Exploring the Quality of the Digital Historical Newspaper Archive KubHist

Yvonne Adesam[1], Dana Dannélls[1], and Nina Tahmasebi[1,2]

[1] Språkbanken, University of Gothenburg, Sweden
[2] Centre for Digital Humanities, University of Gothenburg, Sweden
{yvonne.adesam/dana.dannells/nina.tahmasebi}@gu.se

Abstract. The KubHist Corpus is a massive corpus of Swedish historical newspapers, digitized by the Royal Swedish library, and available through the Språkbanken corpus infrastructure Korp. This paper contains a first overview of the KubHist corpus, exploring some of the difficulties with the data, such as OCR errors and spelling variation, and discussing possible paths for improving the quality and the searchability.

Keywords: Historical newspaper corpus · OCR errors · Spelling normalization

1 Introduction

The past decades have seen a massive increase in digitized, historical documents that have been at the core of a range of different applications, from studies of cultural and language phenomena [14] to temporal information retrieval and extraction. The study of semantic changes, to give one example, has changed character from qualitative studies [17, 16] to automatic detection via topic modeling [13], and word sense induction [15] to methods based on (neural) embeddings [12, 1]. In common for the majority of the existing methods and studies is that they focus on English texts because of the vast amounts of easily available data, for example, through the Google N-gram corpora.

The availability and easy access of datasets like the Google N-gram corpora, and others in full text form, like the Corpus of Historical American English (COHA), and the Penn Parsed Corpora of Historical English, draws researchers to English texts and hence, creates methods developed for the English language.

In Sweden, the amount of digital, historical texts is large compared to many other languages, but still in the shadows of that available for English. There have been few possibilities to make diachronic studies and develop tools for historical Swedish and automatic detection of language changes. The first, large newspaper corpus, KubHist, is a good step towards this goal.

The first version of the KubHist dataset, currently available through the Språkbanken corpus infrastructure Korp [5], contains close to 1.1 billion tokens. Originally going under the name DigiDaily – after the project which led to the digitization of the first batch of historical newspapers, involving the Royal

Swedish Library and the Swedish National Archives (https://riksarkivet.se/digidaily) – the corpus soon changed its name to KubHist (Kungliga bibliotekets historiska tidningar), as more material was added after the end of the project. Språkbanken has not been involved in the process of digitizing the newspapers, and currently only makes the material available, without any post-processing apart from linguistic annotation. The material was OCR processed by the commercial ABBYY Finereader OCR software, using the models that were pre-trained by the software. These models, originally trained on Swedish material seven years ago, were not trained to capture specific features of the newspapers such as layout and typography.

More recently, parts of the material have been re-processed by the Royal Swedish library using an improved OCR process to create data of a higher quality [9]. Additional material has also been added, and the new KubHist corpus, which contains more than 5.5 billion tokens, and will be added to Korp shortly.

Many of the modern methods for detecting language changes rely on neural embedding methods that require large amounts of text (i.e., tokens that are automatically recognized). However, even 5.5 billion tokens is a small amount, considering that it is spread over roughly 200 years. The available tokens per year range from 800 tokens in 1647 to 156 million tokens in 1892, see Figure 1 for an overview of the distribution of tokens over time. In addition to the low amount of data for most years, the quality of the digital text affect the results.

We know that the KubHist dataset, spanning 1645 – 1926, contains a large number of OCR errors, ranging from one misrecognized character in a word – including space, which splits a word into several tokens or joins several words into one token – to gibberish which is not understandable without consulting the image (and sometimes even the image is not enough). Because of this, the number of tokens is just an initial estimate, which will vary during the processing of the material. The texts have been automatically annotated with parts-of-speech and links to lexicon entries. The quality of these annotations varies greatly, due to the annotation tools not being adapted to the historical language variety, bad OCR quality, and spelling variation.

The aim of this paper is to get an initial estimate of the quality of the texts. The material has currently not been processed, apart from the initial OCR process, and some basic linguistic annotation, such as parts-of-speech and linking text words to lexicon entries. No other OCR program, such as for instance Transkribus, has been tested and thus no attempt to compare between the results of other OCR programs, as proposed by [18], has been made. One reason for that is the tremendous amount of data and the time it takes to process the material. Since we do not have a gold standard for this material, we would need to apply unsupervised learning methods to identify and correct errors. The large amount of data prohibits manual approaches, unless we only want to look at a small part (which may or may not be representative). We therefore approach the texts by exploring the lexicon-coverage, using lexical resources available for both historical and modern Swedish. This will help us in our future efforts of improving the OCR through post-processing and by handling spelling variation.

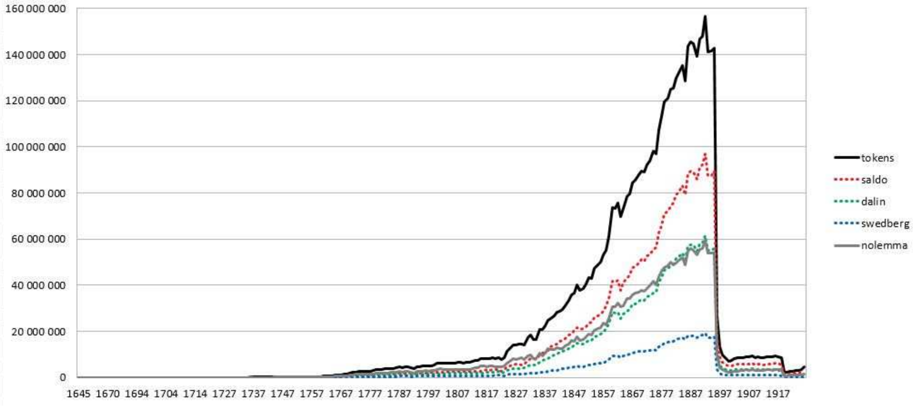

Fig. 1. The number of tokens and annotated in-vocabulary items from the different dictionaries, per year.

The end goal is to use the data for automatically detecting semantic change, after correcting OCR errors and normalizing spelling variation and change. The improvements also have value in themselves, making these diachronic texts better suitable both for manual search and for automatic processing, within, for example, the digital humanities.

2 Basic Annotation

The newspapers in KubHist have been digitized by taking high-quality images of the pages and then applying OCR software, see Section 3 for details. The resulting XML-files have then been processed by the Sparv annotation tools (`https://spraakbanken.gu.se/sparv/`). Sparv produces a range of linguistic annotations, from tokenization to part-of-speech tagging and named entity recognition. Common for all tools in Sparv is that they were developed for modern Swedish and not the language from the KubHist time period. However, the system has a number of historical lexical resources available, and we use them to link tokens in the text to lexicon entries. The dictionaries relevant for the texts at hand are Saldo [4] over contemporary Swedish, Dalin (1853/1855), over 19th century Swedish, and Swedberg [10], over 18th century Swedish [3]. In addition, a morphology is needed (basically a full form lexicon) to match inflected forms [2]. The morphologies for the different dictionaries are at varying level of development.

Figure 1 shows the number of in-vocabulary items from the different lexica, as well as the number of tokens without any match in the lexicon. The sharp drop in available texts from around 1900 is due to copyright restrictions. Although there will be errors in the lexicon links stemming from, e.g., OCR errors, we will focus on the out-of-vocabulary tokens to identify potential OCR errors. However, an out-of-vocabulary item may also stem from words missing in the lexicon

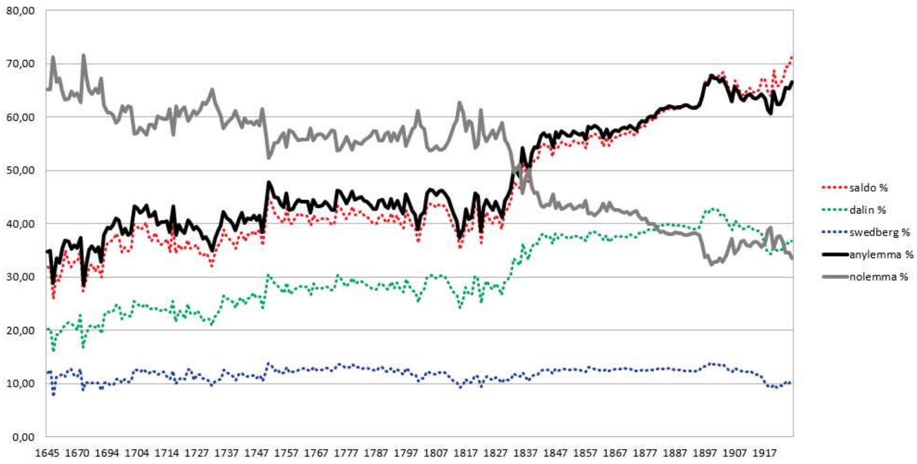

Fig. 2. The number of annotated in-vocabulary items in percent, from the different dictionaries, per year.

(which is especially obvious in the case of names), as well as an underdeveloped morphology. In addition, we will explore cases where we have a Dalin or Swedberg lexicon match, but no Saldo lexicon match, since we would like to increase the diachronic links between the lexica.

In Figure 2, we see the coverage of the different lexica over time, represented as percent of the number of tokens for each year. Swedberg has a fairly stable rate of in-vocabulary items over time, around 12 percent. The generally low percentage comes from the fact that this resource has the smallest morphology attached to it. Saldo and Dalin follow each other over time, although Saldo has by far the largest morphology, which is seen in the gap between the two. However, after 1900, Dalin drops in coverage, most likely because of the spelling reform of 1906, after which the Dalin spelling no longer matches the spelling in the newspapers.

We also see that the links to Saldo and Dalin entries increase from the 1830s, resulting in more tokens having a lexicon match in at least one resource compared to tokens without a match. We assume that e.g. paper quality and fonts (such as a lower number of blackletter articles) decrease the number of OCR errors. The exact reasons still need to be confirmed through closer inspection.

3 OCR errors

There may be several reasons for the low quality of the digital texts after automatic OCR processing. The quality of the paper or print may be low, resulting in smudgy images for the OCR software to work with. Various font sizes, uneven text lines, and a varying amount of columns cause difficulties for the OCR software to analyze the structure of the image. As a result, e.g., points and accents are mistaken for noise, graphic or geometric symbols are interpreted as text, and characters are interpreted as symbols. The mixture of font types, most notably blackletter and roman typefaces, requires that OCR software is properly trained on old fonts and languages. These types of errors are best evaluated with a gold standard [11, 7]. When no gold standard is available, as in our case, it is common to perform qualitative analysis based on different features of the data [6]. Our evaluation of OCR errors presented here are based on the linguistic annotations.

The KubHist material has been processed by an OCR-module that combines the output results from both ABBYY Finereader and Tesseract. The module has been developed in cooperation with the Norwegian software company Zissor, and has been proven to achieve high OCR accuracy, but unfortunately does not process our material with sufficient quality. This becomes apparent when we study the rate of in-vocabulary items. Although there are several reasons for the annotation tools not being able to match tokens in the texts to lexicon entries, a low rate of in-vocabulary items may point to, e.g., blackletter articles. When we explore the 349.608 OCR processed newspaper editions, we find that 27% have an in-vocabulary rate of 50% or lower. Only 3% have an in-vocabulary rate of above 80%. (It should be noted that this does not say anything about the quality of the links to the lexicon entries, it just states that a number of tokens were identified as forms of words in the lexicon by the annotation tools.)

In an initial experiment we examined the top 500 most common out-of-vocabulary tokens, categorizing them according to seven attributes. We found that around 75% of the tokens are numbers or punctuation, which we do not expect to find in the lexicon. Less than 3% contained OCR errors (we would not expect many to show up among the most frequent words), and under 4% required some kind of processing as they were spelling variants which the tools did not recognize. However, as these top 500 words were explored as word types, in isolation, around 10% could not be categorized out of context. Overall, although numbers and punctuation may contain a large amount of OCR errors, which are difficult to detect using the lexicon, this shows that we should not be aiming for 100% lexicon coverage, but that the desired upper bound is much lower (unless e.g. numbers and punctuation are included in the lexicon).

For comparison, we examined two other digitized historical texts that have been manually transcribed and processed by our annotation tools, where we can assume that there are no OCR errors. One contains law text from 1734, the other contains judicial protocols. For both of these texts, the Dalin and Swedberg dictionaries (but currently not Saldo) have been used for matching lexicon entries. We also compared to modern news text, Göteborgsposten of 2013, as well as to texts with more variation, such as the 2017 Bloggmix (various Swedish blogs). For these modern corpora, Saldo has been used for matching lexicon entries. From the results in Table 1 we find that a reasonable upper bound for in-vocabulary items for modern Swedish, using a lexical resource like Saldo, is closer to 80%. For historical texts, the variation is larger, and the upper bound is quite a bit lower than for modern texts.

Table 1. The distribution of in-vocabulary (IV) and out-of-vocabulary (OOV) tokens for different corpora.

Corpus	token count	IV	OOV
Law text	100.000	65%	35%
Judicial protocols	120.000	40%	60%
Göteborgsposten (news)	16.870.000	80%	20%
Bloggmix	1.670.000	78%	22%

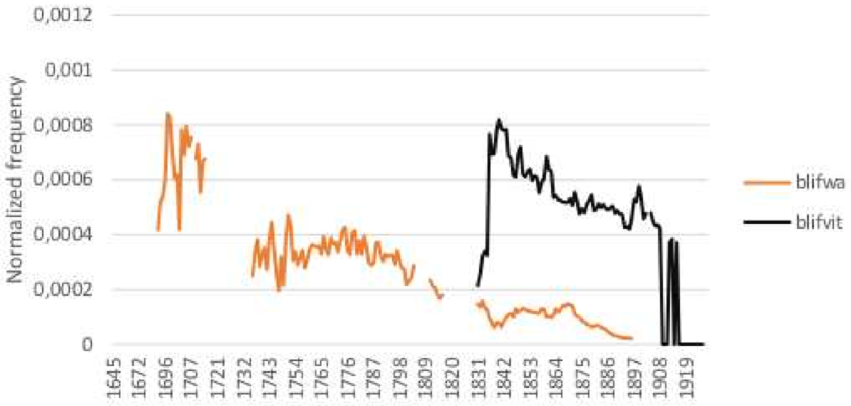

Fig. 3. Two spelling variants with some overlap, but mostly complementary use.

4 Spelling Variation

We split the KubHist dataset into 50-year bins and explore the most frequent words that were out-of-vocabulary. Our hypothesis is that words that appear among the most frequent words in many bins are unlikely to be OCR errors. Instead we expect words that are OCR errors to be less frequent, unless they are consistent with font errors. A word like *massor* ('many', 'masses') could translate to one of "niassor", "iiiessor" etc, and its frequency should be distributed over multiple possible errors, rather than concentrated to one form.

When looking at the most frequent out-of-vocabulary words in all bins, we find that these are different types of punctuation (!"' ()*,-.:;?/») and numbers (*1-9*, *14*), as well as single letters (*M*, *a*, *m*, *n*, *r*, *t*), and a few words (*ägt*, *nied*). Among those that were frequent in only one bin we find names (*Londén*, *Maji*, *Borgholm*), uncommon or short-lived abbreviations (*k.*, *K.*), and possible OCR errors (*ocb/oeh* → *och*, *näget* → *något*). Important to remember is that the first 50-year bin has very little text from only a few sources (only a couple of available newspapers) meaning that a name like *Borgholm* could be e.g. the name of a journalist and not universally important. This remains to be investigated.

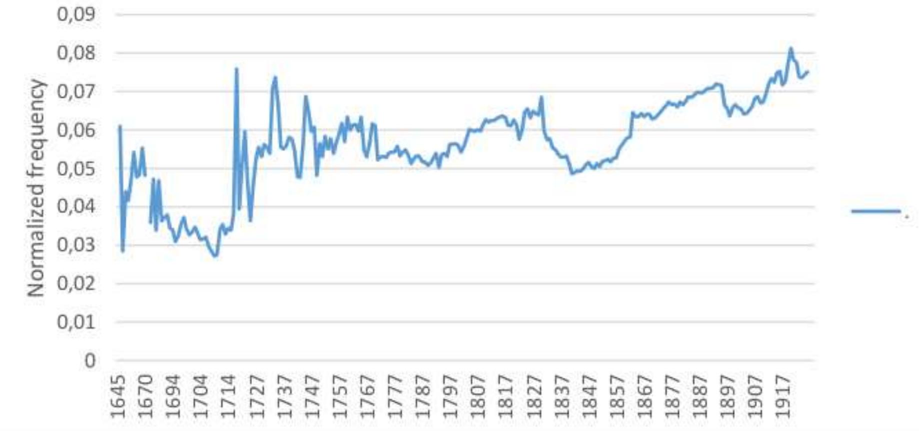

Fig. 4. The frequency of '.' (period) over time. While not present in a dictionary, it should not be categorized as spelling variation or OCR error at a general level (although there definitely are instances of OCR errors).

Among the words that are out-of-vocbulary and appear in three bins, i.e., three 50-year periods, we have words that are common spelling variants, such as *äfven*, *blifvit*, *öfwer*, *blifwa*, *hafwa*, *warit*, *hvilka*, and *hwilka* (where modern Swedish has *v* for *fv*, *fw*, *w*, *hv*, and *hw*). Interestingly, some of these seem to hand over to each other, like in the case of *blifwa* and *blifvit* in Figure 3. The latter has a normalized form *blifva* that is present in Dalin, but due to an incomplete morphological description, the past tense of *blifvit* is not captured. Their frequency seems complementary. Observe that years without a frequency corresponds to an absolute frequency of below 50 occurrences. In the case of a frequently occurring character without an entry in the dictionary, '.' (period) in Figure 4, we see that the frequency is much more consistent across years.

In future work, we intend to use these characteristics to attempt to automatically categorize out-of-vocabulary words as either OCR errors (which we expect to have a low, but possibly consistent frequency), spelling variants (with a higher frequency focused around a specific period), or common characters not included in dictionaries (punctuation, numbers, etc).

5 Conclusions

In this paper we present the first overview of the KubHist corpus, containing-more than 300 thousand Swedish historical newspaper editions. The corpus was recently digitized and OCR-processed by the Royal Swedish library. We explore the texts and some simple methods to identify OCR errors stemming from the digitization process, with the help of historical and modern dictionaries.

We found that a large part of the tokens are numbers or punctuation, for which separating correct tokens from OCR errors is not easily done with a lexicon approach. The results do, however, indicate that a fairly large amount of the errors could be identified with simple processing, such as temporal profiling. Correcting identified errors will increase the amount of data that can be used for automatically detecting semantic change, as well as other research within the digital humanities.

Acknowledgements

The research presented here has been supported by Språkbanken and Swe-Clarin, a Swedish consortium in Common Language Resources and Technology Infrastructure (CLARIN) Swedish CLARIN (grant agreement 821-2013-2003).

References

1. Basile, P., Caputo, A., Luisi, R., Semeraro, G.: Diachronic analysis of the Italian language exploiting Google Ngram. In: Italian Conf. on Comp. Linguistics (2016)
2. Borin, L., Forsberg, M.: Something old, something new: A computational morphological description of Old Swedish. In: LREC Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008). pp. 9–16. Marrakech (2008)
3. Borin, L., Forsberg, M.: A diachronic computational lexical resource for 800 years of Swedish. In: Sporleder, C., van den Bosch, A., Zervanou, K. (eds.) Language technology for cultural heritage. pp. 41–61. Springer, Berlin (2011)
4. Borin, L., Forsberg, M., Lönngren, L.: SALDO: a touch of yin to WordNet's yang. Language Resources and Evaluation **47**(4), 1191–1211 (2013)
5. Borin, L., Forsberg, M., Roxendal, J.: Korp the corpus infrastructure of Språkbanken. In: Proceedings of LREC 2012. ELRA, Istanbul (2012)
6. Cassidy, S.: Publishing the Trove newspaper corpus. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)
7. Clematide, S., Furrer, L., Volk, M.: Crowdsourcing an OCR gold standard for a German and French heritage corpus. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)
8. Dalin, A.F.: Ordbok öfver svenska språket, vol. I–II. Stockholm, Sweden (1853/1855)
9. Dannélls, D., Johansson, T., Björk, L.: Evaluation and refinement of an enhanced OCR process for mass digitisation. In: Digital Humanities in the Nordic countries. University of Copenhagen, Copenhagen (2019)
10. Holm, L. (ed.): Jesper Swedberg: Swensk Ordabok. Skara stiftshistoriska sällskaps skriftserie, Stiftelsen för utgivande av Skaramissalet, Skara (2009)
11. Kettunen, K., Honkela, T., Linden, K., Kauppinen, P., Pääkkönen, T., Kervinen, J.: Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods. In: IFLA World Library and Information Congress Proceedings: 80th IFLA General Conference and Assembly (2014)
12. Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S.: Statistically significant detection of linguistic change. In: WWW (2015)

13. Lau, J.H., Cook, P., McCarthy, D., Newman, D., Baldwin, T.: Word sense induction for novel sense detection. In: EACL. pp. 591–601. Association for Computational Linguistics (2012)
14. Michel, J.B., Shen, Y.K., Presser Aiden, A., Veres, A., Gray, M.K., The Google Books Team, Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Lieberman Aiden, E.: Quantitative analysis of culture using millions of digitized books. Science **331**(6014), 176–182 (2011)
15. Tahmasebi, N., Risse, T.: Finding individual word sense changes and their delay in appearance. In: RANLP (2017)
16. Vejdemo, S.: Triangulating Perspectives on Lexical Replacement: From Predictive Statistical Models to Descriptive Color Linguistics. Ph.D. thesis, Stockholm University (2017)
17. Viberg, Å.: Studier i kontrastiv lexikologi: perceptionsverb. Stockholms Universitet, Inst. för lingvistik (1980)
18. Volk, M., Furrer, L., Sennrich, R.: Strategies for reducing and correcting OCR errors. In: Language Technology for Cultural Heritage. pp. 3–22. Springer, Berlin (2011)