Session 1

Tanja Säily: A large-scale analysis of the use of Scotticisms in the eighteenth century

During the eighteenth century, the standardization of the English language in Scotland became a central topic in public discussions ranging from politeness to cultural and political identity. Standardization was promoted by groups such as the Select Society of Edinburgh. One of the members of the society was the Scottish Enlightenment philosopher David Hume, who published a list of "Scotticisms" in his 1752 work Political Discourses, advocating against their use and providing "English" alternatives. Over the rest of the century, Hume's list was reprinted several times, suggesting that it was seen as important and that it influenced at least some eighteenth-century writers.

This study combines full texts from the massive database of Eighteenth Century Collections Online (ECCO) with bibliographic information in the English Short Title Catalogue to analyse the prevalence and usage of Hume's Scotticisms in eighteenth-century British publications. We calculate the frequency of Scotticisms and their "English" counterparts in first editions of books, examining factors such as genre, author, the author's background, and the place of publication. We track the occurrence of the items throughout the century to see whether the decrease in the use of Scottish items, which was already taking place at the time, accelerated after the publication of Hume's list, which could reflect the influence of the Select Society on bringing the issue of "correct" English to the national consciousness.

Timo Korkiakangas: How was Latin learnt in the Middle Ages? A diachronic corpus-based study of medieval Latin literacies

Latin was the language of writing and civilization in large parts of medieval Europe and effectively a second language that had to be learnt. But how, in fact, was Latin learnt in the Middle Ages? I take the language of medieval texts as the starting point for answering this fundamental question of historiography that is traditionally approached by studying schools, legislation, and manuscripts used for learning. In this way, I turn the attention from what was taught to what medieval Latin writers had actually learnt (i.e., the so-called intake, their Latin2 learner languages).

I investigate corpus-linguistically how medieval writers of history, hagiography, and notarial documents in Italy and Francia used a set of selected ancient Latin forms and constructions that were absent or different from their Romance vernacular. The analysis of these individual usages in the light of what is known of medieval grammar books, schools, didactics, and imitation of ancient literature is liable to reveal mechanisms through which those writers had acquired their Latin literacies. By examining the variation of such mechanisms over time, place, and writers' professional status, I seek to sketch a synthesis of the ways in which Latin was learnt in the period under examination. The scrutiny starts from the 7th/8th century, from where enough textual data survive, and ends in c. 1200, when Latin school grammars meant for second-language teaching came into wide use.

The proposed research refines the present picture of medieval education by showing in detail how the Latin literacies varied in a given century between high clergy, local notaries, and, for example, the barely literate members of a noble family. In my presentation, I will discuss the on-going process of corpus building, including its conversion into a parsebank. A special attention will be paid to the evaluation of the parsed data: in addition to the normal performance scores calculated against a Gold Standard (PoS, morphological features, UAS, LAS), I will report scores for successfully parsed morphosyntactic constructions (e.g., accusative with infinitive, absolute infinitive construction, various agreement phenomena) that are crucial for any corpus-linguistic analysis of Latin.

Panel

Mikko Tolonen: The Relevance of High Performance Computing for Intellectual History

This presentation delves into the transformative impact of High-Performance Computing (HPC), machine learning, and data-driven methodologies on the future of intellectual history. Amid the discourse surrounding revolutionary technologies like BERT and expansive language models, the talk directs its attention toward the pragmatic implications these innovations hold for historical research. It explores the interplay between structured and unstructured data, specifically pertinent in comprehending intricate phenomena like the Scottish Enlightenment. Drawing from a decade of interdisciplinary collaboration within the Helsinki Computational History Group (COMHIS), this presentation outlines the group's research strategy and recent forays into Enlightenment studies. It features a consortium funded by the Academy of Finland, HPC-HD, which synergizes COMHIS with computer/data science groups, including TurkuNLP. This collective effort effectively harnesses HPC to unearth evolving discourses within historical texts. Our vision leverages the potency of deep learning to scrutinize intentions, receptions, and meanings enmeshed within historical texts. Innovative methodologies for identifying shared meanings, transcending language barriers and digitization intricacies, constitute the cornerstone of this computational approach. By pivoting from traditional keyword-centric analysis, this model places context at its core. This approach not only enriches our understanding of intricate historical phenomena but also advances the exploration of material objects in eighteenth-century intellectual history.

Heikki Rantala: Searching and visualizing the plenary speeches of the Parliament of Finland

ParliamentSampo is a data service (https://www.ldf.fi/dataset/semparl) and a web application (https://parlamenttisampo.fi) for presenting open data about the plenary speeches and the parliamentarians of the Parliament of Finland. ParliamentSampo data service makes available the text and other data of all the plenary speeches of the Finnish Parliament from 1907 to 2022 in a harmonized and machine readable format. For the newest corpus from 2015 onwards the speeches have also been annotated using Natural Language Processing techniques: Mentions of MPs and places have been extracted, disambiguated semantically, and linked to corresponding resources; keywords have been added to the speeches; and the texts have been lemmatized. The data is published in Linked Open Data, CSV and XML formats. The data can also be accessed using an open SPARQL endpoint. ParliamentSampo web application is built using this endpoint. It combines faceted search and visualizations, and it can be used by researchers, journalists or interested citizens to easily search and visualize plenary speeches. In my presentation I will tell how ParliamentSampo was created and how it can be used.

Jenni Santaharju: Linguistic diversity revealing our population history

When people move and have contact with each other, it may leave an imprint on different kinds of human diversity such as linguistic, genetic or cultural diversity. Language, genes and culture may spread synchronously or tell different stories about our past. Therefore, looking at different kinds of human diversity will provide us a more thorough view about human population history than just looking at one kind of data.

Temporal and spatial variation of linguistic diversity helps us to understand what has driven or hindered linguistic divergence and convergence. What explains differentiation of human populations and their contacts between them? An excellent area to study these questions is by looking at linguistic diversity is Finland, as due to the ice age its population history is relatively recent if compared to many other areas. In addition Finnish has been well studied both with traditional and quantitative methods. We have used the digitized Dialect atlas of Finnish collected by Lauri Kettunen in the 1920s and 1930s that enables an excellent municipality-wise overview on the diversity in Finnish 100 years ago. I will give a short introduction on our research topics and approaches, and tell what we have learned about the population history from linguistic diversity in Finnish.

Session 2

Mikko Laitinen: Novel computational methods and social networks

Social networks have been observed to play a considerable role in language variation and change, and social network theory has offered a powerful tool in modeling how innovations spread into communities (Milroy 1987). However, existing work on networks, with foundations in linguistic anthropology, leaves open various questions. One of them is that manual participant observation methods effectively limit the analysis to networks with sizes “between 30 and 50 individuals” (Milroy & Milroy 1992: 5). The theory, therefore, relies on evidence from networks that are substantially smaller than natural human networks (e.g. Dunbar 2020). Second, much of the evidence comes from traditionally close-knit urban working-class settings or peripheral rural communities, and it has been suggested that the model cannot be easily operationalized in situations where the population is socially and/or geographically mobile, as is the case in many contemporary societies and communities that are characterized by diversity.

This presentation discusses a computational method that utilizes interaction data obtained from digital social networks. This interaction-based information can be quantified using network parameters (e.g. density, connectedness and similarity) that are adopted from the graph theory but have not so far been used in sociolinguistics or in corpus-based study of variation and change.

This presentation first illustrates the method developed in our group and then in the empirical part establishes network parameters in a dataset of circa 4.8 billion words of English. This dataset has been obtained from a set of 3,935 randomly selected ego networks that contain user-generated texts and interactions from 233,774 individuals in the UK and the US. The median network size in this dataset is 51 individuals. The empirical case study utilizes all the texts from all the individuals in the networks and investigates how two linguistic features currently undergoing change in English are conditioned by network information. The empirical study concentrates on one orthographic feature (contractions of negatives (e.g. not >n't) and verbs (e.g. we will > we'll)) and one grammatical structure (need to + V), both of which are undergoing frequency increases in English, but are driven by differing forces of colloquialization and grammaticalization (Leech et al. 2009; Daugs 2017). Our observations using evidence from data-intensive methods may lead to rethinking the role of social networks in language change.

References:

Daugs, R. 2017. On the development of modals and semi-modals in American English in the 19th and 20th centuries. In Hiltunen, McVeigh & Säily (eds.), Big and Rich Data in English Corpus Linguistics: Methods and Explorations. (Studies in Variation, Contacts and Change in English, vol. 19).

Dunbar, R. 2020. Structure and function in human and primate social networks: Implications for diffusion, network stability and health. Proc. R. Soc. A. 476A.

Leech, G, M. Hundt, C. Mair & N. Smith. 2009. Change in Contemporary English. A Grammatical Study. Cambridge: CUP.

Milroy, J. 1992. Linguistic Variation and Change: On the Historical Sociolinguistics. Blackwell.

Milroy, L. 1987. Language and Social Networks. 2nd ed. Blackwell.

Milroy, L & Milroy J. 1992. Social network and social class: Toward an integrated sociolinguistic model. Lang. in Society 21, 1–26.