Wrangling with Non-Standard Data

Eetu Mäkelä[1], Krista Lagus[1], Leo Lahti[2], Tanja Säily[1], Mikko Tolonen[1], Mika Hämäläinen[1], Samuli Kaislaniemi[3], and Terttu Nevalainen[1]

[1] HELDIG – Helsinki Centre for Digital Humanities, University of Helsinki, Finland
http://heldig.fi
eetu.makela@helsinki.fi
[2] Department of Future Technologies, University of Turku
[3] University of Eastern Finland

Abstract. Research in the digital humanities and computational social sciences requires overcoming complexity in research data, methodology, and research questions. In this article, we show through case studies of three different digital humanities and computational social science projects, that these problems are prevalent, multiform, as well as laborious to counter. Yet, without facilities for acknowledging, detecting, handling and correcting for such bias, any results based on the material will be faulty.
Therefore, we argue for the need for a wider recognition and acknowledgement of the problematic nature of many DH/CSS datasets, and correspondingly of the amount of work required to render such data usable for research. These arguments have implications both for evaluating feasibility and allocation of funding with respect to project proposals, but also in assigning academic value and credit to the labour of cleaning up and documenting datasets of interest.

Keywords: Complexity, Data Issues, Non-Standard Data, Bias, Interpretation, Workflows.

1 Introduction

Research in the digital humanities and computational social sciences requires overcoming complexity in research data, methodology, and research questions. Available datasets have often not been created for research. They are not representative samples of the population of interest, but instead arise from a long process of selection, transformation and evolution, where each step may incur bias due to uneven sampling, changing descriptive practices, or algorithms. For example, in the field of history, a dataset may arise from a set of physical books that remain to us through time, through selective microfilming and automatic transcription, or perhaps a long process of manual curation with different actors and according to different standards. In addition to the problem of what has been preserved, general-purpose corpora and the choices made in processing them do not necessarily cater for the needs of various potential user groups with different research questions [34,12,42]. In the social sciences, datasets can arise from opportunistic and thus incomplete mining of social media. In addition, internet forums and platforms often change structure during their existence, easily biasing temporal subsamples that have been derived from these collections.

Beyond original bias, invariably, a gap exists between what is in the data, and the nuanced human categories of interest. Overcoming this gap requires both inventive application of computational enrichment algorithms as well as qualitative judgements to arrive at a set of derived proxy variables judged to adequately correspond to, and shed light on the phenomenon of interest. Importantly, all such enrichments (such as automated transcription, linguistic or sentiment analysis, or topic modelling) and judgement calls interact with the imbalances in the data in complex ways, necessitating further evaluation for bias and representativeness. For example, hard to parse colloquial language on an Internet forum may be more frequent in certain subforums or for certain user groups, while for historical material, automatic transcription tools often react differently to changing printing practices such as fonts and layouts [19,43].

In this article, we show through case studies of three different digital humanities and computational social science projects, that these problems are prevalent, multiform, as well as laborious to counter. Yet, without facilities for acknowledging, detecting, handling and correcting for such bias, any results based on the material will be faulty [26,4,18,6]. Therefore, we argue for the need for a wider recognition and acknowledgement of the problematic nature of many DH/CSS datasets, and correspondingly of the amount of work required to render such data usable for research. These arguments have implications both for evaluating feasibility and allocation of funding with respect to project proposals, but also in assigning academic value and credit to the labour of cleaning up and documenting datasets of interest.

2 Citizen Mindscapes

Citizen Mindscapes is a multidisciplinary research collective for social media analysis with a focus on Suomi24. Suomi24 is Finland's largest topic-centric social media as well as one of the largest non-English online discussion forums in the world. Two main versions of the dataset have so far been opened for interested researchers through the Language Bank of Finland. The most recent one consists of about 82 million forum comments, which contain about 352 million sentences, or about 4.1 billion tokens written in the Suomi24 forum between 1.1.2001–31.12.2017 [1].

Due to prevalent anonymity, the discussions in Suomi24 are topic-oriented rather than reflecting the identity and image of the author. The material opens unique perspectives for studying the qualitative and quantitative dynamics of interaction, be it emotionally supportive, everyday problem-solving, or conflict-oriented. To explore it, the collective has brought together researchers from social sciences, digital culture, welfare sociology, language technology, and statistical data analysis.

As a ready, "found" dataset, the first task for preparing the dataset for research use was to explore its constitution. Our method of approaching the data included authoring a field guide to the data [22], providing a general introduction to the dataset aimed at researchers. We explain the origins and the context of the data, including the information architecture of the Suomi24 site, which consists of a hierarchy of over 3000 categories and sub-categories, forming a tree of maximum depth 6. An individual may either start a new thread or comment on an existing one. On the topmost level consisting of about

20 main themes, the greatest number of words are written in Society, Relationships and Health.

In the course of exploring the data and the possible ways of asking research questions, the complexity of the dataset, the changing technical and socio-cultural environment related to its production, as well as the lack of understanding within social sciences on how to ask research questions of this type of data, have, while leading to some apparent successes in research topics, also led to identifying a number of challenges that this type of data brings to fore.

First, when we examined the dates of messages in the dataset as well as the ID numbers, the yearly and daily statistics uncovered that a significant part of the comments written by users had gone missing during two years in particular. Our attempts to find out the reasons of this lack from Aller, the owner of the Suomi24 site, proved futile during the first years, necessitating that a warning, and fuller examination of the potential for this omission to cause bias be included in the field guide. However, years later, by happenstance the proper person to explain the disparity at Aller was located, and the missing data recovered, further outlining how relying on external providers for data can be problematic.

A second anomaly in the data discovered through frequency analysis was that some days exhibited distinct peaks in terms of articles written on the site. In discussions with the moderators of Suomi24, the reason was uncovered: during those days a "bot", a program spamming the site, had created lots of new posts.

Concerning our knowledge of the users, within the initial batch of data, in 2001–2016, there were 59 612 registered users, and they wrote 7.2% of the comments on the site [22]. Moreover, a survey of the users of Suomi24 was conducted, to find out who the users are, what purposes they use the forum for, and what are their experiences while on it. Upon closer analysis of the about 1400 respondents, the most typical user turned out to be a middle-aged male who lives with their spouse in a city, and has an academic background. The most typical age groups were 45–54 and 55–64. [10]

All in all, the anonymous nature of the entire dataset as well as the lack of user registration clearly made it challenging data for social scientists, who typically would like to know in detail the demography of the authors, as well preferably take a representative sample selected based on some demographic criteria. How, then, to approach data which does not have reliable information corresponding to this convention? A more exploratory approach needed to be uncovered. It seems apparent that while the dataset is very big, and statistical methods are necessary, the dataset is best approached as very large qualitative data. The question then becomes, how to do qualitative research of over 80 million comments, and how can NLP and statistics help in this?

First, however, we had to make a number of very practical decisions when importing and preprocessing the data. These included: (1) what to do with the X-rated sub-category "Sex", (2) should we split into sentences texts that did not use any sentence delimiters, (3) should we store also the messages deleted by moderators, and (4) which of the available metadata were important to be included in the KORP search engine? In the first import, the "Sex" subcategory was not included. The rationale was that because the subcategory allowed only users who were over 18 years of age, we could not as researchers make it

available for everyone. However, later on we criticized our early decision and the sex discussions were included in an update of the data.

During preprocessing the dataset for the KORP engine, it turned out that some authors did not use any markup for delimiting sentences. So, a form of sentence segmentation was designed and implemented by a researcher from the Language Bank of Finland. Regarding the messages (often entire threads) deleted by the moderators, but nevertheless present in the database, we decided that since they were deleted from the site, they should not be part of our dataset, either. Instead, in the place of the deleted message, in the user interface of Suomi24 appeared a 5-word notice "this message has been deleted". We then considered it as part of the dataset, instead of being noise to be removed. However later, when looking at word-level statistics, the effect of this choice was evident as a frequency peak in the overall word distribution statistics, and in the high peak of messages consisting of 5 words. These examples highlight the difficulty of drawing the line between what is data and what is noise, and of the unforeseen consequences to some later statistical analyses. It also becomes essential for subsequent researchers to know of the early preprocessing decisions made.

Eventually, we began to ask specific research questions of the data. The first such attempt was to examine whether there had been a change in people's attitudes concerning welfare users, during a time when national expenditure on social welfare radically increased due to the recession of 2008. When looking at the relative and absolute frequencies of usage of the term "social bum" (sossupummi), it turned out that the use of the term in its different inflected forms had indeed increased, both relatively and absolutely, during and right after 2008 [36]. Subsequently, qualitative analyses of the discussions identified with this term helped researchers in exploring in more detail what had changed in people's opinions concerning welfare users.

In another study, we examined the national landscape of emotions, focusing in particular on two emotions: happiness/joy and fear/worry [20]. For each, a vocabulary based term expansion was first conducted to obtain a list of alternative terms. Out of these, the ones that were sufficiently frequent as well as accurate in picking out expressions of these emotions (based on a small-scale evaluation made by a human on a random sample) were kept in the vocabulary. Then the relative frequencies of these terms were observed over time, in four ways: as a trend over the years, and concerning the monthly, weekly and daily circular rhythms that the relative frequencies exhibited. A number of interesting findings concerning the rhythms of emotion expressions on the site could be seen. Most joyous times were typically evenings, where the relative frequency of expressing joy increased until about midnight, then a clear decrease was observed. In contrast, the fear and anxiety expressions peaked a bit later, centering in the night hours 00–04. In a related manner, the most joyous times of the year were summer in general, as well as the months leading to Christmas. Fear and anxiety peaked in January as well as in August/September. Due to the anonymous nature of the material however, we did not really know whether our observations described the changing tone of individual emotional landscapes (e.g. due to changes in alcohol consumption and then joy decreasing towards the end of the evening) or rather just the changing tone of the forum itself, during day and night hours when different individuals come, leave their comment, and then again

leave. Nevertheless, these are indeed the time-related general tones when considering such a social media environment as a discussion landscape, with its own rhythms.

In a third example, we studied how people talked about medicines [21,35]. The first article describes the development of a tool for semi-automated design of a proto-ontology of concepts, and the statistical methods for deriving their associations. The second showcases how a social scientist interested in people's discussions about a particular medicine might utilize this tool to understand people's discussions around the medicines – a form of qualitative analysis based on big data, facilitated by a suitable tool that resembles the way in which a social scientist might approach the material. The particular value of this example is the study of medicines and symptoms as experienced by the users themselves. For example, the frequency of cannabis mentions surprised our data analysts. On the other hand, some medical experts were surprised by the depth of the suffering that was revealed in the medicine and symptom discussions. The discussions seem to be a rich form of human experiences that can be brought under study, and that can enrich our understanding of various groups of people whose voices may not currently be heard.

Each of these examples showcase the use of both a statistical analysis of big data in some way, and a qualitative analysis made by a researcher reading through some parts of the dataset. From a social scientists' perspective this big data cannot be considered as "statistically representative" of anything but itself – rather, it is a very big qualitative dataset of what people write anonymously, when trying to understand or express their own lives. However, such as it is, it offers an invaluable and rich lens for observing the human condition. People write for their own reasons and not induced by the viewpoints of the researcher. The question then becomes, how can natural language processing tools, statistics, visualizations, and search engines provide means for both distant reading and close reading of the material, which help the researcher in first formulating and later perhaps answering a specific question concerning the people? In this project, we have made headway, but many questions still remain for further research.

The data is now available in several ways: 1) downloadable by all researchers in full or in parts from the Language Bank of Finland, either in the original JSON format, or 2) in text-only, 3) searchable for everyone interested using KORP[4] and 4) available to be queried from an SQL database hosted by the CSC, particularly suited for use in connection with RStudio and such. Moreover, the Health subsection of 19 million comments has been analysed in terms of medicines and medicine-like names and their connections, and is available for searching using medicine or symptom concepts in the Medicine Radar tool[5]. The code for doing such analyses is provided as open source, see [21].

The act of opening the dataset as well as carefully providing sufficient information of, and tools and examples concerning how to access it has shown to be an effective way to facilitate a larger learning community of researchers. As the number of users of the dataset increases, we are led to believe that multiple research projects besides our own are already researching this data. It is also a topic on some courses teaching social scientists how to work on big data. Finally, engaging a multidisciplinary community of

[4] `http://korp.csc.fi/`
[5] `http://laaketutka.fi/`

researchers around the same open dataset can be seen as a fruitful way to eventually discover new ways of knowledge formation within the social sciences.

3 History of Knowledge Production

In the Helsinki Computational History Group, we have analysed historical patterns of knowledge production in Europe based on document metadata from various library catalogues [25], primarily based on the English Short Title Catalogue (ESTC) [24] and the Finnish and Swedish national bibliographies [44]. These collections include bibliographic records from altogether hundreds of thousands of documents that have been printed in Europe and elsewhere between 1450 and the early twentieth century.

As in the case of Suomi24, for research purposes these databases also have to be considered found data, albeit of quite a different sort. In contrast to Suomi24, while these bibliographic databases are not created for research, they still do aim to be as complete as possible, and follow stringent library cataloguing rules and standards.

Nonetheless, there are many problems in using the catalogues as data for research. The first problem arises from the cataloguing and technical standards themselves, which state that information on e.g. authors and places of publication is to be entered exactly as they are given on the publication title page, instead of any standard notation. This is further compounded by the fact that the MARC21 technical record format used for cataloguing is a flat, book-centric format where all attributes are textual. Taken together, these properties mean that for example the same author can appear in the metadata many times under different name variants, without any explicit links tying them to each other. Similarly, different editions of the same work are in no way formally linked to each other.

Both of these problems are further compounded by the complexity of the historical phenomenon itself. For example, multiple works were often also bound together into a new release, blurring any clear boundaries between what can be considered an edition of the same work, and what should be considered a new one. Problematically, new editions of works also often changed titles significantly, rendering their later joining through metadata alone a much harder task. However, because in the data popular works sometimes add up to tens or even hundreds of distinct editions, it is imperative to account for the bias they may cause, a task towards which the project has expended significant resources [13].

Similarly, while the libraries do maintain authority files containing variant forms for people's names, still the task of identifying the same authors is not trivial due to for example the high number of authors writing their name only in initials on the title page, or the prevalence of publishing anonymously [11]. In addition, for fields such as publisher information, what has been stored are the full publication statements as they appear, actually containing publishers, printers as well as other actors in varying sentence constructions.

Even for cleaning and harmonizing seemingly simple fields such as publication years, the data is problematic due to the uncertainty inherent in cataloguing historical books after the fact [24]. For example, of the multiple ways in which dating uncertainty has been encoded, one has been to just round the approximate year of publication to the

nearest larger number. When looking at the data, this then ends up showing abnormal peaks in the amount of publications every five, ten and 50 years.

Beyond these problems, there are even more fundamental sources of error when library catalogues are taken as proxies for historical reality. Not all printed books remain until today, or are even known to historical records. Importantly for research, the surviving documents can not be considered to be a random, representative sample of what was, as they reflect the interests of collectors. For example, forms of popular print, such as chapbooks and ballads, are less likely to survive compared to popular books, especially if going through multiple editions [30], and occasional fires may have destroyed dedicated libraries that have focused on specific themes. On the other hand, sometimes it can be exactly the small books that survive through a historic quirk. As a concrete example, 80% of the publications between 1640 and 1660 in the ESTC come from the collection of George Thomason, who collected altogether some 22 000 short pamphlets and periodicals related to the English Civil War that was ongoing at that time. Because Thomason was a completionist, many of these individual items are near copies of each other, just printed at slightly different times or by a different printer.

Finally, bias may also arise due to the long time span under which the cataloguing process has been executed. Even disregarding the differences between individual indexers, metadata cataloguing standards themselves have changed considerably from the 1970s when gathering the ESTC was begun, resulting in heterogeneous records.

Yet all in all, the ESTC can at least nowadays be claimed to cover most of the books that remain in significant libraries, even though it is still being updated by the participating libraries discovering and cataloguing more material [41]. This unfortunately is often not the case. For example, when we sought to complement the efforts on the library catalogue metadata with the analysis of full texts collections, we turned to the Eighteenth Century Collections Online (ECCO) dataset, which claims to "contain every significant English-language and foreign-language title printed in the United Kingdom between the years 1701 and 1800"[6]. However, since its initial release of some 135 000 works in 2002, 47 000 further works have been added. The reason for this is that the material in ECCO comes in turn from the progress of a microfilming project "the Eighteenth Century", started in 1982, which itself follows the ESTC cataloguing project. In fact, at present ECCO contains only slightly more than 50% of the eighteenth century works known to the ESTC. ECCO is also by no means unique. See for example [33] for a similar description of the history of the Burney Collection of newspapers.

Even after correcting for sampling bias, the material remains unevenly noisy. In scanning projects that have been going on for a long time, tools ranging from the microfilm scanners to OCR engines have also changed, resulting in different amounts of noise in different sections of a dataset. For example, the digitization process for newspapers at the Finnish national library [32] has used 22 different versions of OCR software during its 13 years of scanning.

It is clear that without facilities for detecting and handling such immense bias, any results based on the material will be faulty. As described in more detail elsewhere [25], we have designed and implemented a dedicated data science ecosystem to harmonize, integrate, and analyse the datasets. This *bibliographic data science* ecosystem is specif-

[6] `http://gale.com/c/eighteenth-century-collections-online-part-i`

ically tailored to interpret the fields of the investigated datasets by using a combination of existing machine learning and data science algorithm libraries and custom code.

The approach is very collaborative, and requires a variety of skills ranging from data sciences to linguistics and humanities [28]. As our experience has shown, active research use helps us to ensure and improve the quality of data harmonization and correct any shortcomings as soon as they are observed. Moreover, as a side product of research, we can constantly extend the workflow to incorporate new data fields. Automation of the workflow is helping us to scale the harmonization efforts up to millions of documents [45], while efficiently monitoring data quality based on unit tests and semi-automated curation processes that identify potential outliers and inconsistencies in the data, such as ambiguous author or institution names, unlikely publication years, or duplicated entries.

Overall, these efforts have specifically highlighted that some of the key elements in ensuring data quality both in metadata as well as full text context include (i) scalable and modular automation of the data processing; (ii) transparent and open documentation and sharing of the algorithmic methods and workflows both within and beyond our own research group; and (iii) the importance of actively maintaining the research workflows, which can best be realized through continued research use and critical examination of the data in the broader context of the study field. Observing and quantifying the effect of potential biases, such as lost document collections, remains challenging. Often the data alone does not contain sufficient information on the presence or magnitude of such effects, but when combined with complementary historical knowledge it can provide a valuable source of information to support or refute alternative hypotheses.

4 STRATAS

The STRATAS project, funded by the Academy of Finland for 2016–2019, is concerned with combining structured and unstructured data in sociolinguistic research on language change. Its key components include (1) evaluating the quality of the previously compiled Corpora of Early English Correspondence (CEEC, 1402–1800) [31] and (2) using the corpora to analyse who creates and adopts new vocabulary in the history of English, and whether there are differences across social groups.

Importantly, in contrast to the other datasets described in this paper, the CEEC was specifically designed for the purposes of historical sociolinguistics, and as such it aims at social representativeness in terms of e.g. gender and social rank. Previous research into the history of English lexis has focused on published texts, which were mostly written by highly educated men. By analysing personal correspondence, a genre that was accessible to anyone who was literate, we are able to consider the language use of women and the lower social ranks as well.

Yet, because the corpus is based on published editions of manuscript letters, it is still somewhat skewed towards well-educated men of the upper ranks, as they were the most literate social group, and it was mostly their letters which were considered important enough to be preserved and later edited. This skew in the archival record decreases over time, which is reflected in the edited record. Nonetheless, even in the eighteenth-century section of the corpus, only about a quarter of the running words were written by women – although this figure rises to a third by the end of the century. The share of middle-

and lower-class writers, belonging to the social ranks below gentry, increases over time as well (see Fig. 1). These changes in the composition of the corpus may be regarded as an additional source of bias that complicates diachronic comparisons, but they also reflect real-world social changes, such as increasing levels of education and the rise of the middle classes in terms of wealth and power in the later eighteenth century [16].

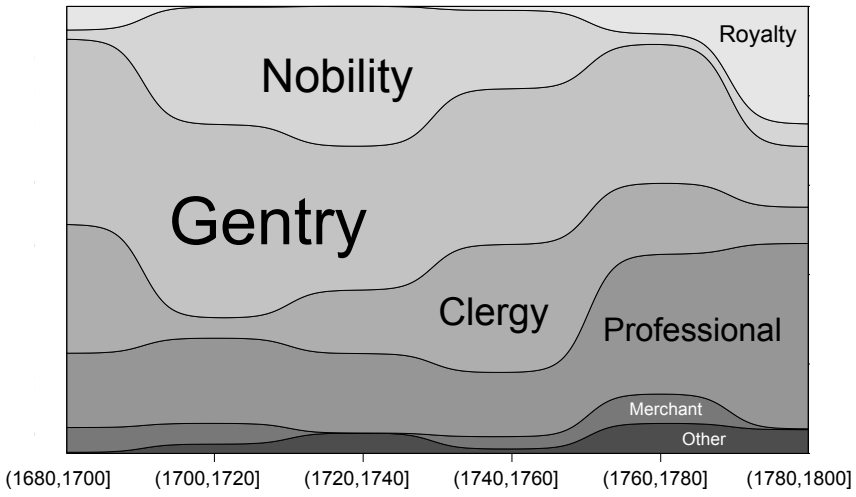

Fig. 1. Density plot of the distribution of letters by social rank in the eighteenth-century section of the CEEC (courtesy of Harri Siirtola). From [16]: Fig. 4.4.

Another aspect of the composition of the corpus and a potential source of bias is the relationship between the sender and recipient of the letter. While letters to friends and nuclear family members comprise the bulk of the corpus, there is a non-negligible amount of letters written to more distant family members and other acquaintances. One factor correlating with sender-recipient relationships is gender. For instance, comparisons between genders may be skewed by the fact that there are more business letters by men and more family letters by women, who were more confined to the private sphere. Moreover, the proportions of these change over time [46].

Importantly for this paper however, the metadata regarding social attributes has been hand-recorded for each letter and person in the corpus. Thus, despite the human interpretation required to determine e.g. an individual's social status, the historical knowledge of the annotators can be said to make this aspect of the data trustworthy up to the level required for research. Therefore, while the collection as a whole may not be perfectly bal-

anced, the bias is always visible and evaluable, allowing researchers to make reasoned subsamples weighing coverage and balance to their liking.

In addition to being socially representative (with the caveats discussed above), for research the CEEC has the advantage of being relatively large, at 1180 writers, 11 713 letters, and 5.2 million running words. While this is nothing compared to the massive historical databases of published texts available for English, it is certainly big data for a collection of private writing. Furthermore, the CEEC is free from optical character recognition (OCR) errors, as the compilers of the corpus who digitized the letter editions proofread the corpus texts multiple times, which they were only able to do thanks to the manageable size of the corpus.

The CEEC compilers wanted to study authentic language use, and thus the corpus is compiled from editions that retain original spelling. English spelling did not develop a prescribed standard until the eighteenth century, and there is great spelling variation in historical English texts – and in the CEEC. However, editions vary in their representation of manuscript sources (e.g. some silently expand frequent abbreviations like *y*^{e} 'the'), which creates a further level of bias into the dataset, raising questions about how authentic the 'original' spelling in an edition is. To remedy this, one part of the STRATAS project was to evaluate the philological reliability of the editions used in the CEEC [40]. The results show that despite great variation in the reliability of its sources, the corpus as a whole does represent authentic manuscript orthography, particularly when used at scale [17].

While the CEEC retaining as much spelling variation as possible is very useful for research on orthography, it poses problems for other kinds of linguistic constructs, particularly those targeting word types, where all different spelling variants of the words should be counted as one. There is a standardized-spelling version of the corpus, but the program used to process the texts (VARD2 [3]) does not normalize low-frequency forms. Yet it is of course the low-frequency forms which are often novel vocabulary and thus interesting.

As an example of the consequences of this, our case studies of the productivity of nominal suffixes *-ness*, *-ity* and *-er* required casting a very wide net to catch all the different spelling variants of these endings. In the case of *-er*, through domain knowledge and experimenting, we determined that the core unit of its spelling was an <r> character towards the end of the word. This unit could be preceded by anything and followed by an optional <e> character as well as optional plural/possessive forms. Possible variants therefore included e.g. *-er(e)*, *-ar(e)*, *-or(e)*, *-our(e)*, *-owr(e)* and *-ur(e)*, and their plural and possessive forms. There were also many *-er* words denoting occupations that were sometimes abbreviated in such a way that the <r> character was lost (e.g. *tailo~* for *tailor* and *carpunt~s* for *carpenters*). These could be captured by searching for the abbreviation marker <~> preceded by <e> or <o>, or followed by plural/possessive forms. But with this wide a query, many more words matched the query than actually in the end represented the phenomena of interest. Concretely, out of the 6800 word types returned by our query for the suffix *-er*, only 1720 turned to actually be *-er* words [37,39,38].

This variation became even more problematic when we shifted our attention to neologisms, or new vocabulary in general. Here, in total, the CEEC contains about 150 000 distinct word forms. By excluding those tagged as foreign, proper nouns or abbrevia-

tions, about 125 000 remain. Our initial idea was to take this vocabulary, and match it against the Oxford English Dictionary (OED) and the Middle English Dictionary (MED) to discover which words appeared in the CEEC earlier than mentioned in the dictionaries. However, even though the OED and MED contain some 375 000 and 260 000 spelling variants for their 280 000 and 60 000 words, only some 36 000 word forms from the CEEC could be mapped to the dictionaries, while 88 000, or three quarters, could not.

Thus, it was clear that more tools for filtering and normalization were required. When looking at the data, we saw that the proportion of words not in dictionaries decreased through time. Thus, for one case study, we cut the material to only those 32 000 appearing in the long 18th century. By using existing methods for normalization such as the variant forms recorded in the OED and a tool called MorphAdorner [5], we were able to match a total of 10 100 words against the OED. Further, we also utilized multiple large databases of published texts, including the British Library Newspapers, Burney & Nichols Collections, Eighteenth Century Collections Online and Early English Books Online (altogether 51.4 billion running words 1400–1900) to filter out spurious matches, requiring that neologism candidates could appear at most 100 times before their CEEC attestation in the published texts (the number 100 was chosen due to the published texts being riddled with OCR errors, which could spuriously filter out neologisms using a tighter limit).

Through this filtering, we finally ended up with 220 candidates, from which we manually identified 82 neologisms which we could then contribute back to the OED. However, from the viewpoint of sociohistorical analysis, this pipeline did not lead to a useful conclusion, as we were still unable to account for 22 000 words, or two thirds of the overall material in the CEEC for this period. Further, we had no idea of how representative the neologisms we did find were of the whole, particularly with the spelling variation not being equally distributed across the social ranks.

By this time, we had further developed and tested multiple different techniques for automatic normalization, finally ending up with a neural machine translation -based approach[7] to discover further candidate matches against the MED and the OED [8,9]. To better understand the reliability of this new version of our pipeline, we conducted a second experiment by subsampling a human-processable number of letters written between 1640–1660 (around the English Civil War), with as equal as possible representation by gender, social rank and relationship. Then, we filtered the sample to include only words that appeared in the CEEC for the first time during this period (as those that appeared there already earlier could not have been neologisms). In the end, we subsampled 836 candidates to evaluate in the dedicated filtering and categorization tool FiCa we developed for this purpose [38], or 9.3% out of the 8954 candidates there were in the entire corpus for 1640–1660.

Upon evaluation, out of these 836 words, 102 turned out to be proper nouns and 3 foreign words, as only part of the sample used in this experiment belonged to the POS-tagged version of the corpus, which would have enabled the elimination of these words beforehand. After removing these, out of the 731 words that remained, 437 (60%) got the right lemma after normalization, while for 114 (16%), no lemma matches were

[7] The normalization model has been released as a part of an open source Python library called NATAS: `http://github.com/mikahama/natas`

found. However, even more worrisomely, 185 or 25% of the words in the sample did get a lemma match, but upon evaluation it turned out to be the wrong lemma! Finally, even out of the 437 that were matched to the right lemma, 77 or 17% were matched to the wrong part of speech, which in this data equates to a possibly wrong earliest attestation date. Actually for this data, even if the lemmatization were perfect, 114 out of the 731 words or about 20% would still get mapped to the wrong entry without further disambiguation based on part of speech or other contextual information.

Taken together, this evaluation showed that despite our best efforts and state of the art approach to discovering candidates, significant error sources still remain that currently preclude any fully automatic application of the pipeline that does not include a manual verification component. Therefore, for the time being, we are still limited to asking more focused questions that can be answered by combining the tools we do have, with a bearable amount of manual labour. For example, by switching from the general question of "Who uses new words" to the much more focused one of "Who uses new words in the 18th century that come into general use later" (as shown by either our comparison corpora or dictionaries), we are left with 2500 candidates instead of the original 32 000. On the other hand, if we're interested in nonce words that are used only once and never seen again, there are some 9000 candidates for such in the 18th century, an amount that is still barely manageable. Importantly for this paper, we believe this state of affairs will continue for some time still, necessitating work to focus not only on improving the automated components of the pipeline, but also on improving the tools for making manual interpretations and decisions in an expedient way.

5 Discussion

As shown by the cases presented here, it is not an exception but the norm that projects in the digital humanities and social sciences have to deal with both unprecedented levels of data complexity, as well as gaps between their objects of interest and what can currently be reliably attained by computational means.

Dealing with all these types of uncertainty and bias will need handling beyond just evaluating the suitability of various statistical significance tests for complex data [27]. In some of the cases of missing data, the situation can be formulated as akin to a quasi-experimental design [2], allowing the scholar to enumerate and rule out alternative explanations based on statistical comparisons between either in-data categories or outside information (such as in the case of ECCO, accounting for selection bias through comparing the types of works in ECCO to those in the ESTC as a whole) [7]. However, due to the data being rich and complex, there are often very many potential confounding variables to consider. Therefore, ruling out alternative causes for phenomena will also require interactional support, such as abilities to quickly highlight data distributions and gaps along each metadata axis, as well as their combinations. It is also not enough to do this only on the level of a whole dataset, but this must be integrated with exploratory user interaction in order to measure the bias for a particular subset of interest to the scholar.

Moreover, because some bias sources are essentially statistically unverifiable due to e.g. comparative data having been lost to history, research must also tackle how to combine statistical analyses with qualitative judgement based on the theoretical under-

standing of the scholar [2,7]. Combined with the current inability of computational approaches to directly produce proxy variables that measure the objects of interest, research by necessity thus needs to include large amounts of interleaved computational inference as well as manual interpretation to produce the final data conclusions are based on [15]. Current tools however make such interplay between computation and analysis arduous, forcing scholarly inquiry into a mould it cannot fit. In short, what would be needed are computational quantitative tools that would not clash with, but truly integrate with, and enhance the process of trustworthy scholarly inquiry.

In practical terms, we wish to highlight open science as a useful approach to solving the data problems described here [14,29]. First, the reliability and quality of research can be improved when the data, methods, and results can be independently investigated, verified, and criticized [14]. Even more importantly, in all of the projects presented here, the datasets under operation have not been collected solely for a single project. On the contrary, they are liable to be of interest to a wide pool of researchers. Yet, while people will want to pose very different questions to the data, all need to uncover and solve the same issues with regard to biases and other problems in the data. Therefore, the overall research process can become much more efficient if people share their data-related discoveries, documentation and clean-up pipelines openly, making them available for reuse and further modification, for instance based on open licensing [23].

To close, we wish to emphasize the fact that none of the problems discussed above have been solved in a general manner yet, and neither are there already existing field guides for most materials. In short, there are no established protocols for the types of research attempted here. Therefore, in the foreseeable future, if they seek to produce trustworthy results, most digital humanities projects will need to enact similar explorations as have been done here. Yet, this fact does not seem to be commonly understood or acknowledged, either by applicants drafting applications, funders evaluating proposals, committees evaluating scientific output or by researchers making use of data cleaned previously by another project. In our opinion, all of these would do well to place more value and weight on the currently largely unseen work that goes into moving from raw data to results.

References

1. Suomi 24 corpus (2017H2), `http://urn.fi/urn:nbn:fi:lb-2019010801`
2. Anderson-Cook, C.M.: Experimental and Quasi-Experimental designs for generalized causal inference. J. Am. Stat. Assoc. **100**(470), 708–708 (Jun 2005).
3. Baron, A., Rayson, P.: VARD2: A tool for dealing with spelling variation in historical corpora. In: Postgraduate Conference in Corpus Linguistics. Aston University, Birmingham (2008).
4. Boyd, D., Crawford, K.: Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. **15**(5), 662–679 (2012).
5. Burns, P.R.: MorphAdorner v2: A Java library for the morphological adornment of English language texts. Northwestern University, Evanston, IL (2013).
6. Cihon, P., Yasseri, T.: A biased review of biases in Twitter studies on political collective action. Frontiers in Physics **4**, 34 (2016).
7. Cuddeback, G., Wilson, E., Orme, J.G., Combs-Orme, T.: Detecting and statistically correcting sample selection bias. J. Soc. Serv. Res. **30**(3), 19–33 (May 2004).

8. Hämäläinen, M., Säily, T., Rueter, J., Tiedemann, J., Mäkelä, E.: Normalizing early English letters to Present-day English spelling. In: Alex, B., Degaetano-Ortlieb, S., Feldman, A., Kazantseva, A., Reiter, N., Szpakowicz, S. (eds.) Proceedings of the second joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL-2018). pp. 87–96. ACL Anthology, Association for Computational Linguistics, Stroudsburg, Pennsylvania (2018), http://aclweb.org/anthology/W18-4510
9. Hämäläinen, M., Säily, T., Rueter, J., Tiedemann, J., Mäkelä, E.: Revisiting NMT for normalization of early English letters. In: Alex, B., Degaetano-Ortlieb, S., Kazantseva, A., Reiter, N., Szpakowicz, S. (eds.) Proceedings of the third joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL-2019). pp. 71–75. ACL Anthology, Association for Computational Linguistics, Stroudsburg, Pennsylvania (2019), https://www.aclweb.org/anthology/W19-2509
10. Harju, A.: Suomi24-keskustelut kohtaamisten ja törmäysten tilana. Media & viestintä **41**(1) (2018).
11. Hill, M.J., Vaara, V., Säily, T., Lahti, L., Tolonen, M.: Reconstructing intellectual networks: From the ESTC's bibliographic metadata to historical material. In: Navarretta, C., Agirrezabal, M., Maegaard, B. (eds.) Proceedings of the Digital Humanities in the Nordic Countries (DHN2019). Copenhagen (March 2019), https://cst.dk/DHN2019Pro/papers/DHN2019_hill.pdf
12. Hiltunen, T., McVeigh, J., Säily, T. (eds.): Big and rich data in English corpus linguistics: Methods and explorations. No. 19 in Studies in Variation, Contacts and Change in English, VARIENG, Helsinki (2017), http://www.helsinki.fi/varieng/series/volumes/19/
13. Ijaz, A., Roivainen, H., Lahti, L.: Analytical edition detection in bibliographic metadata. In: Proceedings of the Digital Humanities (DH2019) (July 2019), in press.
14. Ioannidis, J.: How to make more published research true. PLoS Medicine **11**(10), e1001747 (2014). https://doi.org/10.1371/journal.pmed.1001747
15. Isoaho, K., Gritsenko, D., Mäkelä, E.: Topic modeling and text analysis for qualitative policy research. Policy Studies Journal (2019). https://doi.org/10.1111/psj.12343
16. Kaislaniemi, S.: The Corpus of Early English Correspondence Extension (CEECE). In: Nevalainen, T., Palander-Collin, M., Säily, T. (eds.) Patterns of change in 18th-century English: A sociolinguistic approach, pp. 45–59. No. 8 in Advances in Historical Sociolinguistics, John Benjamins, Amsterdam (2018). https://doi.org/10.1075/ahs.8.04kai
17. Kaislaniemi, S., Evans, M., Juvonen, T., Sairio, A.: "A graphic system which leads its own linguistic life"? Epistolary spelling in English, 1400–1800. In: Säily, T., Nurmi, A., Palander-Collin, M., Auer, A. (eds.) Exploring Future Paths for Historical Sociolinguistics, pp. 187–213. No. 7 in Advances in Historical Sociolinguistics, John Benjamins, Amsterdam (2017). https://doi.org/10.1075/ahs.7.08kai
18. Kaplan, R.M., Chambers, D.A., Glasgow, R.E.: Big data and large sample size: a cautionary note on the potential for bias. Clin. Transl. Sci. **7**(4), 342–346 (Aug 2014).
19. Kettunen, K., Pääkkönen, T., Koistinen, M.: Between diachrony and synchrony: Evaluation of lexical quality of a digitized historical Finnish newspaper and journal collection with morphological analyzers. In: Baltic HLT. pp. 122–129 (2016).
20. Lagus, K., Pantzar, M., Ruckenstein, M.: Kansallisen tunnemaiseman rakentuminen: Pelon ja ilon rytmit verkkokeskusteluissa. Kulutustutkimus.Nyt **12**(1-2), 62–83 (2018), http://www.kulutustutkimus.net/wp-content/uploads/2018/11/KTN_vol12_Lagus-Pantzar-and-Ruckenstein.pdf

21. Lagus, K.H., Ruckenstein, M.S., Juvonen, A., Rajani, C.: Medicine radar – a tool for exploring online health discussions. In: Mäkelä, E., Tolonen, M., Tuominen, J. (eds.) Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2084/short21.pdf
22. Lagus, K.H., Ruckenstein, M.S., Pantzar, M., Ylisiurua, M.J.: Suomi24 – muodonantoa aineistolle (Suomi24 – giving shape to the data set). Valtiotieteellisen tiedekunnan julkaisuja **10** (2016), http://hdl.handle.net/10138/163190
23. Lahti, L.: Open data science. In: Advances in Intelligent Data Analysis XVII. Lecture Notes in Computer Science 11191. vol. 11191. Springer, India (October 2018), conference proceedings.
24. Lahti, L., Ilomäki, N., Tolonen, M.: A quantitative study of history in the English short-title catalogue (ESTC) 1470–1800. LIBER Quarterly **25**(2), 87–116 (12 2015). https://doi.org/10.18352/lq.10112
25. Lahti, L., Marjanen, J., Roivainen, H., Tolonen, M.: Bibliographic data science and the history of the book (c. 1500–1800). Cataloging & Classification Quarterly pp. 1–19 (January 2019). https://doi.org/10.1080/01639374.2018.1543747, special issue.
26. Lazer, D., Kennedy, R., King, G., Vespignani, A.: The parable of Google flu: Traps in big data analysis. Science **343**(6176), 1203–1205 (Mar 2014).
27. Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., Mannila, H.: Significance testing of word frequencies in corpora. Lit Linguist Computing **31**(2), 374–397 (Jun 2016).
28. Mäkelä, E., Tolonen, M., Marjanen, J., Kanner, A., Vaara, V., Lahti, L.: Interdisciplinary collaboration in studying newspaper materiality. In: Navarretta, C., Agirrezabal, M., Maegaard, B. (eds.) Proceedings of the Twin Talks workshop. Digital Humanities in the Nordic Countries (DHN2019). Copenhagen (March 2019), http://ceur-ws.org/Vol-2365/7-TwinTalks-DHN2019_paper_7.pdf
29. Morin, A., Urban, J., Adams, P.D., Foster, I., Sali, A., Bak er, D., Sliz, P.: Research priorities. Shining light into black boxes. Science (New York, N.Y.) **336**(6078), 159–60 (Apr 2012). https://doi.org/10.1126/science.1218263, http://www.sciencemag.org/content/336/6078/159.short
30. Murphy, K., O'Driscoll, S.: Studies in Ephemera: Text and Image in Eighteenth-Century Print. Bucknell University Press, Lewisburg, Pennsylvania (2013).
31. Nevalainen, T., Raumolin-Brunberg, H., Keränen, J., Nevala, M., Nurmi, A., Palander-Collin, M., Kaislaniemi, S., Laitinen, M., Säily, T., Sairio, A.: CEEC, Corpora of Early English Correspondence. Department of Modern Languages, University of Helsinki (1998–2006), http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/
32. Pääkkönen, T., Kervinen, J., Kettunen, K.: Digitisation and digital library presentation system – a Resource-Conscientious approach. In: Mäkelä, E., Tolonen, M., Tuominen, J. (eds.) Proceedings of the Digital Humanities in the Nordic Countries 2018. CEUR-WS.org (2018).
33. Prescott, A.: Searching for Dr. Johnson: The digitisation of the Burney Newspaper Collection. In: Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century, pp. 49–71. Brill (May 2018).
34. Rissanen, M.: Three problems connected with the use of diachronic corpora. ICAME Journal **13**, 16–19 (1989), http://clu.uni.no/icame/journal.html
35. Ruckenstein, M.: Tracing medicinal agencies: Antidepressants and life-effects. Social Science & Medicine p. 112368 (2019).
36. Saari, J., Behm, M., Lagus, K.: Sosiaalipummi! – moraalipaniikki 2010-luvun Suomessa. In: Saari, J. (ed.) Sosiaaliturvariippuvuus: sosiaalipummit oleskeluyhteiskunnassa?, pp. 207–232. Tampere University Press (2017), https://tampub.uta.fi/bitstream/handle/10024/100775/Saari_Sosiaaliturvariippuvuus.pdf

37. Säily, T.: Sociolinguistic variation in English derivational productivity: Studies and methods in diachronic corpus linguistics. No. 94 in Mémoires de la Société Néophilologique de Helsinki, Société Néophilologique, Helsinki (2014), http://urn.fi/URN:ISBN:978-951-9040-50-9
38. Säily, T., Mäkelä, E., Hämäläinen, M.: Explorations into the social contexts of neologism use in early English correspondence. Pragmatics & Cognition **25**(1), 30–49 (2018). https://doi.org/10.1075/pc.18001.sai
39. Säily, T., Suomela, J.: *types*2: Exploring word-frequency differences in corpora. In: Hiltunen, T., McVeigh, J., Säily, T. (eds.) Big and rich data in English corpus linguistics: Methods and explorations. No. 19 in Studies in Variation, Contacts and Change in English, VARIENG, Helsinki (2017), http://www.helsinki.fi/varieng/series/volumes/19/saily_suomela/
40. Sairio, A., Kaislaniemi, S., Merikallio, A., Nevalainen, T.: Charting orthographical reliability in a corpus of English historical letters. ICAME Journal **42**(1), 79–96 (2018).
41. Spedding, P.: "The new machine": Discovering the limits of ECCO. Eighteenth Century Stud. **44**(4), 437–453 (2011).
42. Suhr, C., Nevalainen, T., Taavitsainen, I.: From Data to Evidence in English Language Research. Brill, Leiden, The Netherlands (2019), https://brill.com/view/title/54063
43. Tanner, S., Muñoz, T., Ros, P.H.: Measuring mass text digitization quality and usefulness. D-lib Magazine **15**(7/8), 1082–9873 (2009).
44. Tolonen, M., Lahti, L., Roivainen, H., Marjanen, J.: A quantitative approach to book-printing in Sweden and Finland, 1640–1828. Historical Methods: A Journal of Quantitative and Interdisciplinary History **52**, 57–78 (2019). https://doi.org/10.1080/01615440.2018.1526657
45. Tolonen, M., Marjanen, J., Roivainen, H., Lahti, L.: Scaling up bibliographic data science. In: Navarretta, C., Agirrezabal, M., Maegaard, B. (eds.) Proceedings of the Digital Humanities in the Nordic Countries (DHN2019). Copenhagen (March 2019).
46. Vartiainen, T., Säily, T., Hakala, M.: Variation in pronoun frequencies in early English letters: Gender-based or relationship-based? In: Tyrkkö, J., Timofeeva, O., Salenius, M. (eds.) Ex philologia lux: Essays in honour of Leena Kahlas-Tarkka, pp. 233–255. No. 90 in Mémoires de la Société Néophilologique de Helsinki, Société Néophilologique, Helsinki (2013).