RESEARCH DATA AND HUMANITIES – RDHUM 2019

UNIVERSITY OF OULU, AUGUST 14–16, 2019

BOOK OF ABSTRACTS

Organizers:
Jarmo H. Jantunen (chair)
Sisko Brunni
Maria Frick
Niina Kunnas
Mietta Lennes
Santeri Palviainen
Valtteri Skantsi
Hanna Westerlund
Katja Västi

Plenary speakers

Anna Čermáková (University of Birmingham)

Wicked witches and wise wizards: Children's literature as an interdisciplinary meeting point in digital research

Gender is one of the fundamental structuring principles of our society. Gender is a social construction and as such is changing over time. The change is visible through various social practices and legislation. Reflections of how we conceive of gender and these developments are manifested in the discourse. Essentially, gender construction is reproduced and negotiated through language. Children's literature, is not just "literature" directed at child audiences, it also presents an important formative discourse. In this talk, I am going to investigate the gendered social structure of the 19th century and contemporary children's literature. I will show how a corpus linguistic approach makes it possible to identify different layers of society and how characteristics of fictional social structures that may include the likes of *wicked witches* and *wise wizards* are shared across children's books. There are two major data sources I use. For the analysis of the 19th century I use ChiLit, the 19th Century Children's Literature corpus (4.4 million words, available from CLiC). For the contemporary data, I use a bigger (12.9 million words) corpus of contemporary children's literature drawn from texts published after 2000 by the Oxford University Press (*Oxford Children's Corpus* (OCC). The findings from these two data-sets will be further contextualised by other resources. This paper addresses questions relevant for discourse analysis, literary analysis, stylistics, gender and childhood studies but also social history. It aims to situate corpus linguistics across these fields and within the digital humanities more widely. I will argue in favour of incorporating a qualitative dimension through digital 'close reading' supported by state-of-the-art tools like CLiC.

Arja Kuula-Luumi (Tampereen yliopisto)

Tietosuoja ihmistieteissä

Tietosuojamuutokset ovat näkyneet monella tavalla eri palvelujen käytössä, kun Euroopassa alettiin soveltaa EU:n tietosuoja-asetusta keväällä 2018. Tietosuoja-asetukseen tutustumisen rinnalle saimme lisää opiskeltavaa 1.1.2019, kun asetusta täydentävä ja täsmentävä kansallinen tietosuojalaki astui voimaan. Molemmat vaikuttavat oleellisesti myös tutkimuksiin, joihin sisältyy henkilötietojen käsittelyä.

Esitykseni sisältää keskeisimmät asiat tietosuojasäädösten soveltamisesta henkilötietoja sisältävien tutkimusaineistojen käsittelyyn. Niitä ovat esimerkiksi tutkimuksen rekisterinpitäjän määrittäminen, henkilötietojen käsittelyperusteen valinta ja tutkittavien informoiminen henkilötietojen käsittelystä. Lisäksi selitän tietosuoja-asetuksen mukaisia rekisteröidyn (tutkittavan) oikeuksia ja kerron, miten oikeuksia voi rajoittaa kansallisen tietosuojalain perusteella. Painopisteeni on suoraan tutkittavilta kerättävissä aineistoissa, mutta esityksen lopussa kerron lyhyesti myös tietosuoja-asetuksen soveltamisesta some-datan käyttöön tutkimuksessa. Kuvaan muuttuneen tietosuojan säädösympäristön ottamalla esitykseen mukaan konkretisoivia esimerkkejä.

Veronika Laippala (University of Turku)

From bits and numbers to explanations – doing research on Internet-based big data

Internet is a constantly growing source of information that has already brought dramatic changes and possibilities to science. For instance, thanks to the billions of words available online, the quality of many natural language processing (NLP) systems, such as machine translation, has improved tremendously, and

people's beliefs, cultural changes and entire nations' mindscapes can be explored on an unprecedented scale (see Tiedemann et al. 2016; Koplenig 2017; Lagus et al. 2018). Importantly, almost anyone can write on the Internet. Therefore, the web provides access to languages, language users, and communication settings that otherwise could not be studied (see Biber and Egbert 2018).

Paradoxically, the Internet's extreme size and diversity also complicate its use as research data. Many Internet-based language resources, such as English Corpus of Global Web-Based English (GloWbE) or the web-crawled Finnish Internet Parsebank developed by our research group, are composed of billions of words. Already searching from these databases requires specific tools, but especially the analysis of the search results may not be straightforward. For instance, the Finnish word *köyhä* 'poor' has 209 609 occurrences in the Finnish Parsebank, and its English correspondant has 312 974 hits in GloWbE. These language resources provide easily bits and numbers, but how to explain them?

In my talk, I will present some of the work we have done in our research group in order to bend Internet-based data collections for research questions in the humanities, where numeric results on frequencies are just the beginning of the analysis. In particular, I will discuss our newly-launched project on improving the usability of Internet-based big data, *A piece of news, an opinion or something else? Different texts and their automatic detection from the multilingual Internet*. In the project, the ultimate objective is to develop a system that could automatically detect different text varieties, or *registers* (Biber 1988), such as user manuals, news, and encyclopedia articles, from online data. Currently, for instance a Google search can return an overwhelming number of documents from mostly unknown origins and similarly, the origins of the documents in the web-crawled big data language collections are typically unknown. However, in order to explain research results gotten from these collections, information on the kinds of texts included in the data would be very useful if not mandatory.

Identifying registers from the Internet involves a number of challenges. An essential prerequisite would be information on the registers to be detected. But what kinds of texts is the Internet composed of? A second concern, then, is that online texts do not follow the traditional print media boundaries (see Biber and Egbert 2018). For example, how can one distinguish texts that neutrally report scientific findings from those that use the information to persuade the reader? Additionally, text classification is typically based on manually labeled example documents representing the categories to be detected. However, developing this *training data* is very time-consuming and needs to be done separately for each language. Would it be possible to detect registers without all this manual work?

References:

Biber, D. 1988. *Variation across speech and writing*. Cambridge University Press. Cambridge.
Biber, D. and J. Egbert 2018. *Register variation online*. Cambridge University Press. Cambridge.
Koplenig. A. 2017. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities*, 32(1), 169–188.
Lagus, K., M. Pantzar, and M. Ruckenstein 2018. Kansallisen tunnemaiseman rakentuminen: Pelon ja ilon rytmit verkkokeskusteluissa. *Kulutustutkimus*. Nyt 1-2/2018.
Tiedemann, J., F. Cap, J. Kanerva, F. Ginter, S. Stymne, R. Östling, and M. Weller-Di Marco 2016. Phrase-based SMT for finnish with more data, better models and alternative alignment and translation tools. *Proceedings of the First Conference on Machine Translation*, Volume 2: Shared Task Papers, 391–398. Berlin, Germany. Association for Computational Linguistics.

Workshops

Kansalliskirjaston data haltuun

Kansalliskirjasto on aktiivinen ja tunnettu toimija digitaalisten ihmistieteiden kentällä. Kirjasto tarjoaa monipuolisia data-aineistoja, asettaa niitä tutkimuksen käyttöön ja on osaava tutkimuksen kumppani. Kansalliskirjaston käyttöön asettamista digitoiduista aineistoista ja niistä tuotetuista datapaketeista hyötyvät digitaalisten ihmistieteiden tutkijat eri aloilla.

Digitoinnit, siitä syntyvä data ja datan käsittelyyn tarvittavat välineet antavat datalähtöiselle tutkimukselle uusia mahdollisuuksia, joiden kartoittaminen ja tutkijan odotusten havaitseminen ovat työpajojen tavoitteina. Työpajoissa tarjotaan mahdollisuus keskustella digitaalisten aineistojen käytöstä ja niihin liittyvistä odotuksista Kansalliskirjaston asiantuntijoiden opastuksella.

Työpajoja edeltää Kansalliskirjaston sähköisiä aineistoja esittelevä puheenvuoro ja kolme lyhyttä demoa, joissa esitellään, miten Kansalliskirjaston tuottamaa dataa on hyödynnetty tutkimuksessa. Demoesityksissä käydään läpi 1) Kansalliskirjaston laajimman verkkokokoelman, digi.kansalliskirjasto.fi:n, tarjoamia aineistoja digitaalisten ihmistieteiden tutkimukselle ja 2) uralilaisten kielten verkkokokoelman Fenno-Ugrican käyttömahdollisuuksia ja 3) tutustutaan linkitetyn Fennican sekä ontologiapalvelu Finton sisältöihin.

Työpajoissa tunnistetaan Kansalliskirjaston tarjoaman avoimen datan mahdollisuuksia ja pyritään pohtimaan, miten omaa tutkimusaihetta voisi lähestyä Kansalliskirjaston tuottaman datan avulla.
työpajan otsikko: Kansalliskirjaston data haltuun

työpajan aihe ja tavoite: Tavoitteena on antaa yleiskuvaus Kansalliskirjaston tuottamista aineistoista ja tarjota muutamia käytännönläheisiä esimerkkejä aineistoista tuotetun datan hyödyntämisestä tutkimuksessa sekä haastaa tutkija pohtimaan omaa tutkimusaihettaan suhteessa Kansalliskirjaston aineistoihin.

työpajan kielet: suomi

työpajan järjestäjät: Tuula Pääkkönen, Juha Rautiainen & Jussi-Pekka Hakkarainen

työpajan työskentelytapa: 1) esittelevä yleiskatsaus Kansalliskirjaston tarjoamiin aineistoihin ja dataan 2) kohdennetut demot x 3 (Digi, Fenno-Ugrica, Finto ja Fennica) ja 3) työpajat pienryhmissä

Presenting text similarities by Multidimensional Scaling

In texts, we can measure various properties that help to reveal language use patterns, e.g. the proportion of specific phonemes, words, parts of speech and n-grams. Multidimensional Scaling (MDS) is a group of algorithms for handling such data with a large amount of properties, more specifically for visualizing the level of similarity between objects (in our case, individual texts) based on several parameters. The R and Python programming languages have libraries for MDS, specific modifications for the algorithms can also be written directly in a programming language.

MDS can be used in comparative text analysis to define the variables that best distinguish, for example, texts of different genres (e.g. journalistic and fiction texts), texts written by native speakers and second language learners, texts written by language learners with a different proficiency level or first language, and fiction texts written by different authors. An interactive and entertaining possibility that MDS offers for educational purposes is that students can compare their writings with each other and with diverse marker

texts, such as fiction by well-known authors, newspaper articles of radio broadcasts, so that the distances between texts are shown in scatterplots.

In the workshop, we will show how to use MDS for comparing different texts in the R environment for statistical computing and graphics, how parameter weights can be changed and how they affect the similarities of texts. The duration of the workshop is about 90 minutes. The main language is English, although Finnish can additionally be used for giving instructions, asking and answering questions.

Everyone interested in statistical text analysis is welcome to participate. Former experience of using R is not required, however, the software environment should be downloaded prior to the workshop.

The organizers of the workshop:
Jaagup Kippar
Annika Loor
Kaisa Norak

Tools and services in the Language Bank of Finland

Topic and goals
This tutorial aims to provide the audience with an overview of the Language Bank of Finland, i.e., the collection of corpora, tools and services provided by FIN-CLARIN. The participants will also have a chance to try out some of the tools and new features during a hands-on session. The tutorial is intended for anyone interested in digital research methods.

In addition to versatile corpus search, the Language Bank offers tools for processing and analyzing various types of data that contain text or speech. Many of these tools can be used via an online platform called Mylly. Mylly allows users to upload their own datasets to their personal workspaces and to process, analyze and visualize their data without having to manually type complicated commands. Mylly also keeps track of the user's workflow automatically. Mylly is an open platform where it is possible to add new tools on request.

Further information about the Language Bank of Finland can be found at https://www.kielipankki.fi.

Mode of organization and program design: The tutorial will begin with a general overview of the Language Bank of Finland and continue with a number of demonstrations and examples presented by FIN-CLARIN staff members. After the demonstrations, a hands-on session will be arranged for those participants who are interested in taking a guided tour of the tools. During the hands-on session, it will also be possible to ask more specific technical questions concerning the tools.

Language of the tutorial: English; however, the hands-on session can be flexible between Finnish and English

Organizers and contact persons:
Krister Lindén, FIN-CLARIN / University of Helsinki (Chair)
Mietta Lennes, FIN-CLARIN / University of Helsinki

Challenges and Developments in Preserving and Publishing of Large Audio/Video Data

Introduction
Over the past decades archiving and publishing of datasets, corpora and collections of audiovisual research data has become standard procedure in many academic fields. Building scalable infrastructure and workflows that consider requirements by the scientific target group (e.g. FAIR data principles) has been a challenge for data centres and research infrastructures. One of the first pioneers in this area was The Language Archive Tools (LAT) publishing platform. It was developed at the Max Planck Institute for Psycholinguistics in Nijmegen with the language research community in mind. LAT is used all over the world, also in the locations of the workshop organisers in Cologne (CLARIN-D), Lund (SWE-CLARIN) and Helsinki (FIN-CLARIN). Unfortunately it is no longer maintained and needs to be replaced.

A feasible replacement is open source repository software like Fedora Commons. Although they are geared towards standards in research data management and provide key functionalities they require a great deal of customizing and specialized staff for maintenance. Presently there is no turn-key solution for specialized research data. Repository software does not always integrate well into existing infrastructure and established processes. The University of Cologne has developed a more lightweight approach that builds on basic services at the local computing centre but still fully adheres to standards in research data management.

Topic and goals of the workshop/tutorial
The workshop has three aims:
1. Presentation of the LAT based audiovisual archives in the three CLARIN locations.
2. Discussion of present and envisioned needs of the respective CLARIN centres from a broad angle, for example scientific focus but also resourcing requirements.
3. Exploring the storage solutions and workflows developed at the respective centres to replace their LAT instances.

One focus of the workshop will be the solution developed at the University of Cologne (KA3) for storage, search and publication of audiovisual data. KA3 offers a clear path from possibly large archive copies of data to derived versions for use over the web. It implements the IIIF audio/video API and makes it possible to store original data in the highest possible resolution and access it in compressed formats more suitable for web based applications.

The KA3 frontend and backend are intended to be published as open source.
The technology will be tested at the Language Bank of Finland in Spring 2019, experiences of this pilot will be discussed and mirrored against the requirements in Lund.

Topic coverage
The workshop covers the following desired topics with a focus on the topics in bold.
-Compiling digital databases and infrastructures
-Digital data for less commonly used languages
-Digital data and research as pedagogical resources
-Search engines for digital data
-Annotating digital data
-Future technologies and innovations, e.g. virtual research environments

Intended audience:
-Infrastructure providers that already now provide audiovisual archives like MPI's LAT
-Infrastructure providers that are interested in setting up an audiovisual archive.
-Data providers and researchers that would like to get a deeper understanding about the challenges and solutions for storage and provision of large datasets.

Workshop/tutorial organizers and contact persons:
Jonathan Blumtritt (University of Cologne (Chair)), Felix Rau, (University of Cologne), Jens Larsson (Lund University Humanities Lab), Martin Matthiesen (CSC / The Language Bank of Finland)

Language of the workshop: English

Mode of organization and program design:
-Presenting the archives, with a focus on audiovisual data (Lund University Humanities Lab, Language Archive Cologne, Language Bank of Finland)
-Mode of organization and program design (presentation of the KA3 LAT replacement, presentation of the Language Bank Pilot, discussion)

Tieteen termipankki opiskelijan, opettajan ja tutkijan työvälineenä

Helsingin yliopistossa kehitettävä ja ylläpidettävä Tieteen kansallinen termipankki (http://tieteentermipankki.fi) rakentaa kaikkien Suomessa harjoitettavien tieteenalojen yhteistä, avointa ja jatkuvasti päivitettävää termitietokantaa tiedeyhteisön ja kansalaisten käyttöön. Tutoriaalin tavoitteena on esitellä Tieteen termipankin lähtökohtia, sisältöä, toimintoja ja käyttöä eri näkökulmista. Tutoriaalissa opastetaan ja ohjeistetaan termipankin hyödyntämistä opiskelun, opetuksen ja tutkimuksen tukena sekä sisällön tuottamista termipankin asiantuntijaryhmän jäsenenä.

Termipankki on semanttinen mediawikialusta, jossa eri tieteenalojen asiantuntijat julkaisevat tietoa alansa käsitteistä ja erikoissanastosta: termien suomenkielisiä nimityksiä ja käännösvastineita, määritelmiä ja selityksiä, havainnekuvia ja linkkejä tekstiesimerkkeihin. Termityötä tehdään talkoistamalla. Samalla tarjoutuu mahdollisuus käydä monitieteistä keskustelua käsitteenmuodostuksesta. Keskusteluun voivat osallistua kaikki termipankkiin omalla nimellään rekisteröityneet käyttäjät.

Termipankkiin päivitetään jatkuvasti tieteenalojen ajantasaista termistöä, mikä on elävän ja kehittyvän tieteen kielen perusedellytys. Termityö vakiinnuttaa käsitteitä luomalla yhteisesti sovittuja suosituksia siitä, mitä tieteellisillä termeillä tarkoitetaan. Yhdenmukainen termistö helpottaa ja selkeyttää viestintää sekä vähentää väärinkäsityksiä niin asiantuntijoiden kesken kuin asiantuntijoiden ja maallikoidenkin välillä. Termipankkiin kootut termit auttavat ymmärtämään tiettyä tieteenalaa ja sen piirissä tehtävää tutkimusta, mutta myös tieteenalojen välisiä yhteyksiä.

Tutoriaalin kielenä on suomi ja kohderyhmänä ovat perustutkinto- ja jatko-opiskelijat, opettajat sekä tutkijat. Tutoriaali sisältää termipankin sisältöä ja alustaa esittelevän johdantoluennon demonstraatioineen, lyhyen johdatuksen termityöhön ja käsiteanalyysiin sekä osallistujien toiveiden ja tarpeiden mukaan räätälöitäviä osioita. Tutoriaalin osallistujien toivotaan luovan termipankkiin etukäteen käyttäjätunnuksen (muotoa: Etunimi Sukunimi), minkä jälkeen jatko-opiskelijat ja tutkijat voivat halutessaan pyytää pääsyä termipankissa jo toimiviin asiantuntijaryhmiin ja kokeilla esimerkiksi käsitesivujen luomista ja muokkaamista termipankissa. Osallistujat voivat tuoda tutoriaaliin myös oman tieteenalansa käsitteistöön liittyviä ongelmatapauksia ja erityiskysymyksiä yhdessä ratkottaviksi.

Tutoriaalin järjestäjä:
Johanna Enqvist (Tieteen termipankki, Helsingin yliopisto)

General session

Tuuli Ahonen (University of Eastern Finland)

Audiovisuaalisten tekstien multimodaalisuus

Tutkin väitöskirjassani audiovisuaalisten tekstien, kuten elokuvien ja televisiosarjojen multimodaalista luonnetta sekä sitä, miten tämä piirre vaikuttaa kyseisten ohjelmien kääntämiseen eli tekstittämiseen. Tekstitykset ovat tekstikenttiä elokuvan taikka televisio-ohjelman kuvan alareunassa. Näihin kenttiin kääntäjän pitää mahduttaa se, mitä hahmot ohjelmassa viestivät kielellisesti toisilleen. Usein hahmot puhuvat nopeasti, eikä kaikkea saada mahtumaan tekstitysriveille. Tekstittäjän tehtävänä onkin tiivistää viesti helposti luettavaan muotoon, jossa kuitenkin säilyy viestin ydin. Tässä tehtävässä audiovisuaalisen tekstin multimodaalisuus voi auttaa. Tekstittäjä voi ottaa kuvan ja äänen huomioon kääntäessä ja tiivistäessään puhuttua viestiä näitä moodeja hyväksi käyttäen. Tutkimukseni tarkoituksena on selvittää, miten audiovisuaalisten ohjelmien kääntäjät eli tekstittäjät ottavat kuvan ja äänen huomioon kääntäessään audiovisuaalisia tekstejä.

Tutkimukseni jakautuu osa tutkimuksiin, joissa tarkastelen tätä ilmiötä erilaisista näkökulmista. Tutkimusten tuloksia käytetään Pro gradu -tutkielmaani (2017) varten luomani audiovisuaalisten tekstien analysointimenetelmän kehittämiseen. Loin kyseisen analysointimenetelmän, sillä olemassa olevat analysointimenetelmät eivät vastanneet tutkimukseni tarpeisiin sellaisinaan. Usein ongelmana oli se, että analyysista tuli auttamatta liian pitkä ja raskaslukuinen. Tavoitteeni oli myös nivoa audiovisuaalisten käännösten tutkimusta muihin läheisiin aloihin, kuten elokuvatutkimukseen. Kehittämässäni metodissa ”Multimodaalinen kohtaus- ja jaksoanalyysi” on vielä hiomista ja tämä onkin yksi tavoitteeni väitöskirjatutkimuksessani.

Väitöskirjani osatutkimuksilla pyrin hankkimaan lisätietoa siitä, kuinka tärkeänäammattikääntäjät pitävät kuvaa ja ääntä kääntäessään sekä miten usein heillä on pääsy kaikkiin näihin elementteihin kääntäessään. Pyrin myös tarkastelemaan, miten pääsy erilaiseen määrään audiovisuaalisen tekstin moodeja vaikuttaa kääntämiseen. Tutkimuksen tarkoituksena on selvittää, miten esimerkiksi erilaiset moodien yhdistelmät (pelkkä teksti/teksti+kuva/teksti+ääni/kaikki moodit) vaikuttavat tekstittämiseen sekä prosessina että tämän prosessin lopputulemana.

Ahonen, T. 2017. *Multimodal scene and sequence analysis: condensation and reduction strategies in the subtitles of The Dark Knight*, Pro gradu, Itä-Suomen yliopisto, Filosofinen tiedekunta, Humanistinen osasto, Vieraat kielet ja käännöstiede, http://urn.fi/urn:nbn:fi:uef-20180058

Khalid Alnajjar (University of Helsinki), Mika Hämäläinen (University of Helsinki) & Jack Rueter (University of Helsinki)

A MediaWiki Environment for Curating Dictionaries by Intercomparison and Community Involvement

Lexicographic resources for endangered languages are often fragmented into several different paper publications that have been digitized by different conventions and during different eras. The severely endangered Skolt Sami is no exception. For one part, a massive open source dictionary for Skolt Sami has been made available on the MediaWiki based Akusanat (Hämäläinen & Rueter, 2018) and the Giellatekno infrastructure (Moshagen et al., 2013). For the other, valuable resources such as Sammallahti and Moshnikoff materials (Sammallahti & Moshnikoff, 1991) are not formatted in the same structure and their inclusion in the aforementioned systems poses a challenge.

In order to make all three of the lexicographic resources for Skolt Sami available for use for researchers and non-academic dictionary users alike, we propose a new extension to the MediaWiki based Akusanat dictionary that makes intercomparison of the different materials possible together with a facilitated semantic search functionality. The main goal is to alleviate the workload of dictionary editors when bringing lexicographic data available in the unified system. As community involvement in editing online dictionaries has previously been identified as a viable way of extending the lexicographic data (cf. Everson et al, 2019), we propose a simplified interface for non-technical community members to actively participate in the dictionary editing process.

The intercomparison of the different dictionaries is made possible by our external online toolkit which allows the user to import a set of words for comparison with the Akusanat MediaWiki system data. This juxtaposition of the new data with the existing one in the system, can then be queried with a powerful semantic search functionality, edited and published in Akusanat.

As MediaWiki is highly customizable, we introduce an extension with an intuitive user interface that allows the user to find and filter lexemes. The user can use the extension to find lexemes by their part of speech, source of the entry, translation languages and so on. Once the user has submitted the desired query, the system returns the matching lexemes along with additional properties if requested. These properties are, for instance, inflection category, semantic tag and assonance. Using these meta-information, researchers can easily gather similar words together.

We implement a simplified interface for involving the community speaking Skolt Sami to improve the quality and accuracy of the information presented in our system. The extension allows users to go through words and their translations in the dictionaries masking all the additional technical information of each entry. Users can then verify the correctness of the translations, suggest edits and provide reasons for their opinion. All input feedback along with who has provided it is stored in the system.

Shoju Chiba (Reitaku University)

Theory and practice of enriching a word list: a case study of building a student glossary of Finnish for Japanese learners

Giving a suitable and reliable word list is essential for language learning. Although building a frequency-based word list has become a global standard with the advent of large-scale language corpora, the method still embraces several weeknesses. For example, the lack of the availability of corpora with rich register variation reduces the reliability of the product. In addition, it is not sufficiently recognized that even if a frequency-based word list can picture the unbiased "reality" on the use of the words of a language, it may not serve as the "ideal" tool for the learners of the language. Namely, learners of a foreign language may not always set their sights on acquiring the vocabulary of the same size as of fluent language speakers.

The motivation and background of language learners, living far-off and hence with less contact with natives, may naturally differ from the one in the vicinity of the country where a target language is spoken. In this paper, I first introduce the situation of Japanese learners of Finnish, arguing that we need to set a different goal of mastery of vocabulary for them. What they need will be a more culture-oriented word list, embracing e.g. the travelers' perspective, a view of those who visit the country as foreigners, which are substantially different from the one CEFR-based language materials want to achieve.

This type of enterprise, e.g. building a "reliable" Finnish word list for foreigners, faces methodological difficulties to overcome. The direct recourse to the corpora is nothing less than inadequate. Inbuilding modern language technologies, this paper proposes a cycle of improving such a special kind of learner glossary, where we assess and complement the list using data-oriented methodology.

The main aim of this paper is thus twofold. Firstly, it attempts to show how we can evaluate an existing word list with assistance of modern language technologies like weighting with various frequency calculations and semantic similarity assessment induced by word-vector models. Secondly, this paper attempts to show how the frequency-induced importance ranking of a word in the list can deviate from the viewpoint of native speakers, language teachers and learners. The integration of the different perspectives is of substantive importance for a word list to be improved, hence a theory of building a targeted word list in a reliable way should be desired.

Steven Coats (University of Oulu)

Regional Variation in Speech Rate in American English from YouTube Videos

Features pertaining to the temporal organization of speech such as speaking rate, articulation rate, or pause duration can vary in American English according to dialect or speaker location (Jacewicz et al. 2009, Kendall 2013), but previous studies have mostly not analzyed samples with geographic granularity sufficient for generalizations about regional differences within the United States.

In this study, a corpus of automatic speech-to-text transcripts of spoken American English compiled from more than 29,000 hours of video from YouTube (Coats 2019), mainly of meetings of local government or civic organizations in all 50 states, is used to analyze regional differences in articulation rate. The principal finding confirms a popular conception: speakers from the South articulate slower than average, and speakers from the Upper Midwest more quickly. The study also introduces several methodological innovations: First, a new method for (semi-) automatic corpus compilation from publicly-available YouTube video transcripts is presented. Second, a method is introduced for the calculation of articulation rate using cue and word timestamps from captions files. Third, spatial autocorrelation analysis, used successfully in recent studies of regional variation in written American English (e.g. Grieve 2016), is undertaken for the analysis of articulation rate. Finally, a method for mapping and interactive visualization of articulation rate differences is introduced.

References:
Coats, Steven. 2019. A corpus of regional American language from YouTube. In Costanza Navarretta et al. (eds.), *Proceedings of the 4th Digital Humanities in the Nordic Countries Conference*, Copenhagen, Denmark, March 6–8, 2019. Aachen, Germany: CEUR.
Grieve, Jack. 2016. *Regional variation in written American English*. Cambridge, UK: Cambridge University Press.
Jacewicz, Ewa, Fox, Robert A., O'Neill, Caitlin & Salmons, Joseph. 2009. Articulation rate across dialect, age, and gender. *Language Variation and Change* 21, 233–256.
Kendall, Tyler. 2013. *Speech rate, pause, and sociolinguistic variation: Studies in corpus sociophonetics*. London: Palgrave-Macmillan.

Enum Cohrs (University of Eastern Finland) & Wiebke Petersen (University of Düsseldorf)

Guessing a tweet author's political party using weighted n-gram models

Political parties and their candidates are increasingly using online channels for their electoral campaigns. This was, for instance, observable for the elections for the German parliament (Bundestag) in 2017. But even outside the campaigning time, politicians use Twitter to inform about their work and current topics. For an informed human it is usually easy to guess their political affiliation even if it is not explicitely stated in the tweets. In this paper we present a probabilistic classifier for the political party of a tweet's author and compare different weight configurations, either weighting by word frequency or by part of speech. In

opposition to many existing systems that focus on the US and only work with two-party political systems, our model allows an arbitrary amount of parties. For the German election we included 9 political parties into the analysis and our system achieved an accuracy of 72 % when perusing all tweets published by an author in the specified time interval, or 36 % accuracy when using only one single tweet as input. A random guessing baseline system would have an expected accuracy of 11 % in both cases.

Izabela Czerniak (University of Eastern Finland)

The System of Relativisation in English(es): Some Preliminary Results

This paper is a part of post-doctoral study which examines the environment of the relative clause (RC) in English, including earlier and later stages of development of the language as well as exploring the dynamics of changes in the current varieties. The research derives from the ideas explored in the researcher's PhD study on word order change in Early English (Czerniak, 2016), where many processes and phenomena were identified as being linked to the emerging strict SVO order. One of these phenomena was the rearrangement of relativisation strategies in the English language (e.g. Dekeyser, 1986).

Along with increasing analyticity, there is, among others, a visible difference noted in the choice of relative pronouns when moving from the earlier towards more recent stages of English (e.g. van Gelderen, 2006, p. 173). There are also features particularly characteristic of the current Englishes, with the so-called "doubling or copying" of the relative pronoun (e.g. van Gelderen, 2006, p. 260) being one of the notable instances.

In the present study, the distributions of the majority of common relativisers, including the nonstandard what (e.g. Hermann, 2005, p. 22) and a much preferred ModE as (e.g. van Gelderen, 2006, p. 217) are taken into account. This research also examines the validity of factors behind the rise (or decline) in use of particular relativisers across time. Among these potential factors, as mentioned in various studies on the topic, there are structural ambiguities, processing difficulties, stylistic conditioning, and strong prescriptive tradition (Leech et al., 2009, pp. 228-230). The current stage (1) involves the investigation of the RC environment at the earlier stages of development of English, using the (updated) parsed sections of the Penn Historical Corpora. Some preliminary results of the searches on late Middle and Early Modern English data will be presented. As customary for any corpus research, this study relies on a frequency-based model. Additionally, using a variety of sampling techniques (e.g. controlled (multiple) sampling, stratified sampling, a comparison of sample versus 'population') is necessary in order to not only estimate the impact of factors behind the distribution of relativisers but also to address the issues related to texts of the corpora themselves.

References:

Czerniak, I. (2016). *Anglo-Scandinavian Language Contacts and Word Order Change in Early English*. Publications of the University of Eastern Finland. Dissertations in Education, Humanities, and Theology, no. 85. Joensuu: University of Eastern Finland.
Dekeyser, X. (1986). English contact clauses revisited: A diachronic approach. *Folia Linguistica Historica*, 7.1: 107–120.
Gelderen, E. van. (2006). *A History of the English Language*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Hermann, T. (2005). Relative Clauses in English Dialects of the British Isles. In B. Kortmann, T. Hermann, L. Pietsch and S. Wagner (eds) *A Comparative Grammar of British English Dialects*. Berlin, New York: Mouton de Gruyter, 21-123.
Leech, G., Hundt, M., and Mair, C. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Corpora used:

Kroch, Anthony, and Ann Taylor. (2000). The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4 (http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4).
Kroch, Anthony, Beatrice Santorini, and Lauren Delfs. (2004). The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3 (http://www.ling.upenn.edu/ppche-release-2016/PPCEME-RELEASE-3).

Senka Drobac (University of Helsinki) & Krister Lindén (University of Helsinki)

Optical Font Family Recognition using a Neural Network

Working on OCR (Optical character recognition) of historical newspaper and journal data, we found it beneficial to analyze and evaluate our OCR results based on font family. Our data sets are extracted from a corpus of historical newspapers and magazines (from 1771 until 1874) that have been digitized by the National Library of Finland. Our earlier data is mainly written in Blackletter fonts and later data in Antiqua fonts, while in the transitioning period both font families were used at the same time, even on the same pages. Therefore, in order to make the recognition phase easier and faster, we are building one OCR model, which is able to recognize all fonts represented in the data. In order to make sure that we have sufficient training data for both font families, we need a font family classifier to simplify creation and sampling of training data.

Although there are existing tools for font classification, our problem seems to be overly specific. We only need to distinguish between Blackletter and the other fonts that were printed in Finland between 18th Century and early 20th Century, so the challenge is to find a simple enough font classifier for such a specific task.

Sahare et al., 2017 conveyed a detailed survey of different script identification algorithms. Zramdini et al., 1998 have developed a statistical approach of font recognition based on global typographical features. They report 97 % accuracy of typeface recognition. Brodić et al., 2016 have approached a similar problem as we have, when they do identification of Fraktur and Latin scripts in German historical documents using image texture analysis. The accuracy of their system has been reported to be 98.08 %.

In this work, we build a deep neural network binary font family classifier that for an image of one line of text decides whether it is written in Blackletter or Antiqua typeface. Even with a simple configuration of the network, we get 97.5 % accuracy, leaving space for further improvement.

This font family classifier is specifically created for historical OCR for data printed in Finland. It is useful for collecting and analyzing the data, especially if the OCR is done with line-based software (Ocropy, Kraken, Calamary, Tesseract 4). The font classifier is simple to use, in both the training and prediction phases. It is also easy to change network configurations and parameters.

References:

Sahare P & Dhok SB (2017) Script identification algorithms: a survey. *International Journal of Multimedia Information Retrieval* 6(3): 211–232.
Brodić D, Amelio A & Milivojević ZN (2016) Identification of fraktur and latin scripts in german historical documents using image texture analysis. *Applied Artificial Intelligence* 30(5): 379–395.
Zramdini A & Ingold R (1998) Optical font recognition using typographical features. *IEEE Transactions on pattern analysis and machine intelligence* 20(8): 877–882.

https://github.com/tmbdev/ocropy
http://kraken.re/
https://github.com/Calamari-OCR/calamari
https://github.com/tesseract-ocr/tesseract/wiki/4.0-with-LSTM

Irantzu Epelde (CNRS)

Norantz Corpus, a comprehensive resource for studying Basque language

In language contact situations it is very common to have the less prestigious language being influenced by the most prestigious one, but the contrary is also attested (Heine & Kuteva 2005). Basque is considered a 'vulnerable' language (Moseley 2010), one that lives in a diglossic situation with respect to two of the world's most powerful languages: Spanish and French. Nowadays, all Basque speakers are bilingual with either French or Spanish, and this language contact situation makes Basque very prone to borrowings from these two languages.

The case of phonetic features is of a special relevance in externally-induced variation and change phenomena, given that it is widely assumed that, in cases of language contact, phonetic features are amongst the easiest ones to borrow (Silva-Corvalán 2001). In this paper, we will focus on the addition of a prothetic /e/ in new loanwords produced by older speakers from the French area.

Word-initially adapted borrowings show a prothetic vowel in [sC-] contexts among older speakers. The word-initially inserted vowel is usually e-; e.g. estop 'stop', eslip 'slip', espor 'sport', estok 'stock'… Present-day younger speakers of Basque have no problem articulating a word-initial [sC-], since all of them are bilingual. In French, this prothetic vowel doesn't occur, but according to Basque tradition, until recently, it was generally favored in the adaptation of new borrowings (Mitxelena 1961, Hualde 1991, Hualde & Ortiz de Urbina 2003).

In our data, the addition of this unit appeared to be associated with differences in age: www.norantz.org. We will focus on three age groups: youngs (-30), middle-aged (40-60) and octogenarians (+80) from the traditional provinces of Lapurdi, Low Navarre and Zuberoa (France). All of the informants (60) have the Basque language as their mother tongue and home language, but the older ones received education only in French language and use French in formal (and often informal) situations, in oral and in written communication. The data come from recorded interviews —individual as well as in-group— held in Basque, and from specific questionnaires and word lists. A sample of data is analysed perceptually and acoustically with the Praat speech analysing program in order to measure the occurrence or the absence of this initial /e/.

Language variation can mark stable class differences or stable sex differences in communities, but it can also indicate instability and change. When it marks change, the primary social correlate is age (Chambers 2002), and the change reveals itself prototypically in a pattern whereby some minor variant in the speech of the oldest generation occurs with greater frequency in the middle generation and with still greater frequency in the youngest generation. If the incoming variant truly represents a linguistic change (Labov 1994, Trudgill 1974), as opposed to an ephemeral innovation as for some slang expressions or an age-graded change, it will be marked by increasing frequency down the age scale, as it occurs with the youngest generation in this community.

Olga Gerassimenko (Center of Estonian Language Resources), Neeme Kahusk (Center of Estonian Language Resources), Marin Laak (Estonian Literary Museum) & Kadri Vider (Center of Estonian Language Resources)

Morphologically Annotated Corpus as a Resource for Literary Studies

Literary heritage is being massively digitized in Estonia, making it possible for the researchers to access large amounts of text data (Rand 2018). Yet, digitization often limits itself to merely capturing texts without consistent description of metadata or linguistic tagging of the captured texts. Digitized texts are often left unlemmatized that considerably hinders the corpus search in morphologically rich languages such as Estonian; metadata might be copied from the originals in neither consistent nor computer-readable format. The raw unstructured text as digitization outcome supports the prevalent methodological tradition to work with digitized data through close reading of large amounts of texts. Close reading allows to pay thorough attention to context but significantly limits the possible scope of research. There is a growing need for making full use of the digitized massives of texts in the literary studies but literary scholars do not yet have much experience with linguistically annotated corpora.

In our pilot project of preparing and making publicly available a linguistically annotated literary corpus we aimed to make the corpus search useful for literary scholars (Laak et al. 2019). We selected a correspondence of two Estonian writers, Johannes Barbarus and Johannes Semper, during 1910-1940. The correspondence is unique as it covers several decades of the authors' lives, and the authors are prominent figures in Estonian literature and close childhood friends who discuss cultural and political issues in an exceptionally straightforward and sincere manner. The manuscript letters had already been manually typed-in during their preparation for the printed publication. We have chosen the interface of corpus query system KORP created in Swedish Language Bank Språkbanken and developed in several countries as an open source resource (Borin et al. 2012). KORP allows to search differently annotated corpora simultaneously and is extremely flexible in terms of text and metadata categories included in query and statistics. Our corpus was automatically tagged for sentence and text block boundaries, morphologically analyzed, lemmatized and disambiguated. The metadata was unified semiautomatically: i.e, dates such as "on a third day of Christmas" were customized manually.

We have explored linguistically annotated corpus advantages for literary studies by looking in the literary and societal issues discussed in the letters of Semper and Barbarus. Given the morphological annotation, we can retrieve not only the occurrence and broader context of the lemmatized keywords we are looking for (i.e, political regimes, literary associations, places, names of writers or historical figures) but also grammatically identifiable data associated with those keywords (i.e, epithets and other descriptive words, ethnic groups or proper names, actions, foreign language quotations). Selectively compiled statistics allows us to make full use of metadata in describing the occurrences of keywords: for every corpus query according to the metadata or text categories chosen to compile the statistics, it can be seen at a glance, i.e, whether the keywords occur in the texts of both authors, whether they form longer discussion chains occuring in the immediately related letters and whether they belong to a certain time period. It allows to swiftly check the existing hypotheses and to discover new textual data connections. Our next aim is to prepare a morphologically annotated corpus of Estonian literary criticism.

Simon Hengchen (University of Helsinki), Jani Marjanen (University of Helsinki), Ruben Ros (Utrecht University) & Mikko Tolonen (University of Helsinki)

The omnipresence of the nation

We mine two centuries of digitised newspapers in four languages, and propose a methodologically sound, reusable approach to carry out quality historical research on the changing vocabulary relating to nationhood.Newspaper collections are increasingly used to address historical questions through mining

textual data. They are more seldom used for comparative projects cross linguistic and national boundaries. In this paper, we address the methodological challenges the use of newspapers from different political contexts, languages and datasets (bearing in mind also different data formats) poses, and lay out our approach to tackle a comparative study for the Netherlands, Finland, Sweden, and the UK.

After retrieving the datasets and shaping them in a way that a single pipeline can be reused for all languages and historical realities comes the trade-off between the computational, distant reading of the text, and the actual research question. We focus on the process of nation-building in Europe and test several methods. Whilst historical processes or concepts are complex, and thus cannot be the object of a mere tallying across time, it is obvious that words do. We thus use words as a proxy to study the process of nation-building, and carry that out in several ways. In doing this we also limit the study of nation-building to the development in which the nation became a self-evident frame. As such, we do not trace the theoretical development of the concept of nation or even the intentional processes of shaping Dutchness, Britishness, Swedishness, or Finnishness, but rather focus on the more implicit process in which the nation became a natural frame for conceptualizing societal issues or -- quoting Anderson (2006) -- an imagined community that became inescapable for citizens of any state.

As a first window to the development of this idea, we look at how bigrams starting with the adjective "national" behave in our datasets, in terms of absolute and relative frequencies. This paints a picture of how common the idea of something "national" is mentioned in newspapers in different countries at different periods. We complement this picture with an analysis of the creativity and productivity of the "national bigram": by looking at how "creative" writers are with the linguistic unit, and by looking at how its use evolves across time, we examine part of the vocabulary of the nation, and can identify key junctures in the transformation of this vocabulary. Unsurprisingly, the French revolution, the political ruptures of 1848 and the War of 1870 were particularly important for the diversification in the vocabulary of "national", in all of our cases, but can also show how local political and publishing conditions produced local reactions. The differences also point out how events abroad affected domestic vocabulary, making the development a transnational one (cf. Bos and Gifford 2016). By focussing on bigrams, we open up for a second window of development and trace the domains in which the word "national" was used, thus tracking the process in which a national framing was deployed for almost everything in political and social affairs.

Kerttu Huttunen (University of Oulu)

Tutkimusaineiston kerääminen ja analysointi monipuolisia digitaalisia keinoja hyödyntäen. Esimerkkinä Tunne-etsivät-tutkimushankekokonaisuus

Digitaaliset pelit kuuluvat nykyään kiinteänä osana lasten elämään – niiden pelaaminen on yksi nykylasten leikin muoto. Sähköisiä, vuorovaikutteisia pelejä käytetään yhä enemmän myös kommunikointihäiriöisten lasten kuntoutuksessa, sillä mielenkiintoisten pelien avulla saadaan ylläpidettyä lasten harjoittelumotivaatiota ja harjoitteluun voi sisällyttää loputtoman määrän toistoja. Digitaaliset pelit soveltuvat erityisen hyvin autismikirjon lasten kuntoutukseen.

Monitieteisessä neljän suomalaisen yliopiston hankkeena toteutettavassa Tunne-etsivät-tutkimuskokonaisuudessa tuetaan lasten kielellisiä ja sosioemotionaalisia taitoja, erityisesti lasten kykyä tunnistaa tunteita kasvoilta, puheesta ja sosiaalisista vuorovaikutustilanteista. Tukeminen tapahtuu hankkeessa laaditun, verkossa pelattavan Tunne-etsivät-pelin avulla. Kyseinen peli on tähän saakka laajin ja monipuolisin Suomessa tunteiden tunnistustaitojen harjaannuttamiseen laadittu sähköinen materiaali.

Kommunikointihäiriöisiä lapsia koskevassa hankkeen osassa tutkitaan, voidaanko digitaalisen pelin avulla vahvistaa lasten kielellisiä ja sosioemotionaalisia taitoja ja erityisesti, voidaanko pelin avulla parantaa lasten kykyä tunnistaa tunteita. Jyväskylän yliopistossa väitöskirjahankkeena toteutuvassa osassa taas keskitytään

siihen, miten peli toimii yhteisöllisen oppimisen välineenä: millaista on yhteisöllinen oppiminen ja millaista vuorovaikutusta tyypillisesti kehittyvien lasten kesken syntyy, kun he pelaavat peliä internetissä yhdessä toisen lapsen kanssa.

Tässä artikkelissa kuvataan digitaalisen Tunne-etsivät-verkkopelin kahden tutkimusversion ominaisuuksia. Nämä tutkimusversiot kehitettiin hankekokonaisuuden kahden osahankkeen aineistonkeruuta varten. Lisäksi kuvataan näiden osahankkeiden digitaalista aineistonkeruuta ja analysointia.

Hankkeen tutkimusaineistona on lasten suoriutuminen kielellisiä, kognitiivisia ja tunteiden tunnistamisen ja tuoton taitoja kartoittavissa tehtävissä ja testeissä. Tutkittavana oli 55 lasta, joilla oli joko autismikirjon häiriö, ADHD, kehityksellinen kielihäiriö tai kuulovika. Näiltä lapsilta tutkittiin mm. tuottava sanavarasto, lyhytaikainen muisti, mielen teorian taidot sekä kerrontataidot. Lisäksi tutkittiin keskittymiskykyä ja reaktionopeutta sekä tunteiden tunnistus- ja tuottokykyä. Aineistonkeruussa hyödynnettiin näillä alueilla keskeisinä ärsykkeinä digitaalisia materiaaleja. Lapset mm. tekivät hankkeessa laaditun, tietokoneella tehtävän reaktioaikatestin sekä nimesivät ja tuottivat ilmeitä ja äänensävyjä heille tietokoneella esitettyjen valokuvien, videoleikkeiden ja äänitiedostojen pohjalta. Testaushetket videoitiin ja myös tallennettiin Zoom-äänitallentimella äänitiedostoiksi ennen pelaamisinterventiota ja sen jälkeen. Tunne-etsivät-pelin lokitiedostoista kerättiin tieto kunkin lapsen harjoitusmääristä eli kertyneestä pelaamisajasta ja myös eri tehtävissä onnistumisesta ja kunkin tehtävän suorittamiseen kuluneesta ajasta.

Tausta-aineistoksi kerättiin 109 tyypillisesti kehittyvän lapsen suoriutumistulokset edellä kuvatuissa mielen teorian ja tunteiden tunnistuskyvyn testeissä ja tehtävissä. Vuorovaikutuksen ja mm. paripelaamistyyppien ja niiden muutosten analysointia varten puolestaan videoitiin väitöskirjahankkeessa 16 tyypillisesti kehittyvän lapsen Tunne-etsivät-pelin pelaamista toisen lapsen kanssa parina.
Tutkimusaineisto on nyt kerätty ja se on analyysivaiheessa. Osa tuloksista on jo raportoitu. Videoituja testaustilanteita on hyödynnetty eri testien ja tehtävien pisteytyksessä tai pisteytyksen tarkistamisessa. Niiden avulla on myös litteroitu lasten kuvasarjakerrontatehtävässä tuottamat kertomukset ja lasten sanallinen ja ei-sanallinen vuorovaikutus paripelaamishetkissä. Videoista havainnoidaan myös eleiden käyttöä kuvasarjakerronnassa. Äänitiedostoista puolestaan analysoidaan kuvasarjakerrontatehtävässä sitä, millaista lasten puheen prosodiikka on kerronnan aikana ja käyttävätkö lapset prosodiikkaa tunteiden ja muiden mielentilailmausten korostamiseen.

Edellä kuvatut tutkimushankkeet edustavat monitieteisiä ja moniammatillisia tutkimuskokonaisuuksia, joissa aineistoja kerätään digitaalisesti ja ne myös analysoidaan monipuolisia sähköisiä työkaluja käyttäen. Olennaista on myös eri tieteenaloja edustavien tutkijoiden yhteistyö, toisilta oppiminen ja toisten tieteenalojen tutkimustyökaluihin tutustuminen ja niiden hyödyntäminen. Lopuksi artikkelissa kuvataan laajojen digitaalisten aineistojen keräämiseen, analysointiin ja säilyttämiseen liittyviä kokemuksia

Ali Zeeshan Ijaz (University of Helsinki), Mikko Tolonen (University of Helsinki), Leo Lahti (University of Turku) & Iiro Tiihonen (University of Helsinki)

Analytical determination of editions from bibliographic metadata

Analytical bibliography aims to understand the production of books. Systematic methods can be used to determine an overall view of the publication history. In this paper, we present the state of the art analytical approach towards the determination of editions using the ESTC meta data. The preliminary results illustrate that metadata cleanup and analysis can provide opportunities for edition determination. This would significantly help projects aiming to do large scale text mining.

Laura Ivaska (University of Turku)

Distinguishing translations from non-translations and identifying (in-)direct translations' source languages

The scope of this study is threefold. First, machine learning will be applied to distinguishing translated and non-translated Finnish texts. Then, an attempt to identify the source languages (SL) of the translated Finnish texts will be made. Finally, SL identification will be tested with indirect translations (ITr), that is, with translations made from translations (Assis Rosa, Pieta and Maia 2017). The three underlying research questions are thus: 1) Can translated Finnish be distinguished from non-translated Finnish? 2) Can the SLs of Finnish translations be identified? 3) If the answer to question 2 is yes, then what happens when the method is applied to ITrs; will the analysis identify the ultimate SL, the mediating language or neither?

This study is based on the hypothesis that translated language contains traces of the SL (Toury 1995). The corpus of the study consists in non-translated Finnish prose, Finnish prose literature translations made from English, German, French, Modern Greek and Swedish, as well as ITrs from Modern Greek into Finnish via various mediating languages. The analyses are based on cluster analysis and support vector machines (SVM) using the frequencies of the most frequent words (MFW).

Results show that SVM-based machine learning techniques can distinguish between translated and non-translated Finnish. The SVM-based SL detection, however, proved only partially successful, while a cluster analysis suggested that there is coherence within a group of texts translated from the same SL and variation between the groups of texts with different SLs. Clustering was further tested with ITrs and the results were mixed: six of the 13 tested ITrs custered with direct tranlsations from the ultimate SL, two with translations from their mediating languages and five with neither.

Jarmo Harri Jantunen (University of Jyväskylä) & Samu Kytölä (University of Jyväskylä)

Homot ja uskonnot digitaalisissa diskursseissa

Suomalaisissa kansalaiskeskusteluissa ja digitaalisissa diskursseissa tapahtuvaa homoseksuaalisuuskeskustelua on tutkittu toistaiseksi vain vähän (ks. kuitenkin Charpentier 2000, 2001; Jantunen 2018). Vielä vähäisempää on tutkimus, jossa tarkastellaan homouden ja uskontojen suhdetta näissä diskursseissa; uskontotieteessä, oikeustieteessä ja queer-tutkimuksessa uskonnon ja homoseksuaalisuuden suhdetta on kuitenkin aiemmin tarkasteltu (ks. van den Berg ym. 2014). Jo vuosituhansia homoseksuaalisuuteen on otettu kantaa uskonnollisessa viitekehyksessä (Carlson & Carlson, 2013; Boisvert & Johnson 2011), ja uskonnon avulla on haluttu perustella esimerkiksi homoseksuaalisuuden luonnonvastaisuutta ja syntisyyttä tai esimerkiksi tasa-arvoisen avioliiton kieltämistä (van der Toorn ym. 2017). Diskursseissa näkyvät asenteet, jotka syntyvät usein jo kasvatuksessa ja koulussa; homoseksuaalisuus on esitetty syntinä esim. uskonnonopetuksessa (Lehtonen 2003: 122): vielä 2000-luvun alussa uskovaisuus ja homoseksuaalisuuden vastustaminen korreloivat voimakkaasti (Borg ym. 2007).

Verkkokeskustelupalstat tarjoavat niin sanotulle tavalliselle kansalle tilan tuoda julki näkemyksiä ja kannanottoja, ja näissä digitaalisissa diskursseissa kohtaavat niin maalliset kuin uskonnollisetkin näkemykset. Esityksessämme tarkastelemme homoseksuaalisuudesta ja uskonnosta käytävää keskustelua Suomi24:ssä. Aiemman tutkimuksen (esim. Charpentier 2000, 2001; Bachmann 2011) mukaan kristillinen diskurssi on yksi tavallisimmista homoseksuaalisuuteen liittyvistä diskursseista mm. luonnollisuus-, kriminalisointi- ja sairaus-diskurssien rinnalla (ks. Bachmann, 2011; Baker, 2004). Suomi24-kansalaiskeskusteluissa diskurssi näyttäytyy kristillistä diskurssia (syntiä, irstautta ja pyhää järjestystä, Charpentier 2000, 2001) laajempana ja monipuolisempana uskonto-diskurssina. Sitä tuotetaan esimerkiksi keskustelemalla kirkosta, Raamatusta, Jumalasta ja Jeesuksesta, mutta myös (nais)pappeudesta, ateismista

ja eri uskontokuntiin liittyvistä teemoista, kuten juutalaisuudesta, muslimeista ja islaminuskosta (Jantunen 2018). Uskonnolliset ryhmittymät ja homot nähdään myös esimerkkeinä ongelmallisista vähemmistöryhmistä Suomessa; tämä heijastaa verkkokeskusteluissa tavallista vihapuhetta (vrt. Baker 2005: 68–71). Tarkastelemme esityksessämme Suomi24:n homoseksuaalisuuteen ja uskontoon liittyvää kansalaiskeskustelua korpusavusteisen diskurssintutkimuksen avulla. Lähtökohtanamme ovat Suomi24-korpuksesta tehdyt avainsana-analyysit, joista laajennamme ja syvennämme tarkastelua laadullisen, kriittisen diskurssianalyysin suuntaan.

Heidi Jauhiainen (University of Helsinki), Tero Alstola (University of Helsinki), Aleksi Sahala (University of Helsinki), Saana Svärd (University of Helsinki) & Krister Lindén (University of Helsinki)

Akkadian Cuneiform Texts and Digital Tools

In the Semantic Domains in Akkadian Texts project, we are studying texts originally written in the cuneiform script. We are using language technological methods, such as fastText and PMI, as well as Social Network Analysis to find semantic contexts and relations of words in Akkadian, an East Semitic language that was written and spoken in Mesopotamia c. 3000 – c. 500 BCE. The texts we are analyzing come from the Open Richly Annotated Cuneiform Corpus (Oracc). Oracc is an international cooperative effort containing free online editions of texts from various projects and it is one of the largest electronic corpora of Akkadian texts.

The corpus we have downloaded from Oracc contains 16,487 texts and almost 2 million words. About half of these texts have been tagged as having been written in the Akkadian language. The basic unit of words in Oracc is their transliteration, i.e. the representation of the cuneiform signs in Latin script. More than half of the words have been annotated with, for example, dictionary forms, word senses, and part-of-speech tags. Since Akkadian is an inflecting language, we have opted for using the dictionary forms of the words when analyzing semantic contexts.

Annotation of a text is always an interpretation by a scholar and, as Oracc is composed of a number of subprojects, the way a word is annotated in different texts is not always consistent. Therefore, we have done a lot of preprocessing of the texts, such as normalizing the ways deities and places are referred to. In Akkadian, a word can often have several meanings. We have found a way to distinguish homonyms according to their translations in the annotation. As the annotation of the texts has been done by hand in many different projects, the translations of a word can vary from one to another. We have hence combined the synonymous translations of most of the words.

We have done some of the preprocessing automatically or semi-automatically, but we have also had to do some manual work. In this presentation, we describe the ways we have preprocessed the Oracc data before we can use it for analysing the semantic contexts of the words.

Heidi Jauhiainen (University of Helsinki), Tommi Jauhiainen (University of Helsinki), Krister Lindén (University of Helsinki)

Wanca in Korp: Text Corpora for Under-Resourced Uralic Languages

This paper introduces "Wanca in Korp", a sentence corpora for under-resourced Uralic languages, the pipeline used in creation of the corpora, as well as the various tools used as part of the pipeline.

The Ethnologue recognizes 38 different Uralic languages. The Uralic language group includes mostly linquistically under-resourced languages and only three of the Uralic languages are used as national

majority languages: Hungarian, Finnish, and Estonian. We have chosen all the minority languages of the Uralic branch as languages of interest and created new sentence corpora for most of them using texts openly available on the internet.

For gathering the texts from the internet, we used an open-source web-crawling software, Heritrix, developed and used by the Internet Archive in cooperation with several National Libraries. In addition to conducting our own crawling, we also used the pre-crawled corpus distributed by the Common Crawl Foundation.

In order to determine which pages were written in one or more of the relevant languages, we used a state-of-the-art language identification software developed within the project. For post-processing the identified pages, we developed a web-service, Wanca, where experts and native speakers of each language can participate in manual curation of the crawled links.

The process of sentence corpora creation begins from the pages tagged with relevant languages in Wanca. All the texts available behind existing Wanca links are downloaded. A language set identifier is used for whole texts in order to verify that the downloaded web pages still include texts in relevant languages. Then complete sentences are extracted from the texts. The language of each sentence is again identified and the sentence added to the sentence collection of the respective language. For the most rare languages, a manual curation and additional manual processing is conducted for those pages where possible sentences were found.

The end product is a sentence corpus collection for 28 less-resourced Uralic languages ranging in size from 20 sentences of Vod to 214,226 sentences of North Saami. The corpus is available in the Language Bank of Finland. The work has been conducted within the Kone Foundation funded project "The Finno-Ugric Languages and The Internet" at the University of Helsinki as part of FIN-CLARIN.

Marko Jouste (University of Oulu, The Giellagas Institute), Jack Rueter (University of Helsinki) & Marko Marjomaa (Sámi Giellagáldu)

Challenges in Developing Technology for the Saami Archive Language Resources

The Saami Culture Archive was founded to preserve a notable existing collection of culture and language material describing the historical and present-day Saamis and provide it for wide scientific and cultural use, as well as to gather new material. The archive has been actively developed since 2008 and funded by the University of Oulu and Academy of Finland (FIRI2014). The collection of the Saami Culture Archive consists of a noteworthy amount of sound, video and photograph material as well as digitized documents of Saami traditional culture and various Saami cultural activities. We believe that they can play an important role in the strengthening and revitalization of these minority languages.

In our paper, we will discuss the challenges in developing technology for the Saami Archive language resources. The main concern is how we can provide the language resources to three different Saami language communities in Finland (North Saami, Inari Saami and Skolt Saami) and in each of them, for two types of users, students and linguists.

1) Language technology for students. Technology seems to be the most promising way in the present modern world, where even ordinary people use frequently various applications for example in their smart phones. At present, however, we do not have enough information of how language technology and language resources are used by those who are studying Saami languages. In order to achieve new data, we have operated a query for the students of all three languages asking specific questions on their use of language technology.

2) Language technology for linguists. As a starting point, linguists require language resources for their work and research. The collection of the Saami Culture Archive consists of hundreds of hours of spoken Saami of which a notable amount has been transcribed and entered into an Elan infrastructure. Making spoken language available for language technological tools is crucial, since then it is possible to easily find material in vast collections. The main question is how to create both restricted and open access services which do not compromise the anonymity of the Saami Archive data?

Markus Juutinen (University of Oulu)

Skolt Saami current situation and technological tools

In this presentation I will deal with the current situation of the Skolt Saami language as well as the challenges of the language technological tools supporting the language community.

Skolt Saami is a language spoken some 200-300 speakers mainly in Finland in the municipality of Inari. The language is seriously endangered but with the efforts put to revitalizing proses the language has got dozens of new speakers during the last ten years. The number of second language learners is also rising. A normative writing system for Skolt Saami was developed during the 1970's. Due the new possibilities of studying, the reading skills has got more common during the last 10 years, even though a notable amount of the speakers is still illiterate in their language.

For children there have been language nests in Ivalo and Sevettijärvi from the 90's and some language teaching mainly in the school of Sevettijärvi. The Saami Education Institute gives language education for adult in Inari and via internet. It has also been possible to study Skolt Saami language in the University of Oulu since 2015.

Modern language technological applications, language resources and research of the can also aid the language community and the revitalization of the Skolt Saami language. Various tools have been developed in Giellatekno, Centre for Saami language technology, in the UiT Arctic University of Norway: Skolt Saami dictionary, proofing tool, keyboard, paradigm generator and word form analyzer. Furthermore, there exists archived language resources of more than 100 hours of spoken material in archives in Finland, Norway, Sweden, Russia and Estonia. so far, about 60 hours of this material has been transcribed in ELAN program.

Johanna Kalja-Voima (University of Jyväskylä)

Korpusavusteinen diskurssintutkimus (CADS) tiedeymmärryksen tutkimuksessa (PuS) – analyysiesimerkki tutkijoihin liittyvistä diskursseista

Tämän esitelmän tavoitteena on kuvata, miten tiedeymmärryksen tutkimuksessa (public understanding of science, PuS) (Short 2013) voidaan ja kannattaa hyödyntää korpusavusteista diskurssintutkimusta (corpus-assisted discourse studies, CADS) (Partington 2006; Jantunen 2018a). Esitelmässä kuvataan, mitä tiedeymmärryksen tutkimuksella (sekä laajemmin tieteentutkimuksella) tarkoitetaan, millaisia menetelmiä tutkimuksessa on tähän asti käytetty sekä miksi tieteen ja tiedeymmärryksen tutkiminen on tärkeää. Toiseksi esitelmässä havainnollistetaan, miksi tiedeymmärryksen tutkimukseen kannattaa soveltaa CADS-menetelmää ja millaista uutta tietoa tiedeymmärryksestä saadaan, kun analyysimenetelmänä on kielitieteellinen, korpusavusteinen diskurssintutkimus.

Empiirisessä osuudessa esitelmän tavoitteeseen vastataan tapausesimerkillä tutkijoihin liittyvistä diskursseista. Tapausesimerkin analyysi havainnollistaa, millaisia diskursseja tutkijoihin liitetään Suomi24-

keskustelufoorumilla (Suomi24-korpuksesta ks. Lagus, Pantzar, Ruckenstein & Ylisiurua 2016), kun keskustelijoina ovat pääsääntöisesti ns. maallikot. Analyysissa on sovellettu kollokaatioanalyysia (Stubbs 1996: 172–176, kollokaatioanalyysin soveltamisesta ks. Jantunen 2018b). Lisäksi esitelmässä pohditaan, miten analyysin tuloksia voi hyödyntää avoin tiede -periaatteen toteuttamisessa. Avoimen tieteen yhtenä tavoitteena on mahdollistaa tutkijoiden, päätöksentekijöiden ja kansalaisten osallistuminen tieteen ja tutkimuksen tekemiseen (Avoin tiede ja tutkimus 2014: 2). Tällä hetkellä tilanne lienee sellainen, että tiedeyhteisö ei juurikaan tiedä, mitä kansalaiset tieteestä ja tutkimuksesta sekä tutkijoista ajattelevat. Tähän mennessä kansalaisten näkökulmaa tutkijoista on kartoitettu esimerkiksi virallisella kyselytutkimuksella, Tiedebarometrillä (2016). Tämä tutkimus täydentää edellistä ja havainnollistaa, miten kansalaiset näkevät tutkijat aidossa, epävirallisessa yhteydessä eli keskustelufoorumilla. Kun tiedeyhteisö tietää laajemmin, mitä kansalaiset tutkijoista ajattelevat, voi se hyödyntää esimerkiksi viestinnässään tämän tutkimuksen tuloksia.

Esitelmä on osa laajempaa väitöskirjatutkimustani, jonka tavoitteena on paljastaa ja jäsentää ns. kansalaiskeskustelun digitaalisia diskursseja tieteestä ja tutkimuksesta.

Lähteet:

Avoin tiede ja tutkimus. Avoimen tieteen ja tutkimuksen käsikirja 2014.
https://avointiede.fi/sites/avointiede.fi/files/avointiede-kasikirja.pdf
Jantunen, Jarmo Harri 2018a: Homot ja heterot Suomi24:ssä: analyysi digitaalisista diskursseista. *Puhe ja kieli* 38 (1) s. 3–22. https://journal.fi/pk/article/view/65488/32762 8.2.2019.
Jantunen, Jarmo Harri 2018b: Korpusavusteinen diskurssianalyysi (CADS): analyysiesimerkki homouden ja heterouden digitaalisista diskursseista. – L. Haapanen, L. Kääntä & L. Lehti (toim.), *Diskurssintutkimuksen menetelmistä*. AFinLA-e. Soveltavan kielitieteen tutkimuksia 11 s. 20–44.
https://journal.fi/afinla/article/view/69259
Lagus, Krista, Pantzar, Mika, Ruckenstein, Minna & Ylisiurua, Marjoriitta 2016*: Suomi24. Muodonantoa aineistolle*. Valtiotieteellisen tiedekunnan julkaisuja 10. Helsinki: Kuluttajatutkimuskeskus & Helsingin yliopisto.
Partington, Alan 2006: Metaphors, motifs and similes across discourse types: corpus-assisted discourse studies (CADS) at work. – Walter Bisang, Hans Henrich Hock & Werner Winter (toim.), *Corpus-based approaches to metaphor and metonomy* s. 267–304.
Short, Daniel B. 2013: The public understanding of science: 30 years of the Bodmer report. *The School science review* 95 (350) s. 39–44.
https://www.researchgate.net/profile/Dan_Short/publication/255712425_The_public_understanding_of_science_30_years_of_the_Bodmer_report/links/57c3b8be08aeb95224dbe8e1.pdf
Stubbs, Michael 1996: *Text and corpus analysis. Computer-assisted studies of language and culture*. Oxford: Blackwell Publishers.
Tiedebarometri 2016. Tutkimus suomalaisten suhtautumisesta tieteeseen ja tieteellis-tekniseen kehitykseen 2016. Helsinki: Tieteen tiedotus ry.

Niina Kekki (University of Turku)

Korpusvetoinen tarkastelu synonyymisista adjektiiveista edistyneen suomenoppijan kielessä

Tässä esitelmässä tarkastelen korpusvetoisesti edistyneiden suomenoppijoiden synonyymisten adjektiivien käyttöä ensikielisiin suomenpuhujiin verrattuna. Edistyneellä suomenoppijalla tarkoitan sellaisia suomi toisena kielenä (S2) -oppijoita, joiden suomen kielen taso on eurooppalaisen viitekehyksen taitotasoilla B2–C2 (EVK 2013). Toisen kielen tutkimuksessa on huomattu, että toisen kielen käyttäjät voivat osata oppimaansa kieltä laadullisesti hyvin mutta tehdä käyttökontekstille epätyypillisiä kielellisiä valintoja (Ivaska 2015, 4). Tavoitteeni on selvittää, millaista tietoa kotekstitarkastelun ja frekvenssianalyysin avulla

saadaan niistä konstruktioista, joihin synonyymiset sanat kiinnittyvät S2- ja S1-suomessa. Vertaamalla tuloksia keskenään pyrin havainnoimaan, kuinka S2- ja S1-puhujat hahmottavat lähisynonyymien semanttista sisältöä. Tutkimalla synonyymisten adjektiivien käyttöä korpuslingvistisesti saadaan tietoa siis sekä S2- että S1-puhujien kielellisistä valinnoista.

Keskityn analyysissa synonyymisiin adjektiivipareihin suuri/iso, selkeä/selvä, hankala/vaikea, tärkeä/keskeinen ja niiden kotekstiin eli sanojen välittömässä läheisyydessä olevaan tekstiyhteyteen (vrt. Jantunen 2001). Aineisto on kerätty Turun yliopiston Edistyneiden suomenoppijoiden korpuksesta (LAS2, 345 955 sanetta), joka koostuu suomenoppijoiden akateemisissa yhteyksissä tuottamasta kirjallisesta materiaalista. Aineisto on käytettävissä Korp-hakuliittymän kautta. Vertailuaineistona korpuksessa on suomenkielisten opiskelijoiden opinnäytetöitä (LAS1, 103 848 sanetta, ei vielä käytössä Korpissa).

Analyysista ilmenee, että adjektiiveja käytetään vähemmän S2- kuin S1-aineistossa. Kun tarkastellaan valintaa kahden synonyymisen adjektiivin välillä, huomataan, että S2-puhujien valinnat poikkeavat jonkin verran S1-puhujien valinnoista, mikä voi selittyä akateemisen tekstilajin eritasoisella tuntemisella. Kotekstianalyysin perusteella S2-aineistossa synonyymisten adjektiivien suhteellinen frekvenssi on suurempi sellaisissa kiinteissä konstruktioissa, joissa S2-puhujat ovat oletettavasti nähneet niiden esiintyvän aiemmin, kun taas S1-aineistossa adjektiiveja käytetään vapaammin. Tämä tukee aiempia havaintoja oppijankielen kollokaatiotutkimuksesta (esim. Nesselhauf 2004). Eroa on myös adjektiivien vertailuasteiden käytössä: komparatiivi on yleisempi S1- kuin S2-aineistossa, kun taas superlatiivia käytetään jonkin verran enemmän S2- kuin S1-aineistossa. Näitä huomioita analysoimalla pohdin esitelmässäni, kuinka korpusvetoisella tarkastelulla saadaan edistyneen suomenoppijan kielenkäytöstä sellaista tietoa, joka on sovellettavissa paitsi toisen kielen oppimisprosessin tutkimukseen myös S2-opetukseen.

Lähteet

EVK 2013 = *Eurooppalainen viitekehys: Kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys*. (Common European Framework of Reference for Languages: Learning, Teaching, Assessment 2001.) Suom. Huttunen, Irma & Jaakkola, Hanna. 1.–6. painos. Sanoma Pro Oy, Helsinki
Ivaska, Ilmari 2015. *Edistyneen oppijansuomen konstruktiopiirteitä korpusvetoisesti: avainrakenneanalyysi*. Annales Universitatis Turkuensis. Sarja C osa 409. Scripta Lingua Fennica Edita. Turun yliopisto, Turku.
Jantunen, Jarmo 2001. ”Tärkeä seikka” ja ”keskeinen kysymys”: Mitä korpuslingvistinen analyysi paljastaa lähisynonyymeista? *Virittäjä* 2/2001, 170−192.
LAS2 = Edistyneiden suomenoppijoiden korpus, Turun yliopisto.
Nesselhauf, Nadja 2004. *Collocations in a Learner Corpus*. John Benjamins Publishing Company.

Kimmo Kettunen (National Library of Finland)

Kirjoitetun nykysuomen automaattisesta semanttisesta merkitsemisestä

Tässä julkaisussa esitellään FiST, työn alla oleva kirjoitetun nykysuomen kokotekstien semanttinen merkitsin. FiSTin ensimmäinen versio perustuu vapaasti saatavilla oleviin osiin: 46 226 sanan semanttiseen leksikkoon (Löfberg, 2017; Multilingual USAS) sekä morfologisen analyysin ohjelmiin Omorfiin ja FinnPosiin (Silfverberg ja kumppanit, 2016). FiSTin tämänhetkistä versiota on testattu systemaattisesti erilaisilla suomen aineistoilla, joista laajin on 45 miljoonan sanan elokuva- ja tv-tekstitysten OpenSubtitlesin osa-aineisto. FiST merkitsee teksteihin sanojen semanttisia luokkia noin 82–91 %:n sanastollisella kattavuudella (Kettunen, 2019). Toistaiseksi ohjelmasta puuttuu kaksi merkittävää osaa: semanttisesti monitulkintaisten sanojen käsittely (Robertson, 2019) sekä semanttisesta sanastosta puuttuvien yhdyssanojen kattava käsittely. Digitaalisen humanismin tutkimus perustuu yhdeltä osaltaan laajojen tekstiaineistojen tietokoneavusteiseen analyysiin – tällaiseen tutkimukseen FiST tarjoaa tutkijoille uuden mahdollisen työvälineen.

Timo Korkiakangas (University of Helsinki)

Handwriting quantification for historical linguistics

My poster presents and on-going project in which I quantify early medieval Italian scribes' handwriting to detect whether and how the scribes' competences to produce handwriting and Latin text are connected in extant scribal documents. My preliminary hypothesis is that the scribes' handwriting and linguistic competences are positively correlated with each other, given that both are important components of a successful scribal performance. The study exploits methods developed recently within the framework of digital palaeography and combines them with the corpus-linguistic methods applied to the Late Latin Charter Treebank (LLCT, v.2), a morpho-syntactically annotated dependency treebank which I have created under my previous projects. LLCT contains 1,040 scribal documents (480,000 words) from Tuscany of the 8th/9th century.

I will present preliminary results on how much variation certain characters display across the best and worst spelling scribes of LLCT. The spelling correctness values for each document and scribe derive from Korkiakangas (2017). The variation within the chosen characters will be compared not only to the variation in the spelling correctness, but also to the variation in the use of certain textual and layout features, such as abbreviations, punctuation, and line straightness, which all reflect the scribal performance. Later, the connection to a set of morphological and syntactic features will also be explored.

The main tool to be utilized is the Script Analyzer, an application originally designed by Rajan (2016) for the quantification and comparison of historical scripts. Script Analyzer consists of four modules: The Spline conversion module converts preprocessed images of single characters to a B-spline representation for computational analysis. Based on the B-splines, the Trajectory reconstruction module suggests five trajectories, in the order of their viability. Once a plausible trajectory is chosen, the algorithm proposes a stroke segmentation that is supposed to imitate the hand movement. The Feature extraction module presents some basic features of the character (disjoint/conjoint stroke count, retrace, up-/down-strokes, stroke lengths, size) as well as derived metrics (openness, compactness, breadth/width index, average curvature, disfluency, changeability, divergence, entropy, etc.).

Even at the current stage of the project, it has become obvious that the Script Analyzer succeeds in grasping at the visible reality by quantifying differences that are also perceived by the naked eye. The central intention is, however, to use the tool to detect features that escape the human observer. The preliminary results suggest that those scribes who knew Latin spelling also knew their job in terms of writing more consistent handwriting than those who made spelling mistakes.

Depending on the outcome of tests under way, two other potential tools for handwriting quantification (confidence matrices produced by a handwritten text recognition algorithm and Handwriting Analysis Tool HAT-2) will also be touched on.

References

Korkiakangas, T. 2017. Spelling Variation in Historical Text Corpora: The Case of Early Medieval Documentary Latin, in *DSH* 33, 575–591.
Mohammed, H. 2018. Handwriting Analysis Tool v2.0 (HAT-2). https://www.manuscript-cultures.unihamburg. de/hat.html
Rajan, V. 2016. Quantifying Scripts: Defining metrics of characters for quantitative and descriptive analysis, in *DSH* 32, 602–631.

Niina Kunnas (University of Oulu), Heidi Niemelä (University of Oulu) & Valtteri Skantsi (University of Oulu)

Kielimestari-applikaatio kielitieteellisessä tutkimuksessa ja tiedon välityksessä

Kerromme esitelmässämme kehittämästämme kielisovelluksesta, Kielimestarista, jonka ensimmäinen versio julkaistaan kevään 2019 aikana sovelluskaupoissa. Kielimestarin idea on syntynyt tarpeesta edistää suomalaisten kielitietoisuutta, lisätä heidän tuntemustaan maassamme puhuttavista vähemmistökielistä sekä uudistaa murteiden tutkimusta.

Kielet, joita Kielimestarissa opetellaan, ovat ruotsi, karjala ja pohjoissaame. Kielimestari tulee sisältämään nuorten suosimia tietovisoja sekä moderneja kielenoppimispelejä. Lisäksi sovelluksessa on mikrofonitoiminto, jonka avulla käyttäjä voi tallentaa omaa puhettaan järjestelmään. Tällä tavoin saamme kerättyä ajantasaista murreaineistoa. Kielimestari edistää kielentutkimusta, sillä osa sisällöistä suunnitellaan siten, että niiden vastauksia voidaan hyödyntää tutkimusaineistoina. Näin saamme crowdsourcing-menetelmällä aineistoa suomalaisten kielitietoisuudesta, -asenteista ja -käsityksistä.

Kielten oppimiseen tähtäävät sovellukset (esim. FluentU, Mondly) ovat tulleet maailmalla todella suosituiksi viime vuosina. Paitsi että niitä on tarjolla maailman suurten kielten opiskeluun, niitä on alettu kehittää myös uhanalaisten kielten säilyttämisen tueksi. Murteentutkimukseen tarkoitetut tiedeapplikaatiot (esim. English Dialects, Dialäkt Äpp) ovat pysytelleet maidensa ladatuimpina sovelluksina kuukausien ajan heti ilmestyttyään. Toistaiseksi Suomessa ei ole vielä kehitetty puhelinsovellusta, joka yhdistäisi kaksi edellä mainittua kategoriaa tai joka tähtäisi uhanalaisessa asemassa olevien kielten säilymiseen. Haluamme olla näin uuden teknologian avulla tukemassa sekä kielten säilymistä että lisäämässä niitä koskevaa tietoisuutta. Kielisovelluksemme tarjoaa oppimisen kokemuksia, ja sen kautta välitämme yleistajuisesti myös tutkimustuloksiamme.

Kehitämme Kielimestari-sovellusta yhdessä oululaisen Red Shirt Games -peliyrityksen kanssa. Yritys on erikoistunut viihteellisiin älypuhelinpeleihin, ja sen julkaisemat sovellukset ovat saaneet erinomaisia pisteytyksiä käyttäjiltään. Esitelmässämme kerromme Kielimestarin kehittämisestä havainnekuvin ja esittelemme sovelluksen ensimmäisen version, johon sisältyy yksi kielenoppimispeli neljälle eri kielelle. Luomme myös katsauksen siihen, miten sovelluksen tuoreempien versioiden olisi tarkoitus toimia ja miten sovellusta hyödynnetään kielentutkimuksessa.

Tommi Kurki (University of Turku), Nobufumi Inaba (University of Turku), Annekatrin Kaivapalu (University of Turku), Maarit Koponen (University of Turku), Christophe Leblay (University of Turku), Veronika Laippala (University of Turku), Jorma Luutonen (University of Turku), Maarit Mutta (University of Turku), Markku Nikulin (University of Turku) & Elisa Reunanen (University of Turku)

Digilang: Turun yliopiston digitaalisia kieliaineistoja kehittämässä

Turun yliopiston kieli- ja käännöstieteiden laitoksen Digilang-hankkeessa täydennetään ja kehitetään laitoksen digitaalisia kieliaineistoja. Samalla kieliaineistojen näkyvyyttä lisätään keräämällä ne yhteen ja luomalla yhteinen käyttäjäportaali, jonka avulla tutkijat ja opiskelijat löytävät entistäkin paremmin tarvitsemiaan aineistoja. Yliopiston rehtori on myöntänyt hankkeelle 580 000 euroa aineistojen kehitystyöhön, ja hanke toimii vuosina 2018–2021.

Turun yliopiston kieliaineissa on koostettu, kehitetty ja ylläpidetty digitaalisia aineistoa tutkimuksen tarpeisiin vuodesta 1967, jolloin suomen kielen oppiaineen yhteyteen perustettiin Lauseopin arkisto (LA). LA:n murrekorpus on Suomen ensimmäinen digitaalinen annotoitu kieliaineisto. Varsinkin viime vuosikymmenten aikana alkuperäisen murrekorpuksen rinnalle on kieliaineissa luotu useita muita korpuksia.

Digilang-hankkeessa parannetaan nykyisten aineistojen käytettävyyttä kehittämällä niiden ns. metatietoja, kun esimerkiksi kunkin sanan, lauseen, virkkeen, intonaatiojakson ja diskurssin rakenteesta ja visualisoinnista lisätään tietoja. Näin aineiston käyttäjät pystyvät löytämään helpommin yhä useammasta laajasta puhe- tai tekstimassasta tarvitsemansa tapaukset. Kieliaineistojen saavutettavuutta ja näkyvyyttä lisätään keräämällä ne yhteen ja luomalla yhteinen käyttäjäportaali, jonka avulla tutkijat ja opiskelijat löytävät entistäkin paremmin tarvitsemiaan aineistoja ja saattavat samalla löytää heille entuudestaan tuntemattomia mutta hyödyllisiä aineistoja.

Ensi vaiheessa Digilang-hankkeessa yhdistyy kuusi TY:ssa eri tahoilla kehitettyä kieliaineistoa: Lauseopin arkiston (LA) aineistoja, muita suomen kielen ja suomalais-ugrilaisen kielentutkimuksen oppiaineen kieliaineistoja, kielentutkimusta ja kieliteknologiaa yhdistävän TurkuNLP-ryhmän kehittämiä Universal Parsebanks -aineistoja sekä eri kielten ja kääntämisen tutkijoiden LOG-aineisto.

a) Satakuntalaisuus puheessa -korpus (Sapu; 246 tuntia nykysatakuntalaisen puhekielen äänitteitä, joista 210 tunnista kielitieteelliset transkriptiot, annotoimaton aineisto, osa Lauseopin arkistoa)

b) Suomen kielen prosodian alueellisen ja sosiaalisen variaation korpus (Prosovar; n. 430 puhujalta n. 5700 prosodista äänitekatkelmaa; annotoimaton aineisto; osa Lauseopin arkistoa)

c) Fennougristiset korpukset: Volgan alueen kielten tutkimusyksikön morfologisesti annotoitu mordvalaiskielten korpus Mormula (n. 35 000 virkettä); annotoimattomat tekstikorpukset 7 kielestä (ersä, moksha, mari, udmurtti, komipermjakki, tshuvassi, tataari; yht. n. 17 milj. sanaa); kirjakielen historian korpukset (mari, mordvalaiskielet, yht. n. 1000 tekstiä); paralleelitekstikorpukset (2 tekstiä, 14 kieltä).

d) Akateemisen suomen korpus, joka on rakennettu TY:n strategisen rahoituksen turvin (Edistyneiden suomen oppijoiden korpus ja Ensikielisten suomalaisten akateemisten tekstien korpus sekä nyt koostettava tutkimusartikkelikorpus; osa Lauseopin arkistoa).

e) Universal Parsebanks (UP), joka on 45 erikielistä Internetistä koneellisesti kerättyä aineistoa sisältävä datakokoelma automaattisesti syntaksijäsennettynä. Kielikohtaisten aineistojen koot ovat miljardeja sanoja, ja ne ovat vapaasti käytettävissä. Kehittynein niistä on suomenkielinen Finnish Internet Parsebank.

f) LOG-aineisto, jonka avulla kirjoittamista ja kääntämistä voi tarkastella prosessilähtöisesti: millaisessa prosessissa teksti syntyy, ja esimerkiksi missä järjestyksessä sen osat tuotetaan ja miten sitä muokataan. Aineisto koostuu nyt yhdistettävistä eri tutkijoiden eri kielillä keräämistä tuotoksista.

Leo Lahti (University of Turku), Ville Vaara (University of Helsinki), Jani Marjanen (University of Helsinki) & Mikko Tolonen (University of Helsinki)

Best Practices in Bibliographic Data Science

Bibliographic data science is an emerging field in digital humanities that takes advantage of modern data science in order to facilitate the research use of large-scale bibliographic metadata collections. The conceptual and methodological basis of the field is shaping up, and we aim to clarify some of the key methodological aspects in using library catalogues as quantitative research material. We demonstrate these issues based on four bibliographic collections, including the Finnish and Swedish national bibliographices, the English Short Title Catalogue, and the Heritage of the Printed Book Database. We describe the open bibliographic data science ecosystem that we have designed and implemented to support data harmonization, analysis, and interpretation. We demonstrate how our algorithmic approach supports semi-automated curation and scaling up the research to metadata collections that contain millions of bibliographic records. The analysis highlights the opportunities and prevailing challenges in the

field, and provides examples of metadata-driven harmonization and analysis approach that is generally more widely applicable in related studies in the digital humanities.

Jukka Mettovaara (University of Oulu)

Inari Saami language technology: an overview

Inari or Aanaar Saami (anarâškielâ) is the only Saami language spoken exclusively in Finland, mainly in the areas surrounding Lake Inari. The number of speakers is estimated to be around 350–450, and from the 1950s Inari Saami started rapidly approaching severe endangerment. However, at the beginning of the 1990s, the language community began goal-directed revitalization efforts and were able to stop the language's downfall. The revitalization of Inari Saami has been exceptionally successful among the language revitalization projects of the world: the language has been restored as the language of everyday life in the community and taken over domains it never held before, such as scientific literature. A large part of the success can be attributed to meticulous revitalization planning and thorough assessment of prevailing needs. (Olthuis et al. 2013; Pasanen 2015.)

The age of digital information has undoubtedly accelerated and facilitated the revitalization of many indigenous and minority languages. The development of language technological tools for Saami languages began almost 20 years ago in Giellatekno, Center for Saami language technology, and Divvun, a project and a working group developing Saami language proofing tools, both now part of the University of Tromsø. (Giellatekno 2019.) At the core of Inari Saami language technology is the computational model of inflection, a program that is able to parse a word form and deliver its grammatical analysis and, conversely, generate the entire paradigm of a given word. Inari Saami has a complicated nominal, verbal and derivational morphology with many types of vowel and consonant alternations, which means that tools that can recognize and generate word forms are essential. (Olthuis & Trosterud 2015.)

In my presentation, I will provide a concise overview of the situation of Inari Saami language tools in 2019 and tackle some of the general shortcomings of the tools from the viewpoint of a language learner, user and teacher. I will also present the results of a short survey of Inari Saami language students on what language tools and resources they use and their opinions on them. This is done in order to assess whether there is a need to advertise the language tools more, or provide the speakers and learners more instruction on how to use them. The preliminary results show, for example, that the proofing tool Divvun that is available at the moment in five Saami languages (the Inari Saami version is in beta) for Microsoft Office and LibreOffice remains very underutilized.

References:
Giellatekno 2019: Giellatekno birra. – Giellatekno. giellatekno.uit.no/Giellatekno.sme.html. Last published 14.3.2019, retrieved 17.3.2019.
Olthuis, Marja-Liisa – Kivelä, Suvi – Skutnabb-Kangas, Tove 2013*: Revitalising Indigenous Languages. How to Recreate a Lost Generation*. Bristol: Multilingual Matters.
Olthuis, Marja-Liisa – Trosterud, Trond 2015: Inarinsaamen lingvistinen suunnittelu kieliteknologian valossa. – *AGON*. agon.fi/article/inarinsaamen-lingvistinen-suunnittelu-kieliteknologian-valossa/. Published 12.4.2015, retrieved 17.3.2019.
Pasanen, Annika 2015: *Kuávsui já peeivičuovâ 'Sarastus ja päivänvalo': Inarinsaamen kielen revitalisaatio*. Uralica Helsingiensia 9. Helsinki: Helsingin yliopisto, Suomen kielen, suomalais-ugrilaisten ja pohjoismaisten kielten ja kirjallisuuksien laitos: Suomalais-ugrilainen seura.

Mikhail Mikhailov (Tampere University)

The Extent of Similarity: comparing texts by their frequency lists

Measuring distances between texts can be useful in document search, automated classification, detecting plagiarism, etc. (Hoad, T. C. and Zobel, J., 2003, Gomaa & Fahmy 2013). One of the possible ways to do it is to compare frequency lists of the texts (Kilgarriff 1997, Piperski 2018). In this paper, two methods of such comparison are discussed: the first one is based on top X items from normalized frequency lists and the second compares frequency lists of N random samples of X running words per text.

The research data were literary texts in Russian and in Finnish, original texts and translations. The data included different authors, different genres and translations by different translators. During the first experiment the top 1000 words from lemmatized frequency lists were compared. The frequency lists were merged into a single data frame, a distance matrix was calculated, and finally the cluster analysis was run on the matrix. For the second experiment, 50 random samples of 3000 running words were drawn from each text and the frequency lists of these samples were processed in the similar way as in the first experiment.

The results for both methods were quite satisfactory. The texts by the same authors and texts devoted to the same topics were often clustered together. All retranslations of the same texts were clustered to the same groups. However, the time period, the translator and the source language did not seem to influence the result much. The terminal clusters were more interesting than the upper-level classification. Obviously, a human would have defined the larger groups in a different way. At the same time, this computer-made classification might bring new insights for the researchers.

References:
Gomaa, W.H. and Fahmy, A.A., 2013. A Survey of Text Similarity Approaches. *International Journal of Computer Applications* (0975 –8887). Volume 68–No.13.
<https://research.ijcaonline.org/volume68/number13/pxc3887118.pdf>
Kilgarriff A., 1997. *Using word frequency lists to measure corpus homogeneity and similarity between corpora*. <http://aclweb.org/anthology/W97-0122>.
Kilgarriff A., 2001. Comparing corpora, *International Journal of Corpus Linguistics*, 6(1), pp. 97–133.
<https://www.sketchengine.eu/wp-content/uploads/comparing_corpora_ 2001.pdf>
Piperski, A. Ch., 2018. Corpus size and the robustness of measures of corpus distance. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018"*. Moscow, May 30—June 2, 2018. Moscow, RGGU. http://www.dialog-21.ru/media/4327/piperskiach.pdf

Niko Partanen (University of Helsinki) & Michael Rießler (University of Helsinki)

Automatic validation and processing of ELAN corpora for spoken language data

This presentation demonstrates and discusses workflows developed during the authors' research with spoken corpora of various endangered languages of Northern Eurasia. In our projects, the spoken language material has been transcribed and translated in ELAN, resulting in parallel corpora with aligned audio (and in some cases video). Deeper linguistic annotations, such as morphosyntactic tagging, are also stored in the same files. ELAN is an open source tool that has become a quasi-standard for annotating spoken data in fieldwork-based documentation of endangered languages. For automatic parsing and tagging we have programmed rule-based analysers using the open source infrastructure of Giellatekno (Trosterud 2006), for which we have successfully developed interaction with ELAN (Authors).

This language technology-oriented practice has enabled our projects to focus on the systematization of primary data resources and their transcriptions and translations. Once the transcriptions are edited, the existing morphosyntactic annotations are overwritten with a new analysis matching with the current state of our parser. This solves, to a certain degree, one common problem in spoken language corpora of similar kinds: the annotations at the sentence and word level are hierarchically connected to each other, and editing transcriptions at higher levels will inevitably lead to changes at other levels. Maintaining this is very challenging and easily leads to inconsistencies in annotation.

Furthermore, ELAN allows a relatively lax and manually editable structure, which easily leads to very differently structured files. The software does not report on inconsistencies, which may result in difficulties when attempting to apply corpus queries across multiple ELAN files. The solution we have adapted has been to develop systematic testing scripts that verify that all files in the corpus actually do conform follow the same set of principles. In the case of ELAN, these are related to tier types, tier names and their combinations. This way, errors in the corpus structure can be detected immediately. Eventually, the testing should be conducted with the tools of continuous integration with automatic error reporting.

Besides validation, similar approaches can be applied to corpus parsing. Automatic parsing of corpus annotations and merging them with available metadata allows very rapid analysis of corpus content and structure, which also has numerous benefits when it comes to error detection. Presenting the corpus in a logical data structure within programming languages such as R and Python makes it very easy to convert the corpus into other formats, while also forcing the researchers to be aware of the machine readability of used annotation schemes.

The existence of a variety of annotation schemes and ELAN tier structures makes it very difficult to reuse existing tools for new projects. Creating tools for generalized use beyond the concrete conventions of one project poses a challenge, and often the only feasible approach seems to be to rewrite everything. However, we not only believe that common solutions can be found, but that they would benefit the field at large. Nevertheless, the discussion on how to achieve this is only just getting started. Our presentation is an attempt to progress this development.

Niko Partanen (University of Helsinki), Michael Rießler (University of Helsinki) & Joshua Wilbur (University of Freiburg)

Integrating language technology into work with spoken corpora

Our paper presents on-going work in several language documentation projects on endangered languages spoken in northern Fenno-Scandia and northern Russia. One feature of our work in comparison to many other similar documentation projects has been the systematic application of language technology in order to avoid manual corpus annotation (Authors). This is crucial, since the corpora we work with are so large that manual work can rarely annotate more than a small fraction of the whole corpus. These limitations are present in all parts of the workflow, except potentially the recording process itself. The systematic use of language technology is a logical solution to this manual-work bottleneck. However, its implementation is not necessarily always straightforward, so the approach we propose requires closer collaboration between researchers developing tools for Natural Language Processing and linguists.

Our aim is to use the same system of analysis for spoken and written language, although they both come with their own set of particularities, mainly stemming from the original sources and formats. This requires specific choices concerning spoken language annotation, primarily the use of transcription systems that are compatible with or adaptable to existing tools. However, in our experience the use of contemporary orthographies also works very well with spoken language, especially when transcription accuracy is mainly at the phoneme level.

A technical framework that we successfully integrate into our workflows is EMU (Winkelmann et. al. 2019), which allows reasonably accurate phoneme and word level segmentation from existing utterance-level annotations. From this point of view transcribing utterance level annotations is more than sufficient, as more accurate levels can be derived automatically.

For linguistic annotation we rely primarily on Giellatekno tools (Trosterud 2006). Since these tools exist for, inter alia, a variety of Saamic languages, Finnish and Norwegian, researchers working on northern Eurasian languages should consider whether these tools can be integrated into their workflows. Our approach has been to input analyses as annotations directly added into ELAN files using Python (Authors 2017), and the tools have been implemented in later versions as an external web service using uralicNLP package (Authors 2019; Hämäläinen 2018). We have also conducted tests with various dependency parsers (Authors 2018), but in our low-resource scenario these have not (yet) resulted in viable solutions.

Our third line of work involves speech recognition. The most exciting of these approaches has been the translation of Mozilla's Common Voice platform into languages we work with, and we aim to use this tool to collect larger amounts of spoken data than is possible with normal means of recording and transcription. Since several projects have already reported successful results with speech recognition on endangered languages (Adams 2018; Foley et al 2018), it is only a matter of time before this will be available for our languages. However, since the magnitude of resources needed are beyond what one project can reasonably work with, wider collaboration and data sharing between research projects and institutions will be needed.

References:
(Authors' own papers left out for anonymity)

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird and Alexis Michaud 2018. Evaluating Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation. In *Proceedings of LREC 2018*.

B. Foley, J. Arnold, R. Coto-Solano, G. Durantin, T. M. Ellison, D. van Esch, S. Heath, F. Kratochvíl, Z. Maxwell-Smith, D. Nash, O. Olsson, M. Richards, N. San, H. Stoakes, N. Thieberger and J. Wiles 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (Elpis). In S. S. Agrawal (Ed.*), The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages* (SLTU). 200–204.

Mika Hämäläinen 2018. *UralicNLP* (Version v1.0). Zenodo. http://doi.org/10.5281/zenodo.1143638

Trond Trosterud 2006. Grammatically based language technology for minority languages. In: *Lesser-known languages of South Asia: Status and policies, case studies and applications of information technology*, ed. by Anju Saxena & Lars Borin. Berlin: Mouton de Gruyter, 293–316.

Raphael Winkelmann, Klaus Jaensch, Steve Cassidy and Jonathan Harrington 2019. *emuR: Main Package of the EMU Speech Database Management System*. R package version 1.1.2.

Tuula Pääkkönen (National Library of Finland)

Digital heritage presentation system development + new material types: early findings

Recently the National Library of Finland initiated a development project in order to allow showing also book materials in the presentation and content search system of https://digi.nationallibrary.fi also known as Digi.

This project, nicknamed “Books to Digi” aimed to streamline the digitization process so, that the books could be imported to the presentation system directly after post-processing phase of the digitization.

However, beside the new production process support, there were number of end-user and stakeholder needs that need to be considered or alleviated. We analysed the requirements, identified user segments and condensed the requirements to the core set, for the implementation. Use cases, Pugh matrix (Pugh, 1991) and agile development were used in the system development in order to keep development running and to keep tackling one task after another.

Content, requirements aside, and then third thing to master was the metadata flow within the library. The metadata of the national bibliography uses standardized rules for metadata creation, which have changed throughout the decades and even centuries. For an information system, this requires some needs for normalization of some of the metadata, where we targeted to the publisher names and the publication places in order to make search approachable for end-users. This enrichment of contents was also seen as potential improvement, which in the long run could also be extended to the digitized ephemera, which would then be similarly well-organized as the digitized newspapers and journals.

The project is currently in a beta phase, where the end-users can see the changes and give feedback. One survey was made on 2018 (Pääkkönen & Kettunen, 2018) to get a baseline of user aptitude towards Digi and a new survey will be done on 2019, to evaluate the impact of the new materials and search capabilities. We hope that this project enables people to do interesting findings via full text search capabilities towards the digitized contents.

Acknowledgements
Part of this work is funded by the Academy of Finland project COMHIS – Computational History and the Transformation of Public Discourse in Finland, 1640–1910, decision number 293341.
The development project wants to thank all end-users, colleagues and stakeholders for their feedback and support throughout the project.

References:
Pääkkönen, T., & Kettunen, K. (2018). Kansalliskirjaston sanomalehtiaineistot: Käyttäjät ja tutkijat kesällä 2018. *Informaatiotutkimuksen Päivät 2018*, Retrieved from http://urn.fi/URN:NBN:fi-fe2018110247067

Pugh, S. (1991). *Total design: Integrated methods for successful product engineering*. Wokingham, England; Reading, Mass: Addison-Wesley.

Tuula Pääkkönen (National Library of Finland), Kimmo Kettunen (National Library of Finland) & Jukka Kervinen (National Library of Finland)

Search Options Used in Digitized Serial Publications - Observational User Data and Future Challenges

Easy access to digital data resources is one of the key components of successful data intensive Digital Humanities research. Despite increased use of programming languages, web data services and different digital tools, there is increasing demand for researcher friendly tools that are close to the actual materials.

In the digitized newspaper collection of the National Library of Finland (digi.nationallibrary.fi) there has been continuous efforts to incorporate the internal needs to the functionalities, which are offered to end-users. Therefore during the last development project we utilized user survey and log data to help us to design features for the future.

This paper discusses the key findings of both a user survey and gathered user log data of a digitized historical material Web collection. The aim of the discussion is to pinpoint the most salient features of an interface that users of the digitized historical newspaper and journal collection use. Inclusion of new type of data, digitized books, will bring new challenges to the planning and design of a user interface so that it would serve Digital humanities researchers as well as possible.

Sierge Rasmus (University of Oulu, Giellagas Institute)

Current Situation of the North Saami and Language Technological Tools

North Saami is largest of the Saami languages and it has estimated 20 000 speakers. Most of them are inhabited in Norway and Sweden. Roughly 10 per cent of North Saami speakers live in Finland. Saami languages are protected by law in all these nations, but it concerns only traditional regions of Saami people. North Saami has education from preschool all the way to university, but elementary and secondary schools are also in the traditional Saami region (municipalities of Utsjoki, Inari and Enontekiö, and the northern part of Sodankylä).

Because of the modern trend, where people move from countryside to cities, many of the Saamis don't have equal rights to use their language with authorities or get education in their mother tongue. Technological development can partly enable schooling in Saami languages with distance lectures and electronic materials but it can't fully replace in-class-teacher especially with small children.

In North Saami there are many technological resources available. Many of them are developed by Giellatekno in the Arctic university of Norway (UiT), for example educational websites, proofing tools, word form analysers and generators, corpus based on biblical texts, literature and newspapers, web dictionaries and a speech synthesizer.

In this presentation I view the findings of the survey done on North Saami learners. We (Marko Jouste, Markus Juutinen, Jukka Mettovaara, Marko Marjomaa, Jack Reuter and I, Sierge Rasmus) collected survey data from Skolt Saami, Inari Saami and North Saami students in secondary and higher degree education in Finland. The inquiry asked for example how often students used text books or smartphones in learning the Saami language and what kind of technological resources they used. Preliminary findings show that mobile solutions are more often used than printed books. Even though information technology seems to play an important role in language learning, the vast archived materials are rarely used for learning purposes. Survey data also indicates that teachers' awareness of the technological tools vary, because the survey answers from different schools showed significant variations on how much students use information technology to learn the language.

Toni Ryynänen (University of Helsinki, Ruralia Institute) & Torsti Hyyryläinen (University of Helsinki, Ruralia Institute)

Border Crossing and Trespassing? Expanding Digital Humanities Research to Developing Peripheries with the Novel Digital Technologies

Definitions of and perspectives to Digital Humanities (DH) research tend to deviate amongst the disciplines involved. Typically, DH refers to the application of novel technology and methods in the humanities and social sciences (HSS) research: the usage of data in the context of computational science, collaborative effort combining the expertise from humanities and da-ta sciences as well as examination of digitalisation as a cultural and social phenomenon. We propose an expansion for DH research by discussing an on-going research project funded by the EU's Northern Periphery and Arctic 2014–2020 programme. The Emergreen

project (2018–2021) is utilised here as an illustrative case: is there a role for development-oriented research based on the local needs aiming at producing practical and transnational technological solutions for the stakeholders' (end-users', consumers', companies', the public sector actors') real needs? The article expands the current method discussions about DH research emphasis-ing a need for the future oriented and practically relevant research and development methods. In this context, the digitalisation of the society is analysed from the humanistic perspective aiming at understanding the needs of the public services development. In the conclusions, we propose a concept of "practical digital humanities" for describing research utilising a humanist approach to practical problem solving with digital technology development in the DH context.

Juhana Salonen (Jyväskylän yliopisto, viittomakielen keskus), Anna Puupponen (Jyväskylän yliopisto, viittomakielen keskus) & Tommi Jantunen (Jyväskylän yliopisto, viittomakielen keskus)

Suomen viittomakielten korpusta rakentamassa

Viittomakielikorpusten rakentaminen on lisääntynyt merkittävästi 2000-luvulla: ensimmäiset korpusprojektit käynnistyivät 2000-luvun alussa Australiassa ja Hollannissa, minkä myötä laajoja, koneluettavia aineistokokoelmia on ryhdytty rakentamaan useissa Euroopan maissa 2010-luvulla. Tässä artikkelissa tarkastellaan Suomen viittomakielten, suomalaisen ja suomenruotsalaisen viittomakielen, korpuksen syntyä. Artikkeli esittelee korpuksen rakennusvaiheita eli aineiston keräämistä, käsittelyä, annotointia, pitkäaikaissäilytystä sekä julkaisua tietosuojakysymyksineen. Lisäksi artikkelissa kuvaillaan, miten korpusaineistoa on käytetty ja voidaan hyödyntää viittomakielten tutkimuksessa sekä opetuksessa.

Neljän vuoden mittainen Suomen viittomakielten korpusprojekti käynnistyi Jyväskylän yliopiston viittomakielen keskuksessa vuonna 2014. Projektin aikana kuvattiin keskusteluja ja elisitoituja kertomuksia 91 suomalaista viittomakieltä ja 12 suomenruotsalaista viittomakieltä äidinkielenään käyttävältä, eri puolilla Suomea asuvalta henkilöltä viittomakielisen kuuron projektitutkijan opastuksella. Videomateriaalia kerättiin yhteensä noin 560 tunnin edestä (seitsemästä kamerakulmasta nauhoitetut materiaalit yhteenlaskettuna).

Aineistonkeruun ja editoinnin jälkeen yhteensä 22 suomalaista viittomakieltä äidinkielenään käyttävän kielenoppaan videoaineistoihin on tehty perustason annotaatiot viittoma- ja virketasolla. Annotointivaihe eteni viittomien tunnistamisella, niiden merkitysten erottamisella ja viitotun tekstin ilmauskokonaisuuksien kääntämisellä suomen kielelle. Perusannotointi toteutettiin ELAN-ohjelmalla, jossa viittomia identifioidaan ajallisesti videoon yhteydessä olevien glossien avulla. Annotoinnissa käytettiin lisäksi Suomen Signbank -leksikkotietokantaa, johon ELAN-ohjelman glossit yhdistyvät verkkoyhteyden avulla. Laaja multimodaalinen aineistokokonaisuus täydennettiin metatiedoilla aineiston eri osa-alueista, kuten aineistokokonaisuuden yleisluonteesta, aineistonkeruussa läsnä olleista henkilöistä, videoiden sisällöistä ja video- ja annotaatiotiedostojen muodoista IMDI (ISLE Meta Data Initiative) -standardin mukaisesti. Annotoitu aineisto säilytetään ensisijaisesti Jyväskylän yliopistossa, minkä lisäksi se siirretään maaliskuun 2019 aikana FIN-CLARIN-konsortion Kielipankkiin pitkäaikaissäilytettäväksi sekä julkaistavaksi kielenoppaiden tutkimussuostumusten ja tietosuoja-asetusten mukaisesti. Kielipankissa julkaistava korpusaineisto sisältää noin 14 tunnin edestä kuudesta kamerakulmasta kuvattua videomateriaalia 21 kielenoppaalta sekä videoihin linkitetyt annotaatiotiedostot ja IMDI-kuvaukset.

Suomen viittomakielten korpuksen luonti kehittää molempien viittomakielten kielellisten ja kulttuuristen piirteiden tutkimusta sekä opetusta. Jyväskylän yliopiston viittomakielen keskuksessa korpusaineiston pohjalta on tehty tähän mennessä useita suomalaiseen viittomakieleen keskittyviä tutkimuksia, minkä lisäksi aineistoa on käytetty myös viittomakieliä vertailevassa tutkimuksessa. Kerätty videoaineisto on ainutlaatuinen kokoelma Suomen viittomakielillä tuotettua kerrontaa ja keskusteluja: materiaali sisältää eri-ikäisten ja eri alueilta tulevien henkilöiden viittomista erilaisissa viestintätilanteissa. Systemaattisen

annotoinnin myötä aineisto tulee olemaan merkittävä resurssi tutkimuksen lisäksi viittomakielten opetuksessa, viittomakieliä koskevassa koulutuksessa sekä kielisuunnittelussa.

Maija Saviniemi (University of Oulu)

Iijoki-sarja korpuksena

Oulun yliopiston suomen kielen oppiaine ja Kielipankki ovat parhaillaan muodostamassa ”Iijoki-sarja, Oulun yliopiston Päätalo-kokoelma” -nimistä korpusta. Se on tarkoitus julkistaa kunniatohtorimme 100-vuotissyntymäpäivän kynnyksellä 8.11.2019, jolloin Oulun yliopisto järjestää eri alojen tutkimusta yhdistävän ”Kalle Päätalo tutkijoiden silmin” -symposiumin.

Tekeillä oleva korpus sisältää professori, kirjailija Kalle Päätalon (11.11.1919–20.11.2000) kirjoittaman Iijoki-sarjan 26 romaania, joissa kirjailija kertoo kotiseudustaan ja elämänvaiheistaan 1910–1990-luvuilla. Yhtenä tavoitteena kirjailijalla on ollut kotiseutunsa eli Koillismaan murteen tallentaminen. Iijoki-sarjaan on taltioitu myös monia muita Suomen murteita sekä niitä kommentoivaa metakieltä.

Iijoki-sarjan yhteenlaskettu sivumäärä on yli 17 000. Ensimmäinen osa ”Huonemiehen poika” ilmestyi vuonna 1971 ja viimeinen ”Pölhökanto Iijoen törmässä” 1998. Kielipankkiin tallennettavan korpuksen ensimmäisen romaanin teksti on nyt tokenisoitavana ja jäsennettävänä, minkä jälkeen käsitellään koko aineisto. Korpus on aikanaan käytettävissä Aca-lisenssillä.

Taivalkoskella syntyneen ja Tampereella kirjailijanuransa tehneen Päätalon romaanisarja on omaelämäkerrallinen, ja sitä voidaan pitää maailman suurimpana romaanina. Sarjan keskeinen teema on yksilön sosiaalinen nousu: Päätalo aloittaa tukkijätkänä, työskentelee muun muassa rakennusmestarina ja päätyy lopulta arvostetuksi kirjailijaksi.

Posterissa esitellään Iijoki-sarjaa korpusaineistona ja hahmotellaan muutamia tutkimusaiheita, joita korpuksen valmistuminen mahdollistaa. Olen parhaillani tekemässä tutkimusta fiktiivisen Hermanni Päätalon kielestä Loimujen aikaan -teoksessa, ja aikanaan korpusaineisto mahdollistaa tämän teoksen lähiluvusta liikkeelle lähteneen analyysin laajentamisen vaikkapa koko teossarjaan.

Mari Siiroinen (University of Helsinki)

Vanhojen opinnäytteiden aarteet. Helsingin yliopisto suomen kielen pro gradut 1881–1949

1880-luvusta alkaen on Keisarillisessa Aleksanterin yliopistossa, sittemmin Helsingin yliopistossa maisterintutkinnon vaatimuksiin kuulunut "itsenäinen tieteellinen tutkimuskoe", pro gradu –koe.

Suurin osa Helsingin yliopiston vanhoista pro gradu –tutkielmista löytyy kirjastosta. Aivan kaikki eivät siellä kuitenkaan ole. Esimerkiksi suomen kielen ja sen sukukielten tutkielmat ajanjaksolta 1881–1985 ovat historiallisista syistä edelleen tiedekunnan omissa tiloissa, eivätkä ne löydy kirjaston tietokannoista. Monet vanhat tutkielmat nukkuvat turhaan ruususen unta, vaikka niissä olisi paljon nykypäivänäkin relevanttia tietoa. Monesta aiheesta ensimmäinen tutkimus on pro gradu -työ.

Käsittelen esitelmässäni Helsingin yliopistossa ajanjaksolla 1881–1949 valmistuneita suomen kielen graduja, joita on nyt alettu digitoida. Luon esitelmässäni katsauksen tutkielmien aiheisiin ja niiden kehitykseen, tekijöiden sukupuolijakaumaan sekä otan esille mielenkiintoisia turhaan unohduksiin painuneita aarteita tutkielmien joukosta. Digitoidut opinnäytteet ovat paitsi tutkimusjulkaisuja myös tutkimusaineistoa tieteenhistorioitsijoille.

Saana Svärd (University of Helsinki), Heidi Jauhiainen (University of Helsinki), Aleksi Sahala (University of Helsinki), Tero Alstola (University of Helsinki), Tommi Jauhiainen (University of Helsinki) &
Krister Lindén (University of Helsinki)

Oracc in Korp

Open Richly Annotated Cuneiform Corpus (Oracc) is an international cooperative undertaking providing free online editions of texts written mostly in the Akkadian and Sumerian languages. Akkadian is an East Semitic language whereas Sumerian is a language isolate for which the cuneiform script was originally created. These languages were spoken and written in ancient Mesopotamia, modern-day Iraq, five to two thousand years ago. The text corpora in Oracc have been created by various projects and it is one of the largest electronic resources of Akkadian and Sumerian texts.

Oracc is a valuable tool for an Assyriologist who works with Akkadian and Sumerian texts, but it has been relatively difficult to do searches across all the different Oracc projects. However, when studying the semantic contexts of a word, it is imperative to have all the attestations of the word at hand. An efficient corpus search tool is therefore a necessity. We have added the texts in Oracc to Korp, an online service provided by the Language Bank of Finland. Korp is openly available to all and allows users to make queries in text corpora. The results of the query are presented as concordances, i.e. listing every instance of the word matching the query with all its neighbors line by line. Furthermore, the results are, by default, presented in the so-called KWIC (keyword in context) view, meaning that each result is displayed with the matching words highlighted in the middle of the display. Korp is a useful tool for studying the contexts in which words appear.

In this poster, we present Oracc in Korp. To accommodate the special features of the Oracc data, we have added new fields that are not present in the corpora of other languages in Korp. The ways in which words in different Oracc projects have been annotated differ somewhat. Therefore, we have normalized the dictionary forms of some word classes. We have also unified synonymous translations of words and, subsequently, distinguished homonymous dictionary forms using the translations. Finally, we have added a feature in Korp which allows the user to search for a word using these normalized forms.

Riikka Taavetti (University of Helsinki)

Finns and Foreigners: An Experiment of Textual Analysis with an Archive of Sexual Life Stories

In year 1992, Finnish sociologists Osmo Kontula and Elina Haavio-Mannila gathered a collection of sexual life stories. All together 175 Finns responded the open call for writing that was published in newspapers and magazines. In their life stories, the writers of different ages and backgrounds revealed in varying ways, but often in great detail, their sexual experiences, fantasies as well as disappointments and even experiences of violence and abuse. The material utilised for this study consists of the 149 digitalised sexual life stories that are currently available for research.

In my presentation, I focus on how the authors of the life stories constructed their understandings of Finns and Finnishness with references to their experiences and impressions of foreigners. I will study my material by testing the possibilities of utilising digital text analysis tools with life stories that vary significantly in their forms of writing. My intention is to experiment with different methods for textual analysis to discover whether these tools could reveal something else than a traditional close reading of the life stories does.

The anonymization of these life stories plays an important role for my analysis. As the life stories were archived and made available for future research, they were anonymized and most of the reference to the nationality of people described and the names of the countries the authors have visited were removed.

Therefore, my experimentation with this material also discusses the obstacles caused for the digital analysis of the text by the anonymization. Moreover, I explore the possible new topics of research that the traces of the anonymization process in the texts might offer.

Stefan Veleski (Masaryk University)

"'Doomed' from the Start: The Role of 'Cultural Pollutants' in the 'Cultural Death' of Late Victorian Bestsellers"

This short presentation will apply computational methods in the analysis of the process of divergence that placed some late Victorian novels in the literary canon (thus exposing them to future generations of readers) while pushing others outside of the public eye – a development that could be termed "cultural death". This presentation argues that the large quantity of "cultural pollutants" in the textual fabric of late Victorian bestsellers is a significant factor in this divergence. "Cultural pollutants" are inherent properties of the text linked to the zeitgeist of a particular period that block its vertical cultural transmission across time to new generations of readers – they neatly fit the cultural tastes of a certain period or generation but prevent the novel to thrive when these tastes change. As such, they are direct opposites to Dan Sperber's "cultural attractors", aspects of the cultural product that facilitate its transmission.

The presentation will deal with the following "cultural pollutants": (1) The ratio of "moralizing" (narrative interventions in the text usually seeped in contemporary moral values, which might drastically differ from present-day moral standards or the "natural morality" put forward by literary Darwinists) vs. narrative "action"; (2) the level of saturation with concepts inextricably linked to the sociocultural context of the novels (i.e. references to places, people, professions that are no longer relevant or no longer exist); (3) the inherent stylistic properties of the text that have "gone out of fashion" or undergone changes over the years.

The presentation consists of three parts. The first one will compare a late Victorian bestseller and a canonical novel by using computational close-reading (what Matthew Jockers terms "microanalysis"), in order to provide a "cross-section" of the overall issue. The two novels analyzed in this part are Hall Caine's The Manxman (1894), which despite being a massive bestseller at the time (selling 400,000 copies within a couple of years) is now out of print, and its canonical counterpart – Thomas Hardy's Tess of the d'Urbervilles (1892) (that only sold about 20,000 copies in its early years). Both novels belong to the same genre, and since they were published within a couple of years, they share a nearly identical cultural (external) environment, which "evens the field" for content analysis. The second part continues the synchronic analysis and applies the same methods on ten other such pairs of late Victorian bestselling and canonical novels (paired according to the same criteria). In the third part, the analysis will expand the scope to include all bestsellers from 1891 to 1901 recorded in "The Bookman" – a late Victorian journal that features the most complete listing of bestselling books from the period.

The computational analysis largely relies on LDA topic modeling, utilizing the "tidytext" and "topicmodels" packages in R, as well as simple token distribution analysis. The preliminary findings indicate that late Victorian bestsellers feature markedly high levels of linguistic complexity, heavy saturation with contemporary cultural context, and contain large swathes of moralizing and authorial commentary in their text, which may be perceived as digressions from the main flow of the narrative, especially by modern readers, who are proven to have considerably shorter attention spans than their Victorian counterparts.