Aligning Manual transcriptions to images of Swedish historical tax lists

Handwritten text found in historical records such as tables cannot be automatically transcribed by traditional Optical Character Recognition (OCR) systems. Therefore, Handwritten Text Recognition (HTR) is being used, which usually require large amounts of annotated data for learning. Such annotations can be obtained in a process called alignment, which makes use of manual transcriptions of some part of the document, and are linking corresponding text images to transcribed words. This project aims at implementing text alignment for mainly handwritten historical tables, using different machine learning techniques. The end product will not only be a working alignment algorithm but also a comparison of several such algorithms on historical records.

The project is part of an effort to digitize historical tax lists for Sweden over the 1870 to 1950 period, which is used within the project “The Swedish Transition to Equality: Income Inequality with New Micro Data, 1870–1970”, led by economic historians Jakob Molinder (UU), Erik Bengtsson (LU), and Svante Prado (GU). The research project aims to create new estimates for income inequality in Sweden for the period prior to the existence of digital microdata. The tax lists constitute a unique source on the incomes of all individuals in Sweden required to pay taxes each year. The material is vast, and the research project has so far digitized about 100 000 tax returns using manual extraction, which only constitutes a small sample of the total available. You will have access to the photographs of the tax lists taken by the project, as well as the already hand-digitized records.

The long-run ambition is to digitize the full set of records, which amount to several million individual entries per year; Something that can only be done by automating the procedure. The full set of records will allow for a completely different set of analyses, and will ultimately be linked to individuals in the historical Swedish censuses. The end result will be a dataset of all individuals in Sweden, with information on, among other things, name, occupation, age, and income linked from the tax records. This will provide a unique resource with wide application for research in economic and social history, as well as for genealogists.