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 e.g. the writers’ social status.

STRATAS will create modular open source tools that enable researchers to interactively explore the social embedding of language use by combining texts, metadata and visualizations. The toolkit will be 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 will 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 will need to survey the reliability of available English sources and compile a gold-standard manuscript-based corpus of Finnish 19th-century letters. 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.