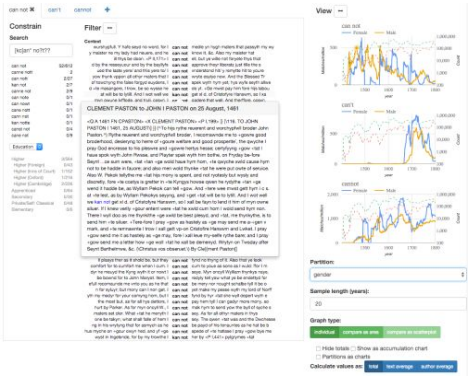
STRATAS: Combining texts and contextual information in historical sociolinguistics

Presenter: Tanja Säily
Team: Terttu Nevalainen, Anni Sairio, Tanja Säily, Samuli Kaislaniemi, Anna Merikallio (Department of Modern Languages, University of Helsinki); Taru Nordlund, Katja Litola, Johanna Marttila (Department of Finnish, Finno-Ugrian and Scandinavian Studies, University of Helsinki); Eetu Mäkelä (Department of Digital Humanities, University of Helsinki); Poika Isokoski, Harri Siirtola (Faculty of Communication Sciences, University of Tampere)

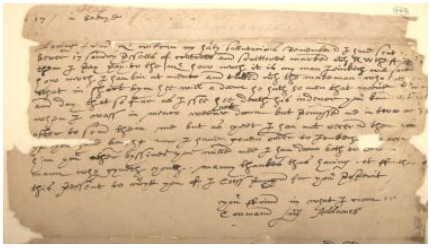
Funded by the **Academy of Finland DIGIHUM programme** for 2016–19, the STRATAS project studies language change by developing tools that enable us to ask questions that have until now been too labour-intensive to answer. These tools are being developed by computer scientists and visualization specialists in collaboration with language historians who study the development of English and Finnish over time.

The project focuses on sociolinguistic variation and change in personal and private writings. Manuscript materials give wider access to the varying communicative needs of individuals in the past than the printed word databases that are more readily accessible to scholars of linguistic, cultural and social history. Private writings are also embedded in a rich sociolinguistic context. From a computer-science perspective this poses a challenge: current tools do not provide an easy way to combine texts with metadata on e.g. the writers' social status.

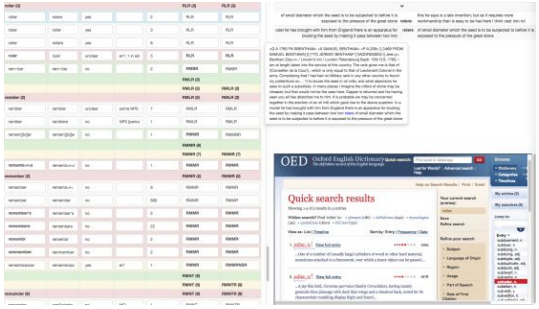


STRATAS creates modular open source tools that enable researchers to interactively explore the social embedding of language use by combining texts, metadata and visualizations. The toolkit is being created in conjunction with a larger set of tools for digital humanities currently being developed in international collaboration. This way, insights from STRATAS can inform wider developments, while the project gains ready-made modules and functionalities for data integration, visualization and exploration. The tools under development include Khepri (Mäkelä et al. 2016a), FiCa (Filtering and Categorization; Mäkelä et al. 2016b), and Text Variation Explorer 2 (Siirtola et al. 2016).

The tools enable **new kinds of research**: this project focuses on social meanings of spelling and word-form variation in historical English and Finnish, and the social embedding of neologisms in earlier English. By analysing original manuscript data, where variation has not been edited away, and by connecting it with external factors (place, time, social status, ideological environment), we can observe how social meanings arise in context. We will also compare manuscripts with printed data to study standardization processes.



To approach these research questions, we need to **survey the reliability of available English sources (subproject ERRATAS)** and **compile a gold-standard manuscript-based corpus of Finnish 19th-century letters (subproject RATAS)**. Most manuscript texts are digitized from modern printed editions, which commonly normalize spellings. To establish the original, authentic spellings, we need to go back to the manuscripts – but we can also chart which features of spelling are usually not modernized, and thus make all editions more reliable for linguistic research.



This presentation will discuss STRATAS as a whole as well as introducing a **case study of the social embedding of neologisms in 18th-century English letters**, along with our plans for expanding the study through new tools and data from large historical corpora and databases, which include the *Oxford English Dictionary* and its *Historical Thesaurus*. The ERRATAS and RATAS projects will also be presented as posters at the Summit.

References

Mäkelä, Eetu, Tanja Säily & Terttu Nevalainen. 2016a. Khepri – a modular view-based tool for exploring (historical sociolinguistic) data. In Maciej Eder & Jan Rybicki (eds.), *Digital Humanities 2016: conference abstracts*, 269–272. Kraków: Jagiellonian University & Pedagogical University. http://dh2016.adho.org/abstracts/226

Mäkelä, Eetu, Tanja Säily & Terttu Nevalainen. 2016b. Developing an interface for historical sociolinguistics. Paper presented at the *Digital Humanities Congress* (DHC 2016), Sheffield, September 2016. https://www.hrionline.ac.uk/dhc/2016/paper/99

Siirtola, Harri, Poika Isokoski, Tanja Säily & Terttu Nevalainen. 2016. Interactive text visualization with Text Variation Explorer. In Ebad Banissi (ed.), *Proceedings of the 20th international conference on Information Visualisation (IV 2016)*, 330–335. Los Alamitos, CA: IEEE Computer Society. doi:10.1109/IV.2016.57