Automating Historical Source Transcription

By Gunnar Thorvaldsen

To cite this article: Thorvaldsen, G. (2021). Automating Historical Source Transcription. *Historical Life Course Studies, 10*, 59-63. https://doi.org/10.51964/hlcs9568

HISTORICAL LIFE COURSE STUDIES

Not Like Everybody Else.
Essays in Honor of Kees Mandemakers

VOLUME 10, SPECIAL ISSUE 3
2021

GUEST EDITORS
Hilde Bras
Jan Kok
Richard L. Zijdeman

EHPS
NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the openjournals website.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, http://www.esf.org), the Scientific Research Network of Historical Demography (FWO Flanders, http://www.historicaldemography.be) and the International Institute of Social History Amsterdam (IISH, http://socialhistory.org/). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at https://openjournals.nl/index.php/hlcs.

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.

The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: http://www.ehps-net.eu.

HISTORICAL LIFE COURSE STUDIES
VOLUME 10, SPECIAL ISSUE 3 (2021), 59-63, published 31-03-2021

Automating Historical Source Transcription

Gunnar Thorvaldsen

UiT Arctic University of Norway & Ural Federal University

ABSTRACT

Transcribing the 1950 Norwegian census with 3.3 million person records and linking it to the Central Population Register (CPR) provides longitudinal information about significant population groups during the understudied period of the mid-20th century. Since this source is closed to the public, we receive no help from genealogists and rather use machine learning techniques to semi-automate the transcription. First the scanned manuscripts are split into individual cells and multiple names are divided. After the birthdates were transcribed manually in India, a lookup routine searches for families with matching sets of birthdates in the 1960 census and the CPR. After manual checks with GUI routines, the names are copied to the text version of the 1950 census, also storing the links to the CPR. Other fields like occupations or gender contain numeric or letter codes and are transcribed wholesale with routines interpreting the layout of the graphical images. Work employing these methods has also started on the 1930 census, which is the last of the Norwegian censuses to be transcribed.

Keywords: Census, Transcription, Machine learning, Population register

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs5968
The article can be downloaded from here.

1 INTRODUCTION

For over four decades, The Norwegian Historical Data Centre has used optical techniques for transcription of printed source materials cooperating with Dutch colleagues. It has also been a pleasure to work with Kees Mandemakers during his leadership of the European Historical Population Samples Network (EHPS-Net) in my role as head of the Subgroup for New Databases. In Norway, we had developed special software to read our 1886 farm tax lists, as described in Kliatskine et al. (1997). Commercial software now reads Latin fonts in printed historical sources, but handles printed Gothic fonts and tight columns only with difficulty. While optical character recognition (OCR) for print is a mature technique, optical transcription of handwriting is precarious and presupposes special programs.

Handwritten names, occupations, etc. cannot be handled easily with OCR since it is difficult to distinguish individual characters. Instead, we cluster whole words mathematically according to image similarity. E.g. many instances of the same birthplace can be transcribed wholesale, if the handwriting style is similar. Since the image and line numbers identify each cell on the questionnaire, we can rationalize the transcription of protocol data cell by cell (Thorvaldsen et al., 2015). This reference describes systems for transcribing censuses and marriage protocols from Barcelona between 1451 and 1905. The Swiss Transkribus project works with text and tries to identify each writer's handwriting by relating some 100 scanned pages to its transcriptions (https://transkribus.eu/Transkribus). Artificial intelligence techniques including (deep) machine learning promise great potential for the future development of optical handwriting recognition, provided we can train appropriate computer models. Thus we now transcribe the first full count historical census in the world before it is opened up for genealogists.

2 THE 1950 POPULATION CENSUS

The Norwegian full count nominative 1801, 1865, 1875, 1891, 1900 and 1910 censuses have been transcribed and the 1920 census follows soon, as the one hundred year blocking expired. Concurrently, we experience much interest in more recent source materials, particularly for medical research. Foremost among these are the 1950 and the 1930 housing and population censuses. The National Archives has scanned these as part of our effort to build a Historical Population Register for Norway. The 1960 and later censuses were transcribed as part of the contemporary aggregation, and went into the building of the Central Population Register from 1964. Adding and linking a transcribed version of the 1950 census with 3.3 million person records provide longitudinal information about the childhood of significant population groups, which are still alive. However, here we receive no help from commercial companies or volunteer genealogists. Finding resources to transcribe the information-rich 1950 census without support is hard, and the same is the case with the similarly structured 1930 census for another decade. The birth records, which are closed from 1935 onwards and the 1946 census are also topical sources not open for genealogists, while access is more open to the local population registers which are now scanned in the National Archives.

The 1950 census manuscript collection consists of 801,000 double-sided questionnaires, which have been scanned to over 1.6 million images in JPEG and JPEG 2000 formats — compressed and full resolution respectively. The forms are large: 29.7 x 70.7 cm. Each form's inside contains up to ten persons with information on education, religious affiliation and previous migration in addition to the usual census variables. Between each nominative line, there are lines with codes for most variables. The reverse side contains address information.

All images have been analysed with the software *Analyseform* developed by the Norwegian Computing Center. It cuts out most fields into separate images, storing them in a database with references to questionnaire, line and row numbers. The program attempts to separate last name, middle names and first names into distinct images, which may fail when there are misplaced spaces or names were written out of sequence.

Each text cell image is analysed mathematically with 30 variables, preparing to cluster similar cells into an interactive process. A standalone GUI (graphical user interface) developed by Kåre Bævre lists e.g. birthplaces similar enough on the computer screen together with a suggested transcription. The operator must click on any erroneously clustered images and OK or change the transcription of the rest. This is an efficient method to control millions of birthplaces compared to ordinary transcription when name variation is limited.

Unfortunately, the birthdate cells were really too small for handwritten day, month and year, rendering automatic interpretation error prone. These images were sent to *Suntech* in India for commercial manual transcription — legal since isolated cells cannot be connected with other census information. The company returned Excel files with cell identifiers and formatted six or eight digit birthdates. Spot tests indicate that the results are within the agreed 3% error rate, most inconsistencies caused by unclarities in the source.

Most personal names occur less frequently in the source than do the municipality names due to the combinations of first and last names. Thus, the GUI has to work with the names after the name elements have been separated by the *Analyseform* program; this splitting works satisfactorily after revisions. The high number of different names even after separation makes the interactive editing of the name clusters a more time consuming procedure than handling birthplace fields. Therefore, we had to devise an alternative procedure for the names, using record linkage.

3 RECORD LINKAGE

We have access to nominative information in the Central Population Register from 1964, the 1960 census and the causes of death register from 1951. Together, these sources contain names, birthdates etc. for virtually the whole population in 1950 — emigration was limited. We cannot identify single individuals uniquely with the transcribed birthdates, but manage many couples and related persons. Thus, we have identified families in the 1950 census where at least two persons had the same birthday as in a family found in the Central Population Register. Using these proto-links and excluding all persons with more than one first name, we guesstimated the likely transcription of the 1950 name images. In this way, even typed names are clustered with handwritten ones making the editing of each name cluster expedite. The operators are instructed to accept into the same cluster names that are written slightly differently such as 'Niels' and 'Nils' or 'Olav' and 'Olaf'; grapheme variants which are often mixed in actual practice. However, they are told to distinguish phonetically distinct names such as 'Maria' and 'Marie'. Also, they are told to not accept strings where elements from different names are displayed together.

There has been a breakthrough in image recognition for many applications using deep learning algorithms reported in papers like Krizhevsky et al. (2012), making this a leading machine learning tool for computer vision. 353,000 images were used to train a deep learning neural network classifying first names. This network was then applied to classify the remaining 3.4 million first names. The network gave the ten most likely names with a score. These are again manually verified using the GUI. This job is rationalized by identifying single persons with a unique birthday and first name in a municipality. With some additional criteria, this led to additional 450,000 images with an improved 'guess' at the first name. Via this and similar methods we are able to reduce considerably the manual job of verifying 3.4 million first names. However, the variation among last names is larger than among first names, increasing the size of this job. In general, larger training sets lead to better classification of the neural network, reducing the need for manual verification.

Deep learning is also used to handle other cells like sex, family position and birthplace in the large cities. We first need to create a training set, then train the deep learning network, followed by applying the trained network for classifying the remaining cells and finally manually verify the classification. Numeric codes are first split into separate digits and then deep learning is used for classifying each digit, since there exist efficient techniques for interpreting single handwritten digits. This method works better when we know the number of digits in the cell since a large ratio of the errors is due to the splitting of numbers into digits. For other cells like occupation, it is more difficult to employ deep learning since there are too few test data for each alternative.

The 1950 census will form part of the Historical Population Register for Norway. Therefore, we want to keep and extend the links to the 1964 central population register, the 1960 census and the cause of death register 1951–1964. Thus, our approach for transcribing the names, as a by-product, provides links needed in the Historical Population Register. We will try to link as many person records as possible based on birthday, name and where available also municipality of birth. While we are waiting for the 1920 and 1930 censuses, the most realistic linkage backwards in time is towards the 1910 census, which was the first to include birthdates for the whole population; in 1891 and 1900 only children under two had their birthdates noted.

4 OTHER VARIABLES FOR RECORD LINKAGE AND ANALYSIS

As a further step in record linkage, the last names and the middle names can be treated by clustering them after looking up the names in newer sources analogous to the first name method. For the surnames, there will be problems with the ca 255,000 women who married between the 1950 and 1960 censuses, taking the husband's surname. Strict name legislation made other name changes rare at the time, but a number of persons switched between multiple first names. Regrettably, the field for previous names are seldom filled in the Central Population Register, and about half of the population lack exact birthplace information.

In addition to the variables above, we have read gender, represented by a single character, automatically from the census forms and checked it against the first names. Naturally, the 1950 census will be more useful in longitudinal and cross-sectional studies if we add the other variables such as occupations etc. Using the clustering method with these alphabetic fields will require significant resources to inspect, edit and encode the information. We prefer to utilize the numeric codes, which Statistics Norway entered for each person on the 1950 census form. These were later on punched onto Hollerith cards, and used to aggregate the census information, but the cards were discarded before Statistics Norway owned tape drives or other media for compact information storage.

In order to digitize the codes we are in the process of interpreting them with established machine learning techniques and libraries (Abadi et al., 2016). IT specialists consider this realistic since a limited number of personnel in-house in Statistics Norway, wrote the codes. The codes are distinguishable because of the red pencils used and we have the codebook so that the number of alternatives are limited in all fields except occupations. Libraries such as Tensorflow provide access to both state-of-the-art image processing methods and efficient implementations for execution on powerful GPUs (Graphical Processing Units). We are presently creating a complementary training set based on a representative sample of digits from the census forms in order to enhance precision. Finding methods to control the resulting database with algorithms that combine these variables in inventive ways and to find rational methods for interactive proofreading, are still challenges.

ACKNOWLEDGEMENTS

This paper is written with input from Kåre Bævre (National Institute of Health), Lars Holden (Norwegian Computing Center), Trygve Andersen (UiT) and Lars Ailo Ballo (UiT). Supported financially by the Norwegian Research Council (Project # 225950).

REFERENCES

The publications including texts in English based on the 1950 census are available at https://www.ssb.no/historisk-statistikk/folketellinger — click on 1950 — the codebook is only in Norwegian.

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B. Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *OSDI'16: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation* (pp. 265–283). Berkeley, CA: USENIX Association.

Kliatskine, V., Shchepin, E., Thorvaldsen, G., Zingerman, K., & Lazarev, V. (1997). A structured method for the recognition of complex historical tables. *History and Computing*, *9*(1–3), 58–77. doi: 10.3366/hac.1997.9.1-3.58

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*(2). doi: 10.1145/3065386

Thorvaldsen, G., Pujadas-Mora, J., Andersen, T., Eikvil, L., Lladós, J., Fornés, A., & Cabré, A. (2015). A tale of two transcriptions. Machine-assisted transcription of historical sources. *Historical Life Course Studies*, *2*, 1–19. Retrieved from http://hdl.handle.net/10622/23526343-2015-0001?locatt=view:master