Contributor/Presenter:
Timo Honkela PhD (Professor of Research in Digital Resources),
University of Helsinki, Department of Modern Languages, Helsinki
National Library, Centre for Preservation and Digitisation, Mikkeli (Finland)

Title:
Improving the quality of a bilingual historical news collection using statistical machine learning methods

Abstract
The National Library of Finland has digitized a large proportion of the historical newspapers published in Finland between 1771 and 1910. This collection contains approximately 1.7 million pages in Finnish and Swedish. In the output of the Optical Character Recognition (OCR) process, errors are common especially when the texts are printed in the Fraktur typeface. The errors lower the usability of the corpus. Automatic spell checking and correction is problematic, for instance, due to the historical spelling variants. This variation leads to the situation that modern morphological analysis tools are not directly applicable. Developing new morphological models for various historical periods would be very costly for a morphologically highly complex language such as Finnish. One approach for improving the quality of the texts after the OCR process is to apply crowdsourcing. In this paper, another path is taken. We experiment with different kinds of statistical machine learning methods. In particular, we apply data driven and mostly unsupervised methods in order to minimize the need for human intervention. Due to the nature of the task, some human guidance for the machine learning process is needed but the basic objective is to reach radical reduction in the human effort needed in the post-processing of the texts. The methods that are used include 1) n-gram models at the level of individual letters as well as for short word-level contexts and 2) the Morfessor method that automatically learns to segment words in a meaningful way. The two main variants of the Morfessor method apply the Minimum Description Length principle and Maximum Likelihood optimization. We present quantitative evaluations of the quality improvements gained with the techniques and provide an estimate of the human efforts required in the process. We also discuss the implications of the work regarning the content analysis of the newspaper collection.

Biography:
Since the beginning of 2014, Timo Honkela has served as a professor for research in digital resources. This field of study includes research into the use of linguistic and other resources in the humanities and social science, as well as development of new kinds of methods and approaches for representing and analyzing humanistic research data. In his position, Honkela works both at the Department of Modern Languages, Helsinki, and at the Centre for Preservation and Digitisation, Mikkeli.
Earlier Honkela served as a chief research scientist at the Department of Information and Computer Science, Aalto University School of Science, where he was the head of the Computational Cognitive Systems research group. Honkela has published more than 160 scientific papers in which methods of natural language processing, text mining, statistical machine learning and cognitive modeling have been applied in different scientific disciplines including linguistics, cognitive science, psychology, educational research, sociology, organizational research, health sciences, religious studies, history, media art, and philosophy