Effect of data quality on the automated identification of register features in Eighteenth Century Collections Online

Aatu Liimatta
University of Helsinki
aatu.liimatta@helsinki.fi

Abstract

Many large-scale investigations of textual data are based on the automated identification of various linguistic features. However, if the textual data is of lower quality, automated identification of linguistic features, particularly more complex ones, can be severely hampered.

Data quality problems are particularly prominent with large datasets of historical text which have been made machine-readable using optical character recognition (OCR) technology, but it is unclear how much the identification of individual linguistic features is affected by the dirty OCR, and how features of varying complexity are influenced differently.

In this paper, I analyze the effect of OCR quality on the automated identification of the set of linguistic features commonly used for multi-dimensional register analysis (MDA) by comparing their observed frequencies in the OCR-processed *Eighteenth Century Collections Online* (ECCO) and a clean baseline (ECCO-TCP). The results show that the identification of most features is disturbed more as the OCR quality decreases, but different features start degrading at different OCR quality levels and do so at different rates.

1 Introduction

Large-scale textual datasets have become increasingly common in computational linguistics and in various subfields of digital humanities. Naturally, such datasets cannot easily compare to the high quality of traditional (but much smaller) linguistic corpora, which have been carefully curated for balance and manually edited to ensure the accuracy of the textual data to as large a degree as feasible. Instead, it is the expectation that when analyzed in aggregate, the large amount of data can smooth over many of the flaws which would prove very difficult to work around with smaller datasets or when focusing the analysis on individual texts or individual instances of items. Even then, the overall lower quality of the data causes issues for many linguistic and other digital humanities analyses, such as in the case of, for example, the analysis of social media data (e.g. Eisenstein, 2013).

However, a major source of large-scale low-quality textual humanities data, particularly in fields such as historical linguistics and computational history, are texts which have been turned from scans of physical documents into machine-readable format using optical character recognition (OCR) technology. For instance, the OCR quality of *Eighteenth Century Collections Online* (ECCO)[1], a collection of over 200,000 mostly English-language works published in the United Kingdom during the 18th century (see Tolonen et al., 2021) and a central resource for DH scholars (Gregg, 2020), is extremely variable. A main contributing factor of the often low OCR quality for ECCO is the OCR being run on bitonal scans of microfilms with OCR algorithms which have not been fine-tuned for eighteenth-century typefaces or trained to recognize e.g. the long *s* character ⟨ſ⟩.

Earlier studies on the effects of OCR errors in ECCO (e.g. Hill and Hengchen, 2019) are largely focused on individual tokens, characters, and n-grams. In contrast, the present study focuses on the effect of the OCR errors in ECCO on the automated identification of a set of more complex linguistic features commonly used for the multi-dimensional method of register analysis (MDA) (see e.g. Biber, 1988; Biber and Conrad, 2009).

[1] https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 45–51
December 1-3, 2023 ©2023 Association for Computational Linguistics

2 Background

2.1 OCR and ECCO

The difficulties caused by dirty OCR have been long recognized in the literature (e.g. Lopresti, 2009; Traub et al., 2015; Vitman et al., 2022). When it comes to the ECCO dataset, Hill and Hengchen (2019) compare the ECCO-TCP[2], a manually keyed version of a subset of the documents included in ECCO, to the same set of documents from the regular OCR-processed ECCO (henceforth: ECCO-OCR) both on the basis of token and type similarity and in a number of bag-of-words approaches used in digital humanities, such as topic modeling and methods of authorship attribution. They find that, for example, the mean OCR precision in their dataset is 0.744, meaning that on the average page, 74% of the tokens are correct, whereas the recall is 0.814, meaning that 81% of the tokens are included in the OCR version.

2.2 Multi-dimensional analysis

MDA, originally developed by Biber (1988), is a corpus-linguistic approach which extracts functional dimensions from a textual dataset; the dimensions describe variation in the communicative purposes and situational concerns between the texts within the dataset (see e.g. Biber, 1988; Biber and Conrad, 2009), each dimension comprising a gradient between two poles with opposing functions.

The central idea behind the MDA methodology is that linguistic features which are better-suited to the function and situational concerns of a text are more likely to be used in the text. Consequently, commonly co-occurring linguistic features can be assumed to share an underlying set of functions. MDA uses statistical methods such as factor analysis on a set of texts to extract co-occurring (and complementary) groups of features which are then interpreted in terms of their function, forming dimensions of register variation.

To give a basic example, *past tense* verb forms and *third-person pronouns* are naturally more common in narrative contexts than in non-narrative contexts. One of the central findings of Biber (1988) is the gradient between "involved" and "informational" production, with the informational pole characterized by features such as *nouns*, *prepositions*, and *attributive adjectives*, and the complementary involved pole by features such as *private verbs*[3], *THAT deletion*, and *contractions*.[4]

While in principle different sets of linguistic features can be and have been used for MDA analyses, it is common to build a MDA feature set on the core set of linguistic features originally compiled by Biber (1988) through an extensive survey of previous linguistic literature.

2.3 Multi-dimensional analysis and ECCO

Many of the features included in the MDA core set of features are much more specific and more complex than those analyzed by Hill and Hengchen (2019). It is a statistically reasonable assumption that a random OCR error is more likely to occur in a longer multi-word construction than in an individual token. Consequently, it could be expected that dirty OCR would make it more difficult to identify such complex features in the data, and that therefore any analysis making use of such features would be severely disturbed or completely prevented if the set of texts being analyzed contains many OCR errors.

In order to test this assumption, Liimatta et al. (2023) evaluate the effects of dirty OCR on the MDA methodology by comparing the results of the analysis run on ECCO-TCP and separately on the parallel set of documents from ECCO-OCR. Perhaps surprisingly, the MDA dimensions Liimatta et al. (2023) acquire from ECCO-TCP and ECCO-OCR turn out to be very similar, which suggests that even if not every instance of each linguistic feature used in the analysis is identified properly in the ECCO-OCR data, enough instances of most features can still be identified for meaningful co-occurrence patterns to be preserved to a degree.

However, it still remains the case that dirty OCR renders many instances of any features of interest unrecognizable by automated processing methods, and as such, there are linguistic features which are better or worse suited for MDA analysis of dirty OCR datasets such as ECCO-OCR and for other analyses using similar approaches. The aim of the present paper is to explore the core set of features used for MDA and see how each of these features is individually affected by the OCR process, to shed light on which kinds of linguistic features can most robustly be used for MDA and other similar analyses, but also more generally to understand

[2] https://textcreationpartnership.org/tcp-texts/ecco-tcp-eighteenth-century-collections-online/

[3] e.g. *think*, *assume*, or *feel*

[4] For the full list of features included in Biber (1988) and their descriptions, see Biber (1988, Appendix II).

the influence of decreasing OCR quality on the integrity of various linguistic structures.

3 Materials and methods

3.1 Data

All of the textual data used in the present analysis is based on ECCO. The *ECCO Text Creation Partnership* (ECCO-TCP) dataset constitutes the clean baseline for the analysis. ECCO-TCP is a manually keyed version of a small subset of the full ECCO dataset. Thanks to the careful editing process, ECCO-TCP, while not perfect, is close in quality to other hand-edited historical datasets in terms of its transcription accuracy, and as such can be considered a clean standard for the included texts (cf. Hill and Hengchen, 2019).

Additionally, in order to be able to estimate the degradation in feature identification caused by dirty OCR when compared to the clean versions of the texts, I have created as a second dataset a parallel subset of the regular OCR-processed ECCO dataset (ECCO-OCR) whose texts match the texts included in ECCO-TCP. Finally, both datasets were tagged for part of speech[5] using spaCy[6].

The OCR quality estimate is based on the confidence levels reported by the OCR engine which Gale, the publisher of ECCO, used to process the texts. The original OCR confidence was calculated on a per-page basis; for the present analysis, the mean of the OCR quality estimate for all of the pages of a work was used as the overall OCR quality of the work.

3.2 Methods

Both datasets were processed with the same feature identification pipeline, which identified the linguistic features of interest using automated means and counted their occurrences in each of the texts in both datasets. The identification of the features was performed using algorithms mainly based on those provided by Biber (1988)[7]. For instance, the algorithm for the feature *WH-clauses* is given by Biber (1988) as follows:

```
PUB/PRV/SUA + WHP/WHO + xxx
(where xxx is NOT = AUX)
```

This means that *WH-clauses* are identified as starting with a word belonging to the classes of public, private, or suasive verbs (e.g. *say*, *believe*, *agree*), followed by a WH pronoun (i.e. *who*, *whose*, *whom*, *which*) or other WH word (e.g. *what*, *when*, *whether*, *why*, *wherever*), followed by a word which is *not* an auxiliary, defined by Biber (1988) to be either a modal verb, or any of the the verbs *to do*, *to have*, or *to be*.[8]

However, the present study uses a different part-of-speech tagset and tokenization scheme than the original study by Biber (1988), which sometimes requires changes or allows for improvements to the algorithms, and consequently the algorithms for each individual feature were not always followed exactly. Furthermore, a handful of the features were excluded from the analysis because their algorithms require manual checking of the results.

After normalizing the observed feature counts to the observed character count[9] of the text, I calculated the proportion of the frequency observed in the ECCO-TCP version of each of the texts which was observed in the ECCO-OCR version of the same text. In other words, because OCR errors will in many cases prevent the implemented algorithms from correctly identifying the features, e.g. because the word form has character errors or because the part of speech has been misidentified, I calculated by how much the observed frequency of each of the features changed from the clean version of the text to the OCR version of the text. This

[5] Automated part-of-speech (POS) tagging is not perfect, and may itself present some problems for the downstream task of automated identification of linguistic features. Furthermore, OCR errors will lower POS tagging quality for the ECCO-OCR dataset. However, POS tagging is a necessary step for the identification of the linguistic features analyzed in the present study, and as such MDA studies even on perfectly clean datasets typically need to make use of imperfect POS tagging. As the aim of the present study is to gauge the effect of OCR quality on the identification of linguistic features, it is reasonable to use a realistic POS tagging as a baseline, and to include the POS tagging quality degradation in the OCR dataset as part of the overall degradation of feature identification caused by dirty OCR.

[6] https://spacy.io

[7] While the MDA methodology has seen extensive use over the years, Biber (1988) still provides the most comprehensive description of the exact patterns by which the individual features can be identified.

[8] For the full list of the features included in the present study, see Appendix A. For the full list of original algorithms and descriptions of the features, see Biber (1988, Appendix II).

[9] While token count or word count would be more typical bases for normalization, I have chosen to use the character count of the text as the normalization basis for problematic OCR data. The dirty OCR often includes many erroneous spaces between strings of characters, which tends to inflate the number of tokens as created by a typical tokenizer. Consequently, preliminary analyses show that the character count of the OCR text, while still wrong, is likely to be proportionally closer to the true character count of the text than the token count is to its true token count.

change was calculated for each text as

$$\frac{f_{ocr}}{f_{tcp}} - 1$$

where f is the normalized frequency of the feature as observed in either the OCR or TCP version of the text; the offset -1 is to have a value of 0 represent no change from TCP to OCR.

4 Results

Figure 1 shows the results of the analysis for each of the features as a function of the OCR quality of the text. Every facet represents a different linguistic feature, with the OCR quality estimate for the text on the vertical axis, higher OCR quality at the top and OCR quality decreasing towards the bottom. The horizontal axis represents the proportional change in the observed frequency of the feature from ECCO-TCP to ECCO-OCR. Because there is a lot of variation in the data, a smoothing trend line[10] was also included for each feature.

In other words, points at 0.0 on the horizontal axis in Figure 1, i.e. on the dark vertical line, represent texts with the exact same frequency in both ECCO-TCP and ECCO-OCR. Such texts are the ideal case, since they have no change in the observed frequency from the clean version to the OCR version of the text, suggesting that all instances of the feature have been correctly identified. However, in practice, as expected, most texts deviate from the ideal. Points to the left of the middle line represent texts which have a lower frequency of the feature in ECCO-OCR than in ECCO-TCP. For instance, in a text at the -0.5 mark, the OCR version has only half the observed frequency of the feature as compared to the clean version. Similarly, the points to the right of the middle line indicate texts with a higher frequency of the feature in ECCO-OCR than ECCO-TCP, with the 0.5 mark meaning a 50% increase in the observed frequency.

Based on Figure 1, it is clear that as the OCR quality decreases, the observed frequency of identification typically also decreases, as evidenced by the trend line tending down and to the left. But different features are also clearly affected differently by the decrease in OCR quality. For some features, the fall towards the left is very direct and rapid, beginning very close to the highest OCR quality texts, meaning that their identification is instantly

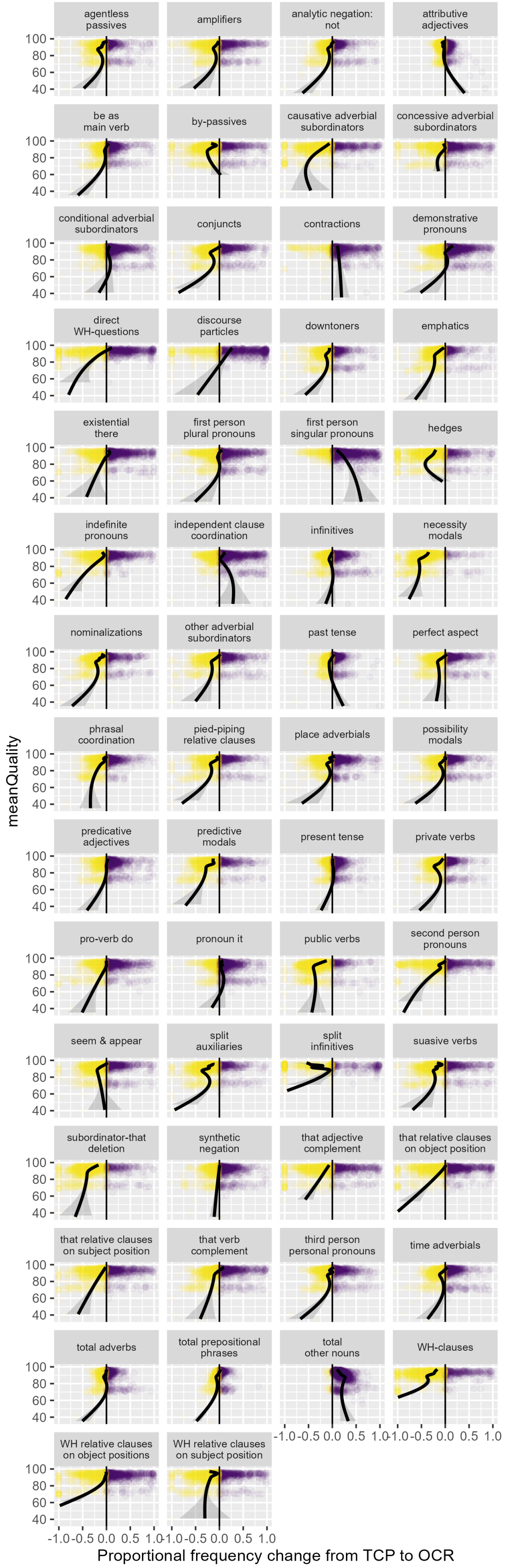

Figure 1: Proportion of ECCO-OCR rate of occurrence of the ECCO-TCP rate of occurrence by OCR quality

[10] ggplot2 (Wickham, 2016): `geom_smooth(method = "gam", formula = y ~ s(x, bs = "cs"))`

affected when the OCR quality decreases. These include features such as *direct WH-questions*[11] and *that relative clauses on object position*[12].

Features which appear to be particularly adversely affected by dirty OCR include *WH-clauses*[13], which have on average lower observed frequencies even in the OCR texts with the best OCR quality, and the frequency drops very rapidly with the OCR quality. Possibly because of OCR problems with the long *s*, *necessity modals*[14] also start at a lower observed frequency even at highest OCR quality, and keep decreasing in frequency with decreasing OCR quality, if somewhat less quickly than WH-clauses do.

In contrast to features of the above type, which start decreasing as soon as the OCR quality decreases, there are a number of features which stay relatively stable for a larger portion of the OCR quality range and only show larger changes in their average observed frequency when the OCR quality drops too much, possibly because they tend to be relatively simple to identify. These features include *present tense* verb forms, which only show a drop below 60-70% OCR quality, *attributive adjectives*[15], which show an increase below a similar OCR quality level, and *predicative adjectives*[16], which show a drop below 70-80% OCR quality. There is also an even larger group of features which do show a small difference from the clean version at higher OCR quality levels but for which this difference is only relatively minor, including e.g. the *analytic negation not*, *first person plural pronouns*[17] and *total adverbs*[18], all of which only show a larger change in their observed frequency below about 70% OCR quality.

Another, rather curious group of features are those whose identified frequencies increase instead of decrease in the OCR dataset. Typically, this increase can be attributed to the OCR process and other processing of the dataset, particularly the tokenization and the part-of-speech tagging. For instance, the *total other nouns*[19] category shows higher frequencies in ECCO-OCR compared to the clean baseline throughout the OCR quality range. This is because the imperfect OCR process creates many strings of characters which the part-of-speech tagger cannot recognize, and most taggers have a tendency of tagging such tokens as nouns, particularly if there are too many unrecognizable tokens in succession to infer a different part of speech from the surrounding structure.

Similarly, *first person singular pronouns* appear to be identified at a close to correct rate in the OCR texts close to the top OCR quality, but there is an increasing number of misidentifications as the OCR quality decreases. This is largely because as the OCR quality degrades, the OCR process produces more and more random *I* characters, which then get misidentified as the first person pronoun *I*.

Finally, there are a few features which appear to be barely affected by the decrease in OCR quality. These are likely relatively simple to identify features which also happen to consist of characters which tend to have been recognized more reliably than average in the OCR process. These features include most prominently *synthetic negation*[20], but features such as *pronoun it* and *infinitives* could also be considered to not be affected very much at all by changes in OCR quality.

5 Conclusion

The results show that, as expected, lower OCR quality leads to lower reliability of identification for most of the features analyzed. However, the effect of OCR on all of the features is not the same, with features which are simpler to identify and less likely to invite OCR errors appearing generally more resistant to lower OCR quality, while features whose identification involves multiple consecutive items from specific lists appear generally more at risk. It is not possible based on this study alone to recommend any single ECCO OCR quality level which could be considered “good enough” for most analyses, as what is good enough depends on the features one is interested in. On the other hand, MDA studies consistently produce “compatible” results regardless of the exact set of features analyzed (McEnery and Hardie, 2012), suggesting that for MDA, focusing on a smaller set of more resilient features may be enough to get better results even from lower-quality textual data.

[11] e.g. *What is that?*

[12] e.g. *the place that they mentioned*

[13] e.g. *I didn't know what it was*

[14] *must*, *should*, and *ought*

[15] e.g. *a blue house*

[16] e.g. *the house is blue*

[17] e.g. *we*

[18] e.g. *quickly*

[19] Nouns not counted as nominalizations.

[20] e.g. *no man*

Acknowledgements

The research was supported by the Academy of Finland under the project *Rise of commercial society and eighteenth-century publishing* (grant number 333717).

References

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge.

Douglas Biber and Susan Conrad. 2009. *Register, genre, and style*. Cambridge University Press, Cambridge.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369.

Stephen H. Gregg. 2020. *Old Books and Digital Publishing: Eighteenth-Century Collections Online*. Cambridge University Press.

Mark J. Hill and Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Aatu Liimatta, Yann Ryan, Tanja Säily, and Mikko Tolonen. 2023. Results from rough data? The large-scale study of early modern historiography with multi-dimensional register analysis. *Digital Humanities in the Nordic and Baltic Countries Publications*, 5(1):297–312.

Daniel Lopresti. 2009. Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(3):141–151.

Tony McEnery and Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice*. Cambridge University Press, Cambridge, New York.

Mikko Tolonen, Eetu Mäkelä, Ali Ijaz, and Leo Lahti. 2021. Corpus linguistics and Eighteenth Century Collections Online (ECCO). *Research in Corpus Linguistics*, 9(1):19–34.

Myriam C. Traub, Jacco van Ossenbruggen, and Lynda Hardman. 2015. Impact Analysis of OCR Quality on Research Tasks in Digital Archives. In *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 252–263, Cham. Springer International Publishing.

Oxana Vitman, Yevhen Kostiuk, Paul Plachinda, Alisa Zhila, Grigori Sidorov, and Alexander Gelbukh. 2022. Evaluating the Impact of OCR Quality on Short Texts Classification Task. In *Advances in Computational Intelligence*, Lecture Notes in Computer Science, pages 163–177, Cham. Springer Nature Switzerland.

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

A List of features included

- agentless passives
- amplifiers
- analytic negation: not
- attributive adjectives
- be as main verb
- by-passives
- causative adverbial subordinators: because
- concessive adverbial subordinators: although, though
- conditional adverbial subordinators: if, unless
- conjuncts
- contractions
- demonstrative pronouns
- direct WH-questions
- discourse particles
- downtoners
- emphatics
- existential there
- first person plural pronouns
- first person singular pronouns
- hedges
- indefinite pronouns
- independent clause coordination
- infinitives
- necessity modals
- nominalizations
- other adverbial subordinators
- past tense
- perfect aspect
- phrasal coordination
- pied-piping relative clauses
- place adverbials
- possibility modals
- predicative adjectives
- predictive modals
- present tense
- private verbs
- pro-verb do
- pronoun it
- public verbs
- second person pronouns
- seem & appear
- split auxiliaries
- split infinitives
- suasive verbs

subordinator-that deletion
synthetic negation
that adjective complement
that relative clauses on object position
that relative clauses on subject position
that verb complement
third person personal pronouns
time adverbials
total adverbs
total other nouns
total prepositional phrases
WH relative clauses on object positions
WH relative clauses on subject position
WH-clauses