open-source-software-based web application containing historical and cultural heritage data on the key landmarks of St. Petersburg. This work contributes to the integration of digital intellectual technologies to systematize and popularize historical and cultural heritage sources through the processing of large amounts of data and the application of machine learning.

The goal of our project is to systematize unstructured humanities knowledge (big data) to create an educational database and web and mobile applications in order to provide the access to this data through its retrospective visualization on an interactive city map. Interactive applications will allow access to the database in online mode while making use of the user's geolocation to track nearby objects. Like in similar applications, it will be possible to get some audio and written information about a specific landmark.

Working on our project, we are analyzing sources and records. In our situation, sources are manuscripts that range from a single paragraph to a multivolume book. Records are source fragments and can range from a single record to hundreds of sections, pages, or paragraphs in a book. We are making a database schema that links people, occasions and dates based on primary sources. Because we use URIs for everything, we also group our data according to location. For the first three categories, we are specifying TEI elements encode groupings based on information about people (birth/death date, birth/death place, occupation, ethnicity, language), events (people involved, locations, dates, keywords), and relationships (family, professional).

To construct the ontology for our project, we are using Linked open Vocab (http://xmlns.com/foaf/spec/) to define the relationships between objects. It describes the FOAF language, creates a dictionary of named properties and determines classes using W3C's RDF technology. To define the relationships between objects on a map, we are using open linked vocabulary http://schema.org/LandmarksOrHistoricalBuildings. Regarding famous people connected to these places, DBpedia ontology http://dbpedia.org/ontology/ is being implemented. Finally, all objects are being mapped onto an interactive city map of St. Petersburg. A user-friendly interface facilitates easy navigation and permits filtering by different categories such as restaurants, music salons and apartments.

Database "Tartu in fiction" and the mobile application "TartuFic"

Reet Auksmann, Halliki Jürma, Katrin Raid, Ingrid Saare, Ülo Treikelder (Tartu Public Library)

The database "Tartu in Fiction" is based on the lovely idea of presenting Tartu by fiction. The database contains texts fragments by mostly Estonian writers who describe Tartu, its people, places, historical landmarks and everyday life as well as the spirit of Tartu and its characteristic phenomena such as students, cafes, jackdaws or slum architecture. The search possibilities include search by names, by book titles and keywords and the results are whole poems, fragments of prose or poetry and photos. The interface includes an interactive map section, linked with the topologic keyword system.

The aim of the project is to preserve our cultural history, to create relations between mental and physical space and to help students, researchers and local historians in finding and linking sources and data. A further goal is to introduce Estonian culture in other countries. First and foremost we aim to inspire people to read and discover.

The latest addition to the database is a mobile application (TartuFic) (Android and IOS platforms). The app includes most topological features of the main application and adds a new facility - the itineraries. The location of the user is displayed on the map, itineraries or routes are linked with the map application and for every point of route different text fragments and photos are displayed. The main purpose of the application is to present our cultural heritage to young people, offering the opportunity to learn the history of our culture in a more habitual environment.

Thursday, 27.09

Keynote lecture: **Detecting language change for the digital humanities; challenges and opportunities**

Nina Tahmasebi (University of Gothenburg)

For the last decade, automatic detection of word sense change has primarily focused on detecting the main changes in meaning of a word. Most current methods rely on new, powerful embedding technologies, but do not differentiate between different senses of a word, which is needed in many applications in the digital humanities. Of course, this radically reduces the complexity, but often fails to answer questions like: what changed and how, and when did the change occur?

In this talk, I will present methods for automatically detecting sense change from large amounts of diachronic data. I will focus on a study on a Historical Swedish Newspaper Corpus, the Kubhist dataset with digitized Swedish newspapers from 1749-1925. I will present our work with detecting and correcting OCR errors, normalizing spelling variations, and creating representations for individual words using a popular neural embedding method, namely Word2Vec.

Methods for creating (neural) word embeddings are the state-of-the-art in sense change detection, and many other areas of study, and mainly studied on English corpora where the size of the datasets are sufficiently large. I will discuss the limitations of such methods for this particular context; fairly small-sized data with a high error rate as is common in a historical context for most languages. In addition, I will discuss the particularities of text mining methods for digital humanities and what is needed to bridge the gap between computer science and the digital humanities.

Morning parallel sessions: Conference hall

Digital humanities first year in Tallinn University

Annika Loor, Kaisa Norak, Jaagup Kippar (Tallinn University)

The presentation is based on accumulated knowledge from the first year of Digital Humanities in Tallinn University. Speakers will introduce the course program, show examples of their works and speak about their experiences. For example, students will present their wordclouds, ggplot charts, interactive PowerBi dashboards, the

interactive map of Estonian feature film movie sets and Estonian film statistical data collection. In addition, speakers will comment on the courses and will attempt to give advice for the future.

7 years of crowdsourcing geodata for historic images – the experience of Ajapaik

Vahur Puik (Estonian Photographic Heritage Society (MTÜ Eesti Fotopärand))

Ajapaik.ee is a crowdsourcing platform for geotagging and rephotography of historic images.The aim is enriching digital heritage content with user generated metadata in order to make content accessible in a map based interface that is more ‘generous’ than the usual text based search. When the pictures have been pinned on the map the next task of rephotography can be performed that has huge educational potential as it helps wide audience to learn more about how places have changed over time, what kind of historic events have taken place in exactly what places etc. Rephotography is didactic both when exercised in person or when studying the repeat photos taken by other users.
Ajapaik has mostly Estonian content and user base, but it is open source and meant as a generic platform for crowdsourcing additional information to pictorial content. By July 2018 more than 74000 images have been geotagged by more than 8000 users and more than 200 users have submitted 10000 rephotographs to historic images.
The biggest issue for the platform is sustainability. It is run and owned by a non-profit organisation and developed with the help of public grant funding from Estonian funding bodies.The platform that was started during a hackathon in 2011 and had around 150000 user sessions (by more than 90000 users) in both 2016 and 2017 (surpassing the traffic to Estonian national central museums database) has accumulated considerable technical complexity and also technical debt.
In addition to geotagging and rephotography there are other tasks about historic images that could be solved in a crowdsourced manner and in combination with natural language processing and/or artificial intelligence and machine learning. For instance the dating information of images is also often very inaccurate and harvesting additional information is needed. Also different categorisation tasks, face detection and recognition, handwritten text recognition (for instance for the messages on historic postcards) are very relevant for pictures in the archives and museums.

Dorpat on the map: virtual city map "Saksa Tartu/Deutsches Dorpat" of (Baltic) German texts of Tartu

Reet Bender, Kadi Kähär-Peterson (University of Tartu)

(Baltic) German texts of Tartu form a collection of various texts from travelogues and fiction to memoirs and anecdotes, which provide vivid descriptions of numerous

themes and topics related to Tartu. The texts could interest the citizens and students of Tartu, as well as visitors of the city, including the Baltic Germans and their descendants, despite of that, due to the decline of German language skills in Estonia, these texts are accessible only for a small number of locals. In addition, even if the texts are readable, it is still difficult to spot the exact place or landmark described in the text. Therefore, the Department of German Studies of the University of Tartu and Tartu City Museum joined hands to create a virtual and bilingual literature map "Saksa Tartu/Deutsches Dorpat". The presentation gives an overview of the project, introduces questions and problems occurred during the process of connecting texts from 18[th] to 20[th] century with a modern city map, and comments the result of the work - more than 150 texts from 40 authors on a virtual map.

Literary Tallinn as a virtual landscape

Maarja Vaino (Tallinn Literary Centre)

Tallinn Literary Centre is responsible for administrating A. H. Tammsaare and Eduard Vilde museums and we take an active part in Tallinn literary life. Our goal and mission is to map the literary legacy of Tallinn (memorial tablets, monuments), but also to carry out research on writers that are essentially connected with Tallinn and to perpetuate the memory of these writers, as well as to dissert the image of Tallinn in the Estonian literature (but also in the foreign literature), to improve international connections, etc. Our dream is to make literature more visible in Tallinn city landscape and to ensure that the literature heritage of Tallinn is well preserved.
Literary map called www.kirjandusliktallinn.ee is a strong step to achieve these goals. Tallinn has been described in literature more often than we can imagine. Best known examples include the Mustamäe of Mati Unt and the Old Town of Jaan Kross and Indrek Hargla, but also Baltic German and Russian literature - for example in Sergei Dovlatov's or Werner Bergengruen's books. The purpose of our literary map is to locate the main actions (of the stories) that take place in the urban open space, as well as the writers homes and memory sites (monuments, museums etc). The presentation will give an overview of the map and gives an insight to our future developments.

Morning parallel sessions: Tõstamaa seminar room

Talking about negotiations: Estonian phrasal verbs as units of natural metalanguage

Haldur Õim (University of Tartu)

Background. Negotiation is a typical form of interaction where people discuss possible ways of reaching a common goal. And the same holds about describing negotiations: it is a common topic in everyday conversations, newspapers, reports of official meetings. Every language has special expressions to refer to structural components of negotiations – participants' communicative and reasoning acts, interests, attitudes. Since these expressions denote verbal acts occurring in the process of communication they can be treated as units of a natural metalanguage: they are themselves expressions of natural language, and second, they reflect culturally specific "folk" understanding of negotiation as a form of interaction.

Theoretical framework and data (see also Õim 2017). The work is based on the SOURCE – PATH – GOAL variant of the conceptual transfer schema COMMUNICATION IS MOTION. Here the PATH component is in focus because here the elements specific to NEGOTIATION come in (Jahansson Falk 2013). The central element here is OBSTACLE; in case of negotiations, typical OBSTACLES are different interests/ideas of participants concerning the ways of reaching the goal. Accordingly, the main type of activities in negotiation can be characterized as "dealing with OBSTACLES". Here the expressions originally denoting actions in the motion domain are regularly used in the domain of negotiation. An obstacle can be removed, eliminated (ignored), one can move around or over it etc. The critical question is: what kinds of actions can be used to deal with what kinds of obstacles. For instance, in Estonian, to reject a proposal (ettepanek), it is 'pushed back' (tagasi lükkama), but an assertion (väide) is 'pushed over, overturned' (ümber lükkama).

Research. In the paper I will give a summary of the results of the analysis and describe the means of formal representation of phrasal verbs of Estonian with the affixal adverbs edasi 'forward', tagasi 'back', kõrvale 'aside', üle 'over', ümber 'around' used to refer to acts in negotiations. For example: edasi lükkma('postone'), tagasi lükkama/võtma, ümber lükkama, kõrvale jätma/lükkama ('leave aside, ignore'). I will concentrate on conceptual representations of OBSTACLEs using the means of qualia theory (Pustejovsky, Jezek 2016): e. g what are the differences between PROPOSAL and ASSERTION that explain the use of different adverbs? **Applications**. As units of metalanguage, the described expressions can be used in applications such as automatic semantic tagging and mining of texts describing negotiations – be it reports of official meetings or literary works.

References

Johansson Falck, Marlene 2013. Narrow Paths, Difficult Roads, and Long Ways: Travel through Space and Metaphorical Meaning. – The construal of spatial meaning: windows into conceptual space. Ed. by Carita Paradis, Jean Hudson, Ulf Magnusson. Oxford : Oxford University Press, 2013, 214-235.
Õim, Haldur 2017. Natural Metalanguage for Describing Negotiations: Dealing with Obstacles (in Estonian). – Book of Abstracts. International Cognitive Linguistics Conference.10—14 July 2017, Tartu, p. 570.
Pustejovsky, James; Elizabetta Jezek 2016. Integrating Generative Lexicon and Lexical Semantic Resources.http://lrec2016.lrec-conf.org/media/filer_public/2016/05/10/tutorialmaterial_pustejovsky.pdf

Word2Vec implementation for Russian text

Olha Kaminska (University of Tartu)

In our time, the task of human language processing is one of the most important branch in the field of Machine Learning. Many tasks require working with text data, and a lot of information is stored in text format. But Machine Learning algorithms work with numbers, not letters, so developing a correct system for converting words to numbers is an important and relevant task.
In this work, the implementation of the algorithm for converting words into vectors, called Word2Vec, is considered. The algorithm is implemented for the Russian text, because, unlike English, for this language exist not so many optimal solutions in the field of Natural Language Processing. The implemented algorithm turns the given word into a 300-dimensional vector, such that the words close in meaning have close coordinates. In general, the described algorithm can be applied to other languages, which can provide enough input text data for processing and vocabulary building.
Work on this project can be divided on the next main steps: collect Russian text data, implement Word2Vec algorithm for Russian text, test obtained results. To check how well implementation works, it was used in movie reviews classification task. All words from train and test data were transformed in vectors, using implemented Word2Vec, then Logistic Regression model with default parameters was applied.

Digitalization of Slovenian Folklore Archive: Collection of Proverbs

Saša Babič (Estonian Literary Museum)

Digital humanities are well established and fast growing field in humanities. In Slovenia mostly linguistics has used the advantages of this field to create open databases and include analytical tools that can be used in collecting of linguistic material and working with data. On the other hand Slovenian folklore and ethnology studies seem to be far behind: only few collections are opened, mostly photography. Institute of Slovenian Ethnology has a big collection of folklore material - short forms of folklore, different tales and descriptions of rituals. One example of such collection is collection of proverbs, which still hasn't been digitized and opened to general public.
Paper will shortly present some more important ethnological and linguistic databases (included in Clarin) and then describe the collection of proverbs and its digitization (recognition, transcription, annotation), which is still in progress. The process of digitization will be presented in the context of logistic and ethic challenges, and compared with digitization of Slovenian riddles (http://www.folklore.ee/Slovenianriddles/), which was carried out on the Estonian Literary Museum within postdoctoral project Tradition and Innovation: Short forms of folklore and contemporary cultural Dialogues (MOBJD33) by Saša Babič.

The Correspondence between J. Barbarus and J. Semper as a Digital Language Resource in Korp

Olga Gerassimenko, Neeme Kahusk, Marin Laak, Tiina Saluvere, Kaarel Veskis & Kadri Vider

(University of Tartu, Center of Estonian Language Resources**,** Estonian Literary Museum)

Written communication between two well-known Estonian authors Johannes Semper and Johannes Barbarus lasted from 1911 to 1940. Both of them were members of outstanding literary circles of Estonia.
The letters and postcards sent by both parties are kept at Estonian Cultural History Archives. The collection consists of more than 600 letters/postcards sent during nearly 30 years, having total more than 240 thousand words.
Such complete collections of correspondence are quite rare, the fact that this collection is digitalised, adds more value to the resource. The process of publishing a commented book is in progress.
This kind of achived material has not only cultural value, but hides incredible linguistic importance as well. By using methods known from corpus linguistics, we can reveal hidden layers of everyday use of language. We are going to describe the process of

converting a digitalised collection into a language recource, that will help to research both linguistic phenomena and cultural context.
Korp is Språkbanken's corpus query tool that lets to view concordances and various statistics in many corpora. Korp is a frontend tool that uses much of IMS Open Corpus Workbench as backend. In order to have a digitalised resource in Korp, one should have text at least split into sentences and words. All kind of additional linguistic analysis is a bonus.
For Estonian, a highly inflected language, one of the most important annotation would be lemmatisation: information about dictionary form of each word. This would enable to search for words in their full morphological richness.
As for the Barbarus-Semper correspondence collection, we are going to describe the steps that were necessary to convert the digitalised collection into a morphologically analysed language corpus provided with essential metadata about the sources.

Afternoon parallel sessions: Conference hall

From history of emotions back to literary history: a case of Soviet realistic prose for children and young adults

Kirill Maslinsky (National Research University Higher School of Economics – Saint Petersburg)

Literary fiction has often served as a convenient source of data and inferences for the history of emotions. Yet the inverse is not true. The results of systematic inquiry into the historical changes of the emotions represented in the literature has hardly been considered as a source for literary history. In this paper, I undertake the experiment in order to draw inferences in this latter direction. The experiment builds upon the idea that natural language processing methods allow to scale the study of the emotional language use up to the level of a considerable corpus of fiction. The corpus of 400 works written in the realistic genre in Russian during 1920s—1980s and addressed to children and young adults is used as data. The objective is to test how the differences between authors in the emotional language use and the historical changes of the emotional language correspond to the conventional style and genre classification and periodization. Emotional language is operationalized by word embedding vectors constructed from the sets of seed words labeling emotions. Emotionally loaded behavior (crying, laughter), “basic” emotions (anger, fear, sorrow) and social-emotional concepts known to be prominent in Soviet culture (happiness, friendship) are included in the analysis. After measuring the share of each constructed emotional category in all of the corpus texts, the latent profile analysis is used to cluster the novels and the authors into larger “styles” employing similar emotional language. Some, but not all of the clusters correspond well to the conventional literary styles or genres, including

Socialist realism and school novel. Results also suggest that the most profound changes in the emotional description are associated with the demise of the socialist realism style after Stalin's death. The method allows to reconsider the conventional landmarks and boundaries in the history of Soviet children's literature, offering an alternative way to partition constant and variable features in the history of literary style.

Automatic linguistic annotation of spoken Pite Saami

Joshua Wilbur (Universität Freiburg)

Pite Saami is a critically endangered Uralic language spoken in northern Sweden, with currently around 40 speakers. Considering its size, the language has a surprisingly large documentation collection: this includes written texts spanning 130 years, and audio recordings from the 1930s through today. The vast majority of these texts are in the spoken-mode (here, I mean 'texts' in the broadest sense, i.e., including exclusively audio media). While this collection of written and spoken texts, taken as a whole, is not sufficient for big-data approaches, it is illustrative of the situation that smaller linguistic and cultural communities around the globe are in, and I hope to show how digital solutions are still quite relevant and important.

In my current project, linguistic annotation of these Pite Saami texts is a priority, ultimately aimed at enabling corpus-based investigations concerning the language's syntactic structures. Rather than annotating by hand, a digital infrastructure has been created to automate the annotation process, thus increasing efficiency and consistency; this presentation will focus on the processes involved and their output. In this, I will show the variety of available texts and how this variety is dealt with in preparing texts for automatic annotation. I will discuss the challenges in working with a language lacking an official orthography (i.e., without any standardized, character-based representation), as well as in dealing with the variation inherent to spoken language.

I will briefly present the language technology tools (a lexical database, a finite state transducer, and a constraint grammar framework) which are used to automatically perform lexical and morphological analyses, and to subsequently remove any resulting ambiguities. I will also provide examples of automatically annotated texts (including tags for morphological categories, part of speech, lexeme assignment and English glosses). In this, I will discuss how this approach can potentially be used in disciplines other than linguistics, such as for automatic thematic tagging relevant to anthropology, history and cultural studies, or others. Furthermore, I will indicate how such digital solutions can also be used to support revitalization efforts.

Digital Language Typology – mining from the surface to the core

Katri Hiovain, Juraj Šimko (University of Helsinki)

The Digital Language typology (DLT) project aims at developing a novel methodology of quantitative assessment of prosody on large spoken language materials, based on Continuous Wavelet Transform (CWT). We combine digital analysis methodology – using speech and language technology tools to process and analyze large data sets – with the framework of linguistic typology, which is about grouping of languages according to their characteristics. Although the field of language typology has succeeded in grouping languages by their differences and similarities in terms of morphology, syntax and phonology [e. g. 2], the analysis of the prosodic features (stress, intonation and durational patterns) can still be used to complement this research area. This contributes to achieving more comprehensive and multi-dimensional knowledge of language change at different aspects of language.
In our research, we focus on large sets of multilingual or dialectal spoken materials and their analysis, which requires automatized processing of the data. For analyzing prosodic features, we use the following procedure:
1) f0 and energy signals are decomposed to three hierarchical components using CWT [1], derivatives of the component signals are quantized to 3 levels (rising, falling and flat) and discretized resulting in finite state space.
2) Unigram models are trained for each language separately. The models capture statistical distributions of states depicting the instantaneous change of f0 and envelope at these three levels in parallel.
3) Perplexity between language models is used as a measure of mutual similarity.
4) The similarities among the languages are summarized using dendrograms.
Four different data sets have been analyzed. As a pilot experiment on several European languages showed promising results on grouping the languages [1], we have continued implementing our methods for investigating the prosodic features of different dialectal materials: North Sámi varieties [3], Swedish dialects [4] and Slavic language varieties.
In all cases, the analyses lead to meaningful groupings readily interpretable in terms of language contact and majority language influence. As the used method does not require any extensive labeling or annotating, it is especially suitable for endangered, under-resourced minority languages, such as the Sámi languages. Overall, our results indicate that prosodic characteristics of dialects can be more strongly susceptible to the majority language influence than other typological features.

[1] J. Šimko, A. Suni, K. Hiovain, and M. Vainio, “Comparing languages using hierarchical prosodic analysis,” in Proceedings of Interspeech 2017.
[2] M. Dryer, S. Matthew & M. Haspelmath (eds.) 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology.

[3] K. Hiovain, A. Suni, M. Vainio and J. Šimko “Mapping areal variation and majority language influence in North Sámi using hierarchical prosodic analysis” in Proceedings of Speech Prosody 2018.
[4] M. Włodarczak, J. Šimko, A. Suni and M. Vainio “Classification of Swedish dialects using a hierarchical prosodic analysis” Proceedings of Speech Prosody 2018

Afternoon parallel sessions: Tõstamaa seminar room

Motivation to engage in creative crowdsourcing: case of campaign "Recite Veidenbaums' Poetry!"

Jānis Daugavietis (University of Latvia)

This paper seeks to analyse the motivation of participants engaged in creative crowdsourcing campaign carried out by the Institute of Literature, Folklore and Art (University of Latvia) in 2017 in which people were asked to recite, record and submit the poems of Latvian poet Veidenbaums. Afterwards, using online questionnaire, `crowdsourcers' were surveyed for their socio-demographic characteristics, motivation to participate and evaluation of the different aspects of the campaign from technical ones to personal ones. Additionally, data available from `Google Analytics' were employed to analyse the process of campaign and various aspects of participation.
From September to December 2017 as a part of 150th anniversary of a Latvian poet Eduards Veidenbaums' Latvian society was invited to virtually participate in campaign by reading one or more of his poems aloud and to record it. Using specially programmed crowdsourcing tool recording was made possible by using everyone's computer or mobile phone and/or uploading it to the site http://lasi.literatura.lv. Another opportunity to participate in campaign was provided by the specially built `Veidenbaums Studio' a mini-recording studio situated at The Latvian National Library.
A conceptual (and also political) problem is fostering accessible digital participation. Within the project one of our aims is to promote social inclusion and support identity-building and value-defining processes of the society, providing a participatory digital environment. Our preliminary analysis displays unequal participation of different social groups in Veidenbaum's campaign (2018 Eglāja-Kristsone, Daugavietis). Overrepresented groups were women, ethnic Latvians, highly educated, students, urban residents; underrepresented - men, Russian speaking community, diaspora, rural residents, manual workers, non-humanitarians, elders.
What were the motives behind engagement or non-engagement? This analysis of user engagement is based on Heather L. O'Brien's model (2008, 2010, 2011 etc.). She distinguishes two basic needs, which makes people to engage in crowdsourcing: utilitarian and hedonic; and four distinct stages: point of engagement, period of sustained engagement, disengagement, and re-engagement (2008). O'Brien is still

improving and modifying her model, and in this paper the model is tested on particular creative crowdsourcing case.

Application of L. Vygotsky and transmediality in the contemporary cultural education

Aleksandr Fadeev (University of Tartu)

In the following article and presentation I'm going to speak about the practical use of digital transmedia technology at a contemporary cultural education from the semiotics point of view. Transmedia is a contemporary semiotical tool that must be used in educational practice to stimulate the meaning making while reading a cultural text. Transmediality is getting its popularity in the cultural field of education curricula and contemporary digital education. One of the reasons for which is our transmedia reality which always forces us to use various media tools. Being involved into the transmedia surrounding it is difficult to receive the necessary diversity of reflection of a particular cultural texts by using one media as an educative and narrative tool, especially if the signified objects are meaningless or unfamiliar for the reader. Thus transmediality can be successfully used for the needs of acquiring own culture and its semiotical background. At the same time in the paper I consider Lev Vygotsky's contribution into contemporary pedagogical framework and the developmental theory from the semiotical and transmedia points of view. The Zone of Proximal Development which can be mediated by the use of various media can be significantly widened by the use of transmedia, digital tools and assisting online resources which at the same time change the role of the teacher in the pedagogical paradigm.

In the experimental part of my paper I am going to talk about our project with the Transmedia Group of the Department of Semiotics at the University of Tartu where we develop educational courses which focus on transmedia technology. Transmedia storytelling helps students translate the story into different cultural languages by using other media which helps them generate more meaning from the narrative. Transmedia narration is another way of reading a text of culture, which includes the necessity to translate the text among different languages, using different media. It doesn't only stimulate the educational interest in the classroom but develops the semiotical apparatus and communication skills.

Creativity in non-digital age: How to collect, analyze and present data in non-digitalized environment(s)

Aleksandar Takovski (South East European University)

Among many differences, unequal pace of general digitalization and especially its application in scientific research is one of the greatest challenges that many societies and research communities are faced with. Macedonia is one of the countries marked by low level of digitalization, which unavoidably affects researchers (deprived of the opportunity to acquire and use digital technologies and approaches). The problems most commonly faced are lack of personal engagement, lack of training opportunities, lack of institutional support (financial and logistic), lack of technical knowledge (e.g. scripting knowledge), and last but not least the lack of digital software adapted to the Cyrillic orthography of Macedonian language. Despite these challenges, some researchers in Macedonia undertake efforts to produce high quality research, unfortunately with little or no reliance on digital approaches to data analysis. This study seeks not only to acknowledge their quixotic efforts, but by identifying and analyzing their challenges, it seeks to offer them help to overcome them.

By relying on personal experiences of researchers form different fields gained through a series of personal narratives and semi-structured interviews, this study tends to: a) identify challenges in data analysis in regard to the use of digital tools, b) present different approaches to textual data analysis undertaken by researchers, c) discuss and suggest potential improvements, especially relevant to non-digitalized environments.

Beside the research objectives, this project additionally seeks to help Macedonian researchers by bridging the gap in the use of digital approaches and technologies in data processing between them and researchers with more advanced skills in digital methodologies. In this respect, the conference is an excellent venue for networking and knowledge transfer that will give rise to future actions aimed at enhancing Macedonian research knowledge and skills in the realm of digital methodologies of data analysis.

Keynote lecture: **Fake it till they believe it? A quest for authenticity on social meedia**

Andra Siibak (University of Tartu)

In this technology saturated society the members of older and younger generations alike struggle to find the right balance between the advantages of constant connectivity and the fact that this very connectivity inevitably gives others access to information that was previously considered private. Due to the context collapse on networked publics, social media users are experiencing a never-ending tension between imagined audiences, actual receiving audiences and invisible audiences.

The personal publics that we have created to ourselves on social media usually challenge the users to maintain equilibrium between a contextual social norm of personal authenticity that encourages information-sharing amongst your ideal audience (peers, close friends), with a need to conceal some information from the eyes of the nightmare readers (e.g. parents, teachers, employers). User's discrepancies between one's imagined and actual online audience, however, may lead to unwanted consequences and may thus jeopardize the physical, mental or social well-being of a person.

In the current talk I will make use of the findings from different qualitative case-studies to illustrate how the imagined audiences on social media interpret and reflect upon the online identity constructions they have come across on social media. Focus groups with young followers of micro-celebrities (N= 51), and students lurking on their teachers on social media (N=43); as well as interviews with employers carrying out online background checks on job applicants (N=30) will be used to explore what authenticity means in the context of online impression management.