Proceedings of LR4SSHOC: Workshop about Language Resources for the SSH Cloud, pages 27–32
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Mining Wages in Nineteenth-Century Job Advertisements
The Application of Language Resources and Language Technology to study Economic and Social Inequality

Ruben Ros,[1] Marieke van Erp,[2] Auke Rijpma,[1,3] Richard Zijdeman[3,4]
[1]Utrecht University
[2]KNAW Humanities Cluster DHLab
[3]International Institute for Social History
The Netherlands
[4] University of Stirling
Scotland
r.s.ros@uu.nl, marieke.van.erp@dh.huc.knaw.nl,
a.rijpma@uu.nl, richard.zijdeman@iisg.nl

Abstract

For the analysis of historical wage development, no structured data is available. Job advertisements, as found in newspapers can provide insights into what different types of jobs paid, but require language technology to structure in a format conducive to quantitative analysis. In this paper, we report on our experiments to mine wages from 19th century newspaper advertisements and detail the challenges that need to be overcome to perform a socio-economic analysis of textual data sources.

Keywords: historical newspaper, job ads, occupations, income analysis, information extraction

1. Introduction

Economists and sociologists draw on historical data to study long term trends. Information from archival records is used, for example to assess intergenerational mobility (Knigge, 2016). Overall, relative to contemporary data, historical quantitative data are hard to come by, and seldom the result of a research oriented data gathering process. As a result, researchers have to deal with what is available: a signature to indicate someone's literacy, relative height to proxy health and the rounding of numbers ('age heaping') as an indicator of a population's numeracy. In that sense even occupation is a proxy for a person's human, economic and social capital.

Historians draw, next to quantitative sources, on qualitative sources. For example, by manually studying 617 memoirs for reasons why people put their children too work, or rather try to keep from work (Humphries, 2003). Another example is (Schulz et al., 2014) who painstakingly derived ascribed and achieved characteristics from 2194 job employment advertisements. However, compared to the derivation of occupational information, mining such qualitative data is even more labour intensive and will therefore always lead to relatively small samples, reducing the statistical power of hypothesis tests.

In this paper, we expand on these sources by using a computer assisted method, text mining, to extract from a qualitative source an occupation specific characteristic: wage. Wages are one of the most important long-run data series being gathered. They are a headline measure of long-run living standards and economic leadership (Allen, 2001; Allen et al., 2011; Feinstein, 1998; de Zwart et al., 2014). The Malthusian view of pre-industrial societies being characterised by long-term income stagnation relies on wages and prices for its empirical support (Clark, 2005). Furthermore, wage data are a key component in current explanations for the timing and location of industrialisation and the transition to sustained growth (Allen, 2009).

However, there has been substantial criticism at the methodology used in such research. Daily wages from a narrow set of occupations (construction workers) are made comparable through a standardised CPI and working days estimate. The degree to which such estimates are representative over all occupations and the places for which we have wages, as well as the assumptions about hours worked are increasingly criticised as biasing results (Stephenson, 2018; Humphries and Schneider, 2019; Humphries and Weisdorf, 2019). Overall, the call is to broaden the horizon in the type of sources being used to obtain historical wage data.

In this paper, we present a text mining use case for the social science and humanities domain. We describe a rule-based classifier that is used on a large corpus of job advertisements to analyse the development of wages. With this approach, we can automatically extract wages from a wide range of occupations and places. We describe our experiments and results, as well as the challenges in working on non-digital born resources and with a corpus that displays diachronic language variation.

The remainder of this paper is organized as follows. In the next section, we describe related work followed by the data sources used in Section 3. We then describe our approach in processing and analysing the wage information from digitized newspapers in Section 4. Our evaluation is presented in Section 5. followed by our conclusions and future work in Section 7. The code used in this project, as well as the lists of occupation titles and qualitative wage indicators can be found on: `https://github.com/rubenros1795/mining-job-ads`

2. Related work

To the best of our knowledge, advertisements have only occasionally been used in historical sociology and economic history. Schulz et al. use job advertisements to study whether people are hired for their ascribed or achieved characteristics (Schulz et al., 2014). Gray also collects quantitative data from newspaper advertisements, in this case rental prices to study the New York rental market (Gray, 2018). However, so far all these studies rely on manual entry to extract information from advertisements. One exception is Cummins, who uses regular expressions to extract information on wealth at death from the printed volumes of the Principal Probate Registry of the United Kingdom between 1892–2016 (Cummins, 2019). Automating this process further, and evaluating the results systematically, can greatly increase the usability of this kind of source material.

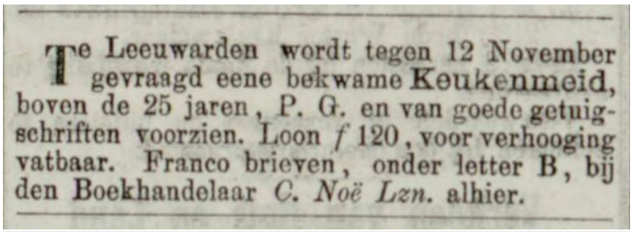
Automatically extracting information from text is a well-developed task in natural language processing research (Weischedel and Boschee, 2018). However, occupations are not a category of concepts that are generally included in such tasks, which often focus on named entities and relations between them. Many such approaches are based on machine learning methods, but annotating training data for this work was out of the scope of our project. Furthermore, the non-digital born nature of the source material used in our study also affects the quality of automatic methods (van Strien et al., 2020).

3. Data

The Dutch National Library has undertaken several large-scale digitization projects, resulting in a collection of over 15 million pages of digitized newspapers spanning the period between 1618 and 1995.[1] Newspapers printed before 1876 are free of copyright restrictions. For the post-1876 newspapers, access was granted by the National Library. Besides the text of the newspaper articles (made machine-readable with Optical Character Recognition), various metadata fields such as dates, newspaper titles and links to the original images are also available.

The Dutch National Library has assigned four categories to different newspaper articles: news, advertisements, personal announcement, and illustrations. This makes it relatively straightforward to extract advertisements. For this project, we focused on advertisements in all newspapers present in the National Library database printed between 1850 and 1890. Although newspaper advertisements appear as early as the seventeenth century, it was only the mid-nineteenth century that saw the rise of large-scale newspapers, structured in a relatively consistent way and including relatively stable categories of advertisements.

After downloading the advertisements, they were preprocessed by removing all non-alphanumeric characters, followed by lowercasing and tokenization. We decided not to lemmatize the corpus because we suspected that the low OCR quality of the ads would hinder proper lemmatization. The statistics of the (preprocessed) data are summarized in Table 1. We measure OCR quality using the type-token ratio (TTR), which is the relative number of unique words divided by the relative number of words. Because faulty OCR produces many non-existent words, it is expected that texts with a low OCR-quality have a relatively high number of unique words (types). The ratio between the number of words and the number of unique words in a given year thus forms an approximation of OCR quality. Since a higher number of articles often leads to a higher number of types we divide both the number of tokens and the number of types by the number of articles.


Te Leeuwarden wordt tegen 12 November gevraagd eene bekwame Keukenmeid, boven de 25 jaren, P. G. en van goede getuigschriften voorzien. Loon *f* 120, voor verhooging vatbaar. Franco brieven, onder letter B, bij den Boekhandelaar *C. Noë Lzn.* alhier.

rpo Leeuwarden wordt tegon 12 November I gevraagd eene bekwamo Keukenmeid, boven de 25 jaren, I. O. en van goedo getuigschriften voorzien. Loon / 120, voor verhooging vatbaar. Franco brieven, onder letter B, bij den Bockhandolaar C. Nof Len. alhier.

In Leeuwarden is by 12 November asked one skilled kitchen maid, over 25 years old, P. G. and with good reference letters. Salary f120, increase possible. Post paid letters, under letter B. at the Bockshop C. Noë. Lzn. in this town.

Figure 1: Example of a non-preprocessed job advertisement in *Leeuwarder Courant* (19-7-1878) and its translation in *italics*

According to (Reynaert, 2008), a TTR of around 44% is expected for twentieth-century newspapers. Wevers reports type-token ratios between 5% and 25% in the period 1890-1910 (Wevers, 2017). The quality of the OCR in the ads used in this study is quite low, especially compared to similar measurements in earlier studies. The low scores for the material used in this study are likely to result from the visual complexity of advertisements (compared to news articles). This also explains the decline in type-token ratios in the period between 1850 and 1879. In this period, an increase in the number of advertisements and the growing size of newspaper pages led to more complex page compositions that are harder to process for OCR engines.

Figure 1 shows an example of a job advertisement published in 1878. The photographed newspaper is provided, as well as the translation in *italics*. OCR errors are marked in red.

4. Approach

The approach to extract information from job advertisements consists of two tasks. First, the job advertisements (currently not labelled as such by the National Library) need to be identified from the overall set of advertisements. The second step is to extract ”wage indicators” from the advertisements and connect them to specific occupations.

4.1. Extracting Job Advertisements

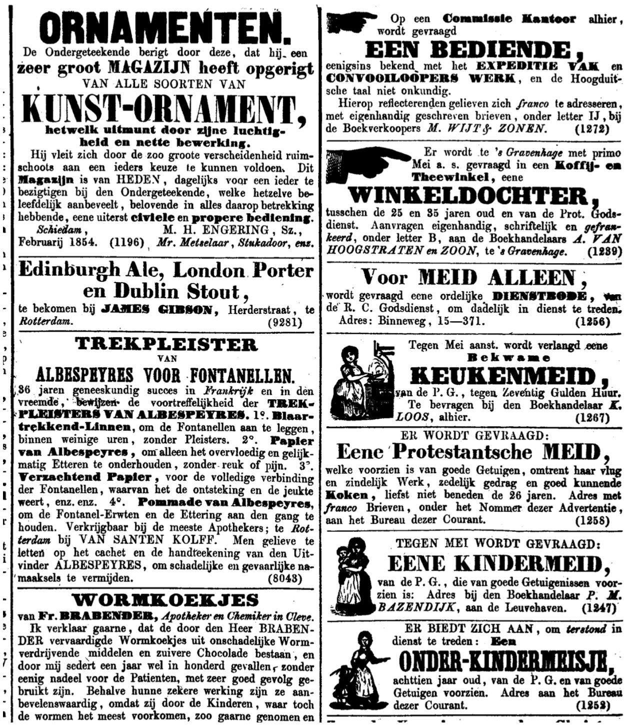
Because a single advertisement (identified as such by article segmentation metadata provided by the National Library) often contains multiple advertisements (Figure 2) we need to detect potentially multiple occupations and associated wages per article. Especially job advertisements are

[1] https://delpher.nl

	1850-1859	1860-1869	1870-1879	1880-1889
pages	1,491,391	3,321,334	3,534,270	3,427,923
articles	303,631	562,584	645,721	525,792
tokens	$1,1E+08$	$1,7E+08$	$1,72E+08$	$1,35E+08$
types	6,048,407	10,344,452	13,227,808	10,233,301
TTR	18%	16%	13%	13%

Table 1: Description of the dataset: the number of pages, the total number of articles, the total number of tokens, the total number of types and the type token ratio (TTR). The TTR is calculated as the ratio between the relative number of tokens (words) and the relative number of types (unique words).

sensitive to poor segmentation because they are generally short and do not have a consistent visual appearance. From all the advertisements that contain occupation titles, around 80% contain multiple titles. This problem could be solved by implementing a new segmentation procedure or by training a classifier to find relevant segments within the overall text, but this was out of the scope of this project.



ORNAMENTEN.

De Ondergeteekende berigt door deze, dat hij een **zeer groot MAGAZIJN heeft opgerigt** VAN ALLE SOORTEN VAN

KUNST-ORNAMENT,

hetwelk uitmunt door zijne luchtigheid en nette bewerking.

Hij vleit zich door de zoo groote verscheidenheid ruimschoots aan een ieders keuze te kunnen voldoen. Dit **Magazijn** is van HEDEN, dagelijks voor een ieder te bezigtigen bij den Ondergeteekende, welke hetzelve beleefdelijk aanbeveelt, belovende in alles daarop betrekking hebbende, eene uiterst **civiele** en **propere bediening.**

Schiedam, M. H. ENGERING, Sz.,
Februarij 1854. (1196) *Mr. Metselaar, Stukadoor, enz.*

Edinburgh Ale, London Porter en Dublin Stout,

te bekomen bij **JAMES GIBSON,** Herderstraat, te *Rotterdam.* (9281)

TREKPLEISTER

VAN

ALBESPEYRES VOOR FONTANELLEN.

36 jaren geneeskundig succes in *Frankrijk* en in den vreemde, bewijzen de voortreffelijkheid der **TREKPLEISTERS VAN ALBESPEYRES.** 1°. **Blaartrekkend-Linnen**, om de Fontanellen aan te leggen, binnen weinige uren, zonder Pleisters. 2°. **Papier van Albespeyres**, om alleen het overvloedig en gelijkmatig Etteren te onderhouden, zonder reuk of pijn. 3°. **Verzachtend Papier**, voor de volledige verbinding der Fontanellen, waarvan het de ontsteking en de jeukte weert, enz. enz. 4°. **Pommade van Albespeyres**, om de Fontanel-Erwten en de Ettering aan den gang te houden. Verkrijgbaar bij de meeste Apothekers; te *Rotterdam* bij VAN SANTEN KOLFF. Men gelieve te letten op het cachet en de handteekening van den Uitvinder ALBESPEYRES, om schadelijke en gevaarlijke namaaksels te vermijden. (8043)

WORMKOEKJES

van **Fr. BRABENDER,** *Apotheker en Chemiker in Cleve.*

Ik verklaar gaarne, dat de door den Heer BRABENDER vervaardigde Wormkoekjes uit onschadelijke Wormverdrijvende middelen en zuivere Chocolade bestaan, en door mij sedert een jaar wel in honderd gevallen, zonder eenig nadeel voor de Patienten, met zeer goed gevolg gebruikt zijn. Behalve hunne zekere werking zijn ze aanbevelenswaardig, omdat zij door de Kinderen, waar toch de wormen het meest voorkomen, zoo gaarne genomen en

Op een **Commissie Kantoor** alhier, wordt gevraagd

EEN BEDIENDE,

eenigsins bekend met het **EXPEDITIE VAK** en **CONVOOILOOPERS WERK**, en de Hoogduitsche taal niet onkundig.

Hierop reflecterenden gelieven zich *franco* te adresseren, met eigenhandig geschreven brieven, onder letter IJ, bij de Boekverkoopers *M. WIJT & ZONEN.* (1272)

Er wordt te *'s Gravenhage* met primo Mei a. s. gevraagd in een **Koffij- en Theewinkel**, eene

WINKELDOCHTER,

tusschen de 25 en 35 jaren oud en van de Prot. Godsdienst. Aanvragen eigenhandig, schriftelijk en *gefrankeerd*, onder letter B, aan de Boekhandelaars *A. VAN HOOGSTRATEN en ZOON*, te *'s Gravenhage.* (1239)

Voor MEID ALLEEN,

wordt gevraagd eene ordelijke **DIENSTBODE, van** de R. C. Godsdienst, om dadelijk in dienst te treden.
Adres: Binneweg, 15—371. (1256)

Tegen Mei aanst. wordt verlangd eene **Bekwame**

KEUKENMEID,

van de P. G., tegen Zeventig Gulden Huur.
Te bevragen bij den Boekhandelaar *K. LOOS*, alhier. (1267)

ER WORDT GEVRAAGD:

Eene Protestantsche MEID,

welke voorzien is van goede Getuigen, omtrent haar vlug en zindelijk Werk, zedelijk gedrag en goed kunnende **Koken,** liefst niet beneden de 26 jaren. Adres met *franco* Brieven, onder het Nommer dezer Advertentie, aan het Bureau dezer Courant. (1258)

TEGEN MEI WORDT GEVRAAGD:

EENE KINDERMEID,

van de P. G., die van goede Getuigenissen voorzien is: Adres bij den Boekhandelaar *P. M. BAZENDIJK*, aan de Leuvehaven. (1247)

ER BIEDT ZICH AAN, om *terstond* in dienst te treden: **Een**

ONDER-KINDERMEISJE,

achttien jaar oud, van de P. G. en van goede Getuigen voorzien. Adres aan het Bureau dezer Courant. (1252)

Figure 2: Two columns of advertisements in Nieuwe Rotterdamsche Courant (15-2-1854) that are identified as a single advertisement.

Job ads were extracted based on an expansive list of occupations obtained from the historical international classification of occupations (HISCO) database (Mandemakers et al., 2019; van Leeuwen et al., 2002).[2] We considered expanding the list by generating spelling alternatives based on string edit distance (e.g. "bakcer" and "dakker" as a way to detect wrongly OCR'ed instances of "bakker"). However, this introduced a new type of noise to the dataset because these alternatives contain many words that are no longer connected to the occupation title. Moreover, the list obtained from the HISCO dataset already includes 12,671 unique occupation titles.

After selecting the advertisements based on the list of occupation titles the problem of wrongly segmented advertisements was circumvented by selecting a specific window of words around the occupation title. This introduced the problem of multiple advertisements being grouped together. However, by measuring the average number of words between occupation titles, trying different window sizes and close readings of individual advertisements we found a window of twelve words to the left, and forty words to the right of the occupation title to be the most effective. Overall, this method would result in 175,209 extracted windows on a total of 1,515,179 advertisements.

4.2. Extracting Wage Information

Information about wages and compensations for advertised occupations in nineteenth-century job advertisements comes in two forms. First, many jobs are advertised with reference to qualitative indicators such as *hoog loon* ("high pay") and *behoorlijk salaris* ("reasonable salary"). This category is also marked by wage indicators that are related to the applicants' capabilities. Especially the phrase *loon naar bekwaamheden* ("wage by capabilities") frequently appears in the advertisements in the period 1850-1870. The second category of wage information is quantitative information. Despite the 'messiness' of digitized advertisements, numerical wage indicators are relatively consistent in form. Low-skilled jobs were often paid the same: a hundred guilders a year or two guilders a week. This reflects in the advertisements. Furthermore, OCR errors are also relatively consistent. Often, "10" is recognized as "lo" and the guilder sign "ƒ" is frequently recognized as an "f". The classifier is therefore designed in such a way that the most frequently occurring OCR-errors are recognized and corrected.

4.2.1. Extracting Qualitative Wages

Qualitative wage classifiers were extracted using a list of frequently occurring indicators (such as "good wage", "wage by capabilities", "good salary" etc.). This list was composed manually on the basis of an extensive close reading exercise. We observed how only a limited range of qualitative indicators were used in the advertisements and that the use of a manually composed vocabulary of indicator was justified. Of course, OCR might have affected the identification of the indicators, but because all of the indi-

[2] https://datasets.iisg.amsterdam/dataset.xhtml?persistentId=hdl:10622/MUZMAL.

cators concerned short words, the effect of OCR-noise was minimal.

4.2.2. Extracting Numerical Wages

The extraction of the numerical wages was the central and most complex part of the project. We tackled this task by designing a rule based classifier. The reason for doing so is twofold. First, as mentioned earlier, the quantitative information appears in a relatively consistent form. Second, such an approach does not require large amounts of labeled training data.

The first step in identifying numerical wages was to extract all tokens with numerical characters and with a length of less than six and appearing in the selected context windows around occupation titles. Because a significant portion of the wages comprised the numbers 10 or 100, we also included tokens containing the character combinations “lo” and “loo” as a way to capture the most frequent OCR errors. Additionally, a separate vocabulary of spelled out numbers (“hundred”, “fifty”) was used to extract quantitative information in non-numerical form.

After selecting a list of wage candidates a rule-based classifier was used to determine the likelihood of the selected token expressing a wage. The classifier uses the following features:

- whether the first character of the token, or the preceding token is an “f” or “ƒ”.
- whether the preceding token is the word *van* (“of”) or *tegen* (“against”), since wages are often discussed as *loon van 5 gulden* (“wage of five guilders”).
- whether the token that follows the numerical candidate is the word *gulden* (“guilder”).
- whether one of the three words before or after the numerical candidate is either *loon* (“wage”), *salaris* (“salary”), *beloning* (“remuneration”) or *jaarwedde* (“a year’s wage”).
- whether the first two characters of the numerical candidate are “18” or the token after the candidate is a month. In that case, the candidate is probably an indicator of time.

The features were selected on the basis of a close reading of a sample of over a hundred advertisements per decade. This showed that the appearance of job advertisements was relatively stable over time. Terms such as “salary” and “guilder” consistently appeared alongside occupation titles and the inner structure of advertisements was did not change significantly. This led us to select the above mentioned lexical features for the classifier.

5. Evaluation

The extraction and identification method was evaluated by comparing the extracted results with a set of manually annotated advertisements. A sample of 150 advertisements was drawn from the advertisements published in the years 1851, 1856, 1861, 1866, 1871, 1876 and 1881, resulting in a set of 1050 advertisements. Because at this stage we only wanted to evaluate the identification of wages and not the identification of occupations, we generated the windows using the method outlined in Section 4.. This resulted in a total of 620 job advertisements that formed the evaluation set.

The subsequent annotation procedure comprised the tagging of occupations, qualitative wage information, and quantitative wage information. In Table 2 below, all annotation options are listed. Initially, software (WebAnno) was used for annotation but given the fairly straightforward entities that needed annotation it proved much faster to use a hand-built Python script.

The following rules were set for annotating the windows:

- If two occupation titles are mentioned *reiziger of secondant* (“traveler or assistant”), both titles were annotated only if one of them is not detected in another advertisement. The script that creates the windows based on the occupations generates separate advertisement windows for every detected occupation, so in the case of “reiziger of secondant” both *reiziger* and *secondant* could have their “own” window. Therefore, if both are detected and processed as separate windows only one is annotated.
- Wages can come in the form of a range, such as: *ƒ 100 tot ƒ 120* (“f 100 tot f 120”). In this case, both indicators are annotated (separate by a space). However, when there is extra money to be earned *ƒ 100 inclusief f5 waschgeld* (”f1 including f5 laundry allowance”), only the ‘main’ wage number is anntated.
- qualitative and quantitative information sometimes cooccurs. In that case they are annotated as: quan [token(s)], qual [token(s)].

During the annotation it turned out that half of the advertisements could either not be classified as a job advertisement, or did not contain any wage indicators. The first problem arose from occupation titles such as *boekhandelaar* (“bookseller”), or *burgemeester* (“mayor”) that were mentioned in different contexts (See Figure 1). The second problem, was mostly caused by high-skilled occupations (teachers and civil servants) that were not accompanied by wages.

In Table 3 we report the results of the evaluation procedure. Because the results vary based on whether we count all ads, ads that are only classified as job ads, or ads that contain wage indicators, we differentiated the evaluation results. With f_1 scores between 0.624 and 0.724 our classifiers works reasonably well in extracting wage information. Given our original corpus of around 1.000.000 advertisements, containing around 175.000 occupations, a recall of 70% would give us 122.500 correctly classified advertisements. Here, we have to keep in mind the issue of true negatives: many of the windows are extracted on the basis of occupation titles that do not relate to job advertisements, but to other types of advertising. For example, the word ”baker” could also be used in the context of bread advertisements. The classifier correctly discards these windows, because they are not job advertisements. However, in the

Feature	Tag	Situation
Occupation	[occupation title]	if the extracted occupation is correct
Occupation	“na”	if the extracted occupation is incorrect because the ad is not a job ad
Occupation	[correct occupation]	if the extracted occupation is incorrect because another occupation is advertised
Occupation	“np”	if the extracted occupation is incorrect because no occupation is mentioned
WAGE	“qual” [token(s)]	if the extracted wage is indicated by qualitative indicators
WAGE	“quan” [token]	if the extracted wage is indicated by quantitative indicators
WAGE	“na”	if no wage indicators are present because the ad is not a job ad
WAGE	“np”	if no wage indicators are present

Table 2: Explanation of the tags used in the annotation.

	All Ads	Job Ads	Ads with Indicators
precision	0.715	0.717	0.641
recall	0.754	0.737	0.607
f_1	0.734	0.727	0.624

Table 3: Evaluation results for three different categories of advertisements in the evaluation set.

evaluation process, these “true negatives” are counted as correct classifications. When translating the recall of 70% to the expected number of extracted advertisements, these true negatives must be taken into account.

6. Preview: Wages in Domestic Services

To illustrate the types of analyses a structured dataset on wages can facilitate, we include a small example here of the combined wages of four similar occupations: kitchenmaid, maid, servant and housekeeper. Figure 3 shows the extracted quantitative wage indicators in the period between 1850 and 1879, combined for all four occupations. The resulting nominal wage series is fairly flat, but for the 1850–1875 period this is in line with other work (Allen, 2001).[3]. Moreover, the variation in wages in any given year confirms that there is substantial wage heterogeneity, even within a set of similar occupations.

7. Conclusions and Future Work

In this paper, we presented an approach and experiments for extracting wages and occupations from historical newspaper ads. We highlighted the main challenges in working with non-digital born data, as well as artefacts of the digitization process such as OCR and segmentation errors. Our observations can serve as guidelines for other researchers taking on text-driven analyses on historical newspapers.

Our method of extracting information about wages and occupations in historical advertisements proves to be a promising line of research. It sheds a light on the socio-economic history of professional categories that generally lack quantitative evidence. By using large-scale digital corpora and relatively straightforward text classification methods, this evidence can be gathered.

Of course, several problems need to be resolved before we can use wage indicators as reliable quantitative evidence. Three practical problems need further attention. First, occupations need disambiguation. In the example “baker searches apprentice”, both occupations, along with their windows, are extracted. In the current situation, both occupations would be connected to a wage indicator that only refers to “apprentice”. Tackling this problem could be done by, for example, considering the verbs that surround a specific occupation or by simply removing occupation titles that are seldom advertised. A second problem is the occurrence of multiple quantitative indicators in the advertisements. Wages were often negotiable or dependent on capabilities or background. In that case, we might encounter “f50-60”, “100, rising to 110” or “f100 with a bonus of f10 a month”. Currently, the classifier selects only the ”top candidate” from the ad. This problem could be resolved by a second classifier that disambiguates multiple numerical indicators extracted from the ads.

Thirdly, we could only perform an evaluation on a sample of the dataset. For large-scale data enrichments, further evaluations using different samples, or a human-in-the-loop annotation approach are recommended. Here, additional information about the system’s decisions such as its confidence or the text quality can help distinguish ‘easy’ from ‘difficult’ cases, enabling a setup where human experts are only presented those cases that the computer cannot solve. We foresee many benefits from hybrid annotation efforts and hope our experiment provides a first inspiration for such experiments in different contexts.

References

Allen, R. C., Bassino, J.-P., Ma, D., Moll-Murata, C., and Van Zanden, J. L. (2011). Wages, prices, and living standards in China, 1738-1925: in comparison with Europe, Japan, and India. *The Economic History Review*, 64:8–38, February.

Allen, R. C. (2001). The Great Divergence in European Wages and Prices from the Middle Ages to the First World War. *Explorations in Economic History*, 38(4):411–447, October.

Allen, R. C. (2009). *The British industrial revolution in global perspective*. New approaches to economic and social history. Cambridge University Press, Cambridge [etc.].

Clark, G. (2005). The Condition of the Working Class in England, 1209–2004. *Journal of Political Economy*, 113(6):1307–1340, December.

Cummins, N. (2019). Where Is the Middle Class? Inequality, Gender and the Shape of the Upper Tail from

[3] http://www.iisg.nl/hpw/data.php#europe

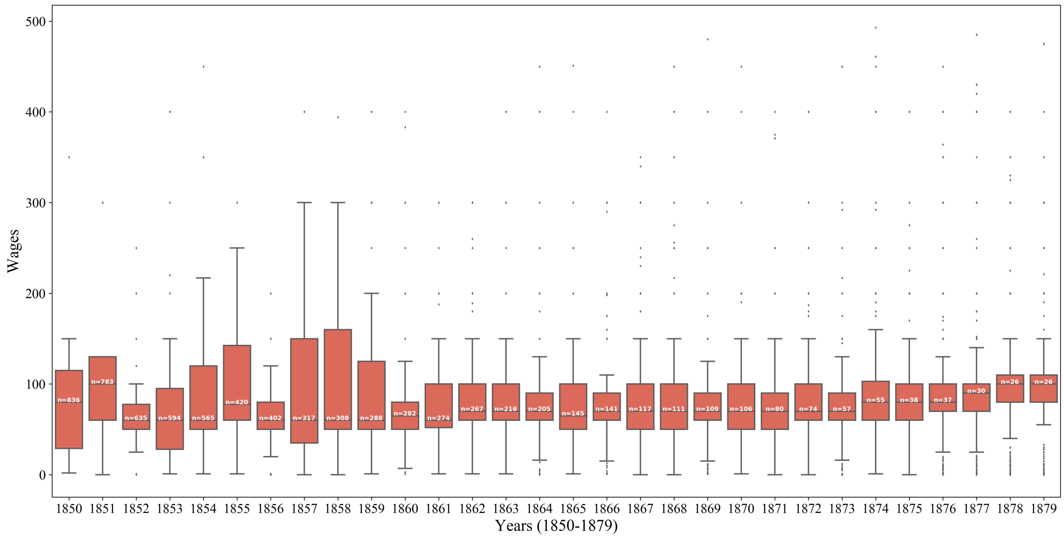


Figure 3: Box plots of quantitative wages extracted for the occupations *keukenmeid* (“kitchenmaid”), *meid* (“maid”), *huishoudster* (“housekeeper”) and *bediende* (“servant”) in the years between 1850 and 1879. Inside the boxplots, the number of observations for every year is included.

60 Million English Death and Probate Records, 1892-2016. Technical report, London School of Economics & Political Science (LSE).

de Zwart, P., van Leeuwen, B., and van Leeuwen-Li, J. (2014). Real wages since 1820. In *How Was Life? Global well-being since 1820*, pages 73–86. Organisation for Economic Co-operation and Development, October.

Feinstein, C. H. (1998). Pessimism Perpetuated: Real Wages and the Standard of Living in Britain during and after the Industrial Revolution. *Journal of Economic History*, 58(03):625–658.

Gray, R. (2018). Selection Bias in Historical Housing Data. Technical report, Queen’s University Belfast, Belfast.

Humphries, J. and Schneider, B. (2019). Spinning the industrial revolution. *The Economic History Review*, 72(1):126–155.

Humphries, J. and Weisdorf, J. (2019). Unreal Wages? Real Income and Economic Growth in England, 1260–1850. *The Economic Journal*, 129(623):2867–2887.

Humphries, J. (2003). Child labor: Lessons from the historical experience of today’s industrial economies. *The World Bank Economic Review*, 17(2):175–196.

Knigge, A. (2016). Beyond the parental generation: The influence of grandfathers and great-grandfathers on status attainment. *Demography*, 53.

Mandemakers, K., Mourits, R., and Muurling, S. (2019). HSN_hisco_release_2018_01, December. Publisher: IISH Data Collection type: dataset.

Reynaert, M. (2008). Non-interactive ocr post-correction for giga-scale digitization projects. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 617–630. Springer.

Schulz, W., Maas, I., and van Leeuwen, M. H. D. (2014). Employer’s choice – Selection through job advertisements in the nineteenth and twentieth centuries. *Research in Social Stratification and Mobility*, 36:49–68, June.

Stephenson, J. Z. (2018). ‘Real’ wages? Contractors, workers, and pay in London building trades, 1650–1800. *The Economic History Review*, 71(1):106–132.

van Leeuwen, M. H. D., Maas, I., and Miles, A. (2002). *HISCO: Historical international standard classification of occupations*. Leuven University Press, Leuven. OCLC: 49628570.

van Strien, D., Beelen, K., Ardanuy, M. C., Hosseini, K., McGillivray, B., and Colavizza, G. (2020). Assessing the impact of ocr quality on downstream nlp tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020)*, pages 484–496, Valleta, Malta, Feb. SCITEPRESS.

Weischedel, R. and Boschee, E. (2018). What can be accomplished with the state of the art in information extraction? a personal view. *Computational Linguistics*, 44(4):651–658.

Wevers, M. (2017). *Consuming America: A Data-Driven Analysis of the United States as a Reference Culture in Dutch Public Discourse on Consumer Goods, 1890-1990*. Ph.D. thesis, University Utrecht.