Journal of Data Mining and Digital Humanities

EPI sciences overlay journals

OCR quality and the resilience of algorithmic identification of linguistic register features in Eighteenth Century Collections Online

Aatu Liimatta[1]

[1]University of Helsinki, Finland

Corresponding author: Aatu Liimatta , 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 in a clean baseline (ECCO-TCP). The results show that the identification of most features is disturbed more as the OCR quality decreases, but that some features can be particularly resilient against the degradation in OCR quality.

Keywords

multi-dimensional register analysis; optical character recognition quality; Eighteenth Century Collections Online

I 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].

II 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][2]. When it comes to the ECCO dataset, Hill and Hengchen [2019] compare the ECCO-TCP[3], 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

Register analysis is a field of linguistics which is focused on *registers*, functional varieties of language defined by the situation and/or purpose of the language use. Contemporary computational work on register analysis in corpus linguistics commonly builds on or is inspired by the framework of multi-dimensional register analysis (MDA), originally developed by Biber [1988]. MDA is a methodological framework for extracting functional dimensions from a textual dataset; the extracted 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

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

[2] OCR technology has continued to improve, which can reduce the number and impact of OCR errors in more recent and upcoming datasets. However, any remaining OCR errors will cause similar data quality degradation, even if its overall effect is lower. Furthermore, dirty OCR is not the only potential source of errors in textual data. As degradation of data quality for the purposes of computational analysis may also be introduced by variation in spelling conventions, typographical errors, and other similar sources, the present analysis has potential further implications for datasets ranging from historical texts to modern social media postings.

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

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*[4], *THAT deletion*, and *contractions*.[5]

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

III 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

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

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

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 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 tokenized, split into sentences, and tagged for part of speech[6] by feeding the full texts through spaCy[7].

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][8]. 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 verbs *to do*, *to have*, or *to be*).[9]

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[10]. Furthermore, a handful of the features were excluded from the analysis because their algorithms as given by Biber [1988] require manual checking of the results,

[6] 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 potentially imperfect POS tagging for the baseline TCP dataset, as would also be the case for a real-world analysis of the data, and to include the degradation of the POS tagging quality in the OCR dataset as part of the overall degradation of feature identification caused by dirty OCR.

[7] https://spacy.io

[8] 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.

[9] 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].

[10] The algorithms were originally developed for the analysis of contemporary English. For the present study, the implementation details of a number of the algorithms have been adjusted for technical reasons. However, in order to provide a baseline reference measurement of the influence of OCR errors, the algorithms have not been changed either to better accommodate eighteenth-century English or to account for common OCR errors.

which would be infeasible with a large dataset the size of ECCO: the excluded features are *gerunds*, *present participial clauses*, *past participial clauses*, *present participial WHIZ deletion relatives*, and *sentence relatives*. Furthermore, *type-token ratio* and *mean word length* were excluded because the focus of the present study is on feature frequencies. As a change from the original feature set used by Biber [1988], *first person pronouns* were split into *first person singular pronouns* and *first person plural pronouns*, which have different functions and as such behave differently from each other.

In order to compare the occurrence of features between texts which differ in length, the observed number of occurrences of each feature is normalized to some base, commonly the token count or word count of the text. However, the problem with using such measures is that dirty OCR commonly includes mistaken strings of characters which cause changes in tokenization, in particular erroneous space characters, and sometimes missed space characters. Excessive space characters cause typical tokenizers to interpret individual words in the original text as multiple tokens, and missing spaces conflate multiple words into a smaller number of tokens. Because of this discrepancy between the real number of tokens in the text and the number of tokens extracted by a tokenizer from a dirty OCR text, the normalization basis can also be wrong, and is likely to be more wrong the lower the OCR quality is. In other words, even if the count of the feature of interest as extracted from the OCR version of the text is itself correct, the base used for normalization, i.e. the token count, is likely to be at least slightly off from its real value, resulting in skew in the normalized frequencies of the items. In order to gauge the overall influence of the two effects of dirty OCR (viz. the inability to identify features correctly and the inability to calculate the normalization base accurately), I consider in the present study the combined effect of the two by comparing normalized values instead of simple counts of feature occurrences.

However, token count is not the only possible normalization base. Primarily, the observed character count of a dirty OCR version of a text could be expected to be proportionally closer to the true character count than the observed token count is to the true token count. Crucially, the character count is not affected by the insertion or removal of spaces. Based on an analysis of the effect of OCR quality on the observed token count and character count of the text, the results of which are presented in Section 4.1 below, I have opted to use the character count of the text as the normalization base.

After normalization, I calculated which proportion of the frequency observed in the ECCO-TCP version of each of the texts 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 change was calculated for each text as

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

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.

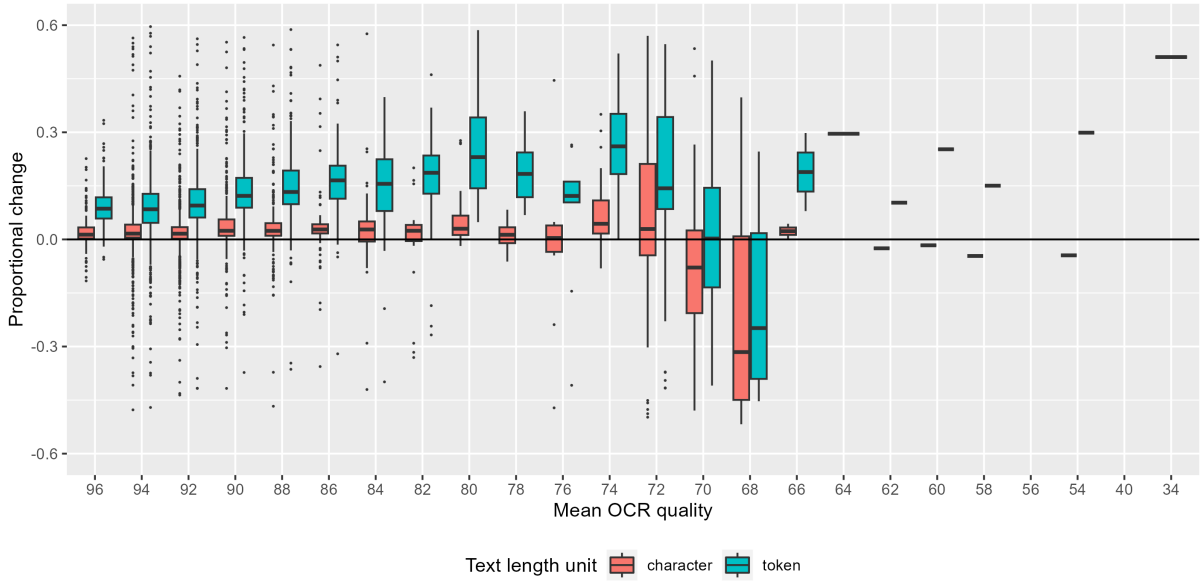

Figure 1: Proportional change in observed text length from ECCO-TCP to ECCO-OCR by OCR quality

IV RESULTS

4.1 Normalization base

Figure 1 shows the proportional change of the observed text length from the clean TCP baseline to the dirty OCR version of the text. Two measures of text length are analyzed: token count (blue) and character count (red). The texts have been combined into bins of two percentage points of OCR quality. The black horizontal line represents the ideal case of no change from TCP to OCR. The larger the deviation is from this line in either direction, the greater the discrepancy between the observed text length in the clean version and the dirty OCR version of the text.

Perhaps the most striking feature of Figure 1 is the fact that below a mean quality of roughly 75% to 80%, both measures appear to completely break down: the observed character and token counts both spread out and plummet. Since this breakdown implies a complete inability of the OCR engine to recognize the tokens and characters of the text with any level of correctness, this is also likely to be the quality range below which most analyses of ECCO data will become ineffective. In the OCR quality range above this breakdown zone, the token and character counts follow more stable trajectories.

However, based on the figure, even in this more stable region, the measured token counts are clearly higher in the OCR version when compared to the clean TCP texts. Even the texts with the highest OCR quality show considerably higher token counts, and the discrepancy only grows as the OCR quality decreases. This means that not only is the token count off by a significant amount in analyses making use of OCR texts, but the degree of error also changes with OCR quality, increasing the unreliability of the measure.

On the other hand, the character counts in Figure 1 tend to follow the zero line much more closely. This indicates that the character counts experience less proportional change from the TCP to the OCR version of the text until the breakdown zone. This result is in line with and further refines the findings by Hill and Hengchen [2019, 4], who report that in particular the total character counts of the ECCO-OCR corpora are “remarkably accurate”.

Furthermore, following the development of character counts in the stable range in Figure 1, decreasing OCR quality is not associated with increased deviation from the zero-line. This stability means that even if the character counts are off by a small amount, any normalized frequencies tend to simply be offset from the true value by similar amounts, and their comparisons within the same dataset should still work relatively well. In summary, character count appears to remain a relatively accurate measure of text length throughout the higher OCR quality range.

4.2 Feature frequencies

Figure 2 shows the proportional change in the observed normalized feature frequencies for each of the analyzed features as a function of the OCR quality of the text. In the figure, every facet represents a different linguistic feature. Within each facet, the horizontal axis represents the proportional change in the observed normalized frequency of the feature. The black vertical line represents no change from TCP to OCR. Values to the left of the line mean that the observed frequency is lower in the OCR version compared to the TCP version of the text, whereas values to the right mean a higher frequency in the OCR version.

The vertical axis within each facet represents the OCR quality estimate of the ECCO-OCR version of the texts, with the highest OCR quality at the top and the quality decreasing towards the bottom of the facet. In the figure, the texts are divided into brackets of five percentage points of OCR quality, the range of variation within each bracket represented by an individual box in the boxplot.[11]

In other words, points at 0.0 on the horizontal axis in Figure 2, 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.

A cursory visual inspection of Figure 2 already makes it clear that, as expected, the vast majority of the features are affected by the OCR quality, with only a handful of features staying close to the vertical line indicating no change from the clean TCP text to the OCR version. At the same time, it is equally clear that not every feature is affected identically by the decrease in OCR quality: while there are certain repeating patterns, different features often follow different trajectories.

Perhaps the most central description of the patterns observed in the data is the direction of the effect of decreasing OCR quality. That is to say, the features can roughly speaking be divided into three main categories: 1) decreasing frequency; 2) stable frequency; and 3) increasing frequency. Of these, features exhibiting a decreasing frequency constitute an overwhelming majority. In Figure 2, a decreasing frequency pattern can be visually identified as boxplots trending down and to the left, meaning a growing decrease in the proportion of identified instances of the feature as the OCR quality decreases. The decreasing pattern is the prototypical expected effect of decreasing OCR quality, since the lower the OCR quality becomes, the more likely it

[11] For clarity, outliers are not shown in the figure.

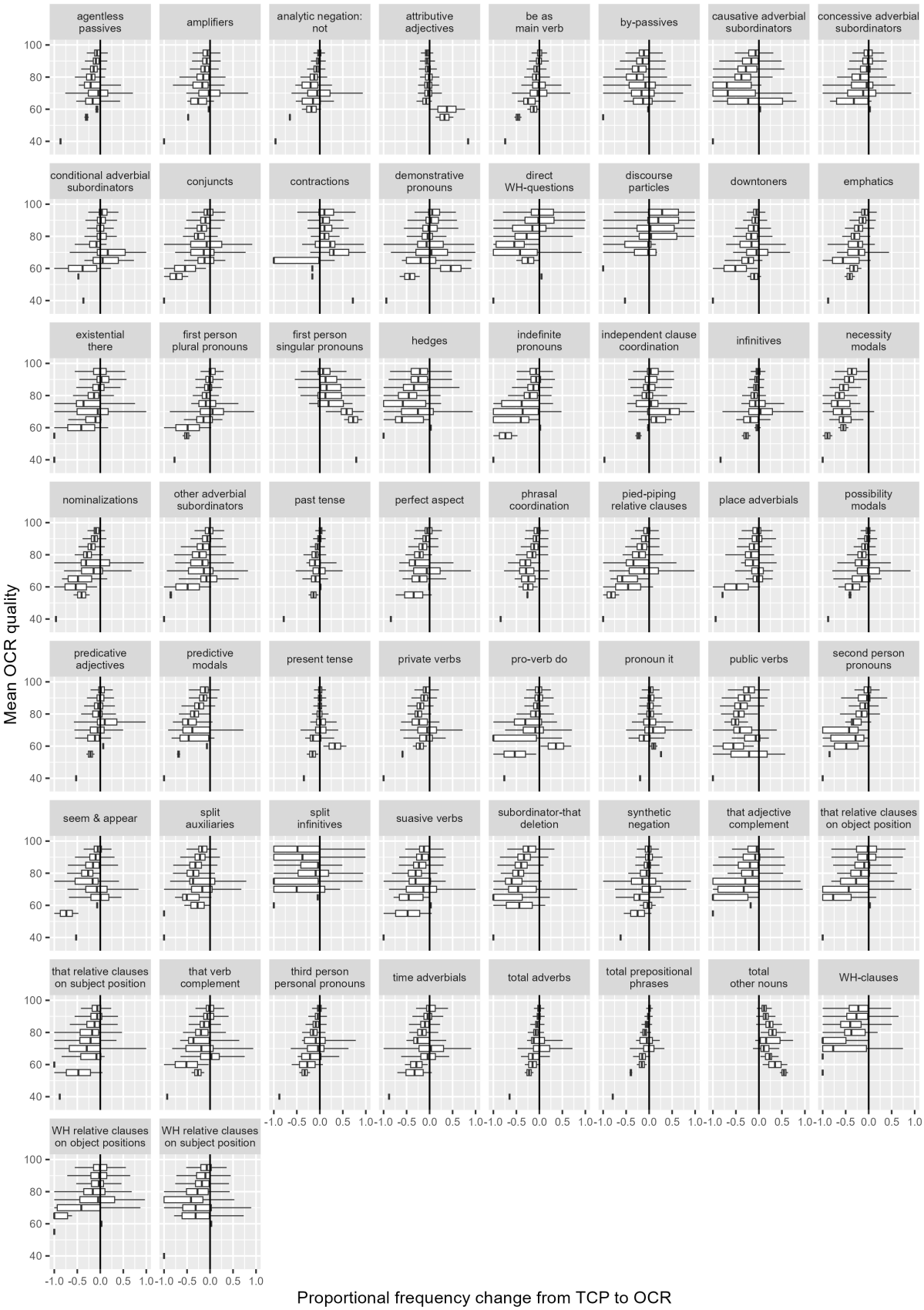

Figure 2: Proportional change from ECCO-TCP rate of occurrence to ECCO-OCR rate of occurrence by OCR quality

is that an OCR error hits an instance of the feature in a way which prevents the identification algorithms utilized from properly identifying it. As such, it is no surprise that a decreasing frequency tendency is typical for most of the features analyzed.

Features following the second pattern of stable frequency are the most promising in terms of usability in analyses across a wider spectrum of OCR quality levels. Visually, these features can be identified as following a vertical line straight down with a relatively narrow range of variation. Of course, it is impossible that there would be any linguistic feature which could always be identified more or less correctly regardless of the OCR quality level of the text. Still, the stable frequency group of features could be considered to include those features which are relatively unaffected by the decrease in OCR quality for a large range of the OCR quality spectrum, particularly down to the mid-70% OCR quality levels where even the normalization base ultimately starts to break down, as shown above.

The features which could be said to follow the stable frequency pattern include most prototypically (in alphabetical order) *attributive adjectives*, *present tense* verb forms, and *pronoun it*. It might also not be unreasonable to include in this class features such as *be as main verb*, *conditional adverbial subordinators*, *infinitives*, *past tense* verb forms, *predicative adjectives*, *pro-verb do*, *synthetic negation*, and *total adverbs*, for at least a part of the higher OCR quality range.

In general, the features following the stable frequency pattern appear to fall into two main classes: features covering a very large variety of items, and features targeting a very specific small item. Of the top three features, *attributive adjectives* and *present tense* belong in the former class, as both of them are based on the identification of almost any word belonging in certain large word classes (adjectives and verbs, respectively), which are relatively easy for a part of speech tagger to identify correctly, albeit the features are further limited to only include adjectives immediately preceding a noun (e.g. *red house*) and verbs in the present tense. On the other hand, the third feature, *pronoun it*, simply made up of the token *it*, belongs to the second class of specific small items. These items are protected by their shortness, since a random OCR error is more unlikely to hit them than a longer word consisting of more characters. The other features listed above also follow roughly these two classes, although their identification is often slightly more complex and depends on more of the context being identified correctly, which somewhat decreases the accuracy of their identification. For instance, *conditional adverbial subordinators* and *pro-verb do* mostly target specific short words within specific contexts, while *predicative adjectives* and *total adverbs* include large classes of items.

A further factor helping the features following the stable frequency pattern is that they tend to be common in texts. Since the features are already relatively well identified, as explained above, and they come in large numbers, even a few missed identifications will typically not constitute a particularly large change in the observed frequency of the feature in proportional terms.

The third and final main group of features is the small but curious group of features which show an increase in the observed frequency of the feature as the OCR quality decreases. At first, this result may seem contradictory to the overall expected effects of decreasing OCR quality, i.e. a decrease in frequency, but a closer inspection of the items following this pattern makes the reasons behind this pattern clear. The only features which clearly follow this pattern are *first person singular pronouns* and particularly *total other nouns*. The imperfect OCR process often creates strings of characters which a part-of-speech tagger cannot recognize, and most taggers have a tendency to tag such tokens as nouns, particularly if there are too many unrecognizable

tokens in close proximity to infer a different part of speech from the surrounding structure. This overtagging of nouns causes an overcount of supposed nouns in texts with lower OCR quality. Similarly, the imperfect OCR process may produce individual *I* characters which then are erroneously identified as the first-person pronoun *I*, though the effect is likely somewhat suppressed in the present analysis because of the requirement that the *I* has also been (mistakenly) tagged as a personal pronoun.

Feature	ρ	**Feature**	ρ
total prepositional phrases	-0.523	⋮	⋮
past tense	-0.390	infinitives	-0.192
emphatics	-0.388	WH-clauses	-0.188
nominalizations	-0.373	direct WH-questions	-0.175
private verbs	-0.371	amplifiers	-0.175
necessity modals	-0.353	WH relative clauses on subject position	-0.171
public verbs	-0.342	demonstrative pronouns	-0.163
perfect aspect	-0.340	first person plural pronouns	-0.159
possibility modals	-0.340	independent clause coordination	-0.144
conjuncts	-0.312	existential there	-0.143
analytic negation: not	-0.308	hedges	-0.143
time adverbials	-0.308	that adjective complement	-0.142
other adverbial subordinators	-0.294	indefinite pronouns	-0.133
split auxiliaries	-0.293	that relative clauses on subject position	-0.128
pied-piping relative clauses	-0.289	that relative clauses on object position	-0.118
seem & appear	-0.287	synthetic negation	-0.101
that verb complement	-0.285	place adverbials	-0.099
subordinator-that deletion	-0.279	predicative adjectives	-0.086
predictive modals	-0.266	concessive adverbial subordinators: (al)though	-0.085
causative adverbial subordinators: because	-0.265	pro-verb do	-0.085
by-passives	-0.265	conditional adverbial subordinators: if. unless	-0.082
agentless passives	-0.263	WH relative clauses on object positions	-0.059
total adverbs	-0.245	pronoun it	-0.026
suasive verbs	-0.231	present tense	-0.020
phrasal coordination	-0.210	discourse particles	-0.017
be as main verb	-0.210	contractions	0.074
downtoners	-0.203	attributive adjectives	0.083
third person personal pronouns	-0.201	split infinitives	0.136
second person pronouns	-0.199	first person singular pronouns	0.185
⋮	⋮	total other nouns	0.428

Table 1: Spearman's rank correlation coefficient (ρ) between decrease in the OCR quality and the change in the observed frequency of each feature.

Table 1 lists all features included in the analysis ordered by the text-level Spearman correlation of the proportional change in frequency from ECCO-TCP to ECCO-OCR with the OCR quality decrease[12] of the text. In other words, larger correlation values, either negative or positive, represent a higher tendency of the feature to (respectively) decrease or increase in frequency in the OCR version of the text when compared to the clean version, which means that the normalized frequencies of the feature have a higher tendency of becoming more unreliable with decreasing OCR quality. Conversely, correlation values close to zero mean that the feature is

[12]In order to make the interpretation of the correlation values more intuitive, the perfect OCR quality of 100% is taken as the baseline, and the correlation is calculated based on the amount of decrease of the quality percentage from the perfect baseline. For instance, an OCR quality of 80% is taken as a decrease in quality of 20%. This approach makes negative correlation values represent a decrease in observed feature frequency when the OCR quality decreases.

largely unaffected by decrease in OCR quality. Because of the effect of OCR quality on text length measures, the correlations were calculated for texts with mean OCR quality of 75% or higher.

An analysis of the correlation coefficients provides a different, complementary view of the relationship between OCR quality decrease and changes in observed feature frequencies. Most of the items mentioned in the above analysis can be found in Table 1 roughly where one would expect based on Figure 2, with features whose frequencies are more strongly affected by OCR quality having higher absolute correlation coefficients. However, some of the features differ from this pattern. As the most obvious case, *past tense* is the feature with the second-highest negative correlation with decrease in OCR quality, even though above I classified it potentially as a feature with a stable frequency across much of the higher OCR quality range. This is because of the fact that correlation is simply a measure of the consistency of the relationship, not of its magnitude. In other words, the high negative correlation value shows that the observed frequency of *past tense* verb forms does decrease consistently with decreasing OCR quality. However, at the same time, Figure 2 shows that in numerical terms the decrease appears to be quite small particularly in the highest quality levels. As such, in practical terms, the high negative correlation is not particularly impactful, and the problem with OCR quality is not as large with *past tense* verb forms as it is with most of the other features. Conversely, a feature which has a correlation close to zero, indicating little consistent change with regards to OCR quality, may still have a wide range of observed frequencies across all OCR quality levels. This pattern would mean that the calculated frequencies for the feature are unreliable regardless of the OCR quality level. Consequently, when only the very best-behaving features are needed, both the correlation and the actual magnitude of the change should be as close to zero as possible.

V 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 quality is not the same for all features. Features which are simpler to identify and cover large classes of words appear more likely to be identified correctly. Features relying on words short enough that they are unlikely to be hit by random OCR errors also tend to be more resilient. On the other hand, single-character words such as *I* start appearing erroneously as OCR quality decreases. Features with a higher rate of occurrence are also less affected by low OCR quality since a small number of misidentified instances does not change the overall picture much. At the same time, features which are more complex, requiring the correct identification of multiple consecutive items, are more at risk. Similarly, features based on limited lists of items are more easily affected by OCR errors. Furthermore, features which appear only rarely can be affected by random OCR errors much more easily than more common features.

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. The analysis of normalization bases shows a complete breakdown below the mid-70% range of OCR confidence, suggesting a potential minimum OCR quality for ECCO analyses. It must however be kept in mind that ECCO has been OCR processed in two parts using different OCR methods, which means that the OCR confidence scores between the two parts are not comparable with each other [Tolonen et al., 2021, 29].

The analysis of the difference between tokens and characters as normalization bases shows that

character count is more resistant to a decrease in OCR quality, supporting its use instead of the token count when dirty OCR data is being analyzed. When it comes to the selection of features to be analyzed, more frequent features covering large open classes of items (such as different word classes, excluding nouns) are likely the most resilient in the face of OCR errors.

On the other hand, multi-dimensional analyses of register consistently produce "compatible" results regardless of the exact set of features analyzed [McEnery and Hardie, 2012]. While the low quality of feature identification is not inherently a problem for analysis methods like MDA [Liimatta et al., 2023], focusing on a smaller set of more resilient features may lead to better results even from lower-quality textual data when using such methods.

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. *Variation across speech and writing*. Cambridge University Press, Cambridge, 1988. ISBN 0-521-32071-2. doi: 10.1017/CBO9780511621024.

Douglas Biber and Susan Conrad. *Register, genre, and style*. Cambridge University Press, Cambridge, 2009. doi: 10.1017/CBO9780511814358.

Jacob Eisenstein. 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, 2013. URL https://aclanthology.org/N13-1037.

Stephen H. Gregg. *Old Books and Digital Publishing: Eighteenth-Century Collections Online*. Cambridge University Press, December 2020. ISBN 978-1-108-76741-5 978-1-108-72069-4. doi: 10.1017/9781108767415.

Mark J. Hill and Simon Hengchen. 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, 2019. doi: 10.1093/llc/fqz024.

Aatu Liimatta, Yann Ryan, Tanja Säily, and Mikko Tolonen. 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, October 2023. ISSN 2704-1441. URL https://journals.uio.no/dhnbpub/article/view/10668.

Daniel Lopresti. Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(3):141–151, September 2009. ISSN 1433-2825. doi: 10.1007/s10032-009-0094-8.

Tony McEnery and Andrew Hardie. *Corpus linguistics: Method, theory and practice*. Cambridge University Press, Cambridge, New York, 2012. doi: 10.1017/CBO9780511981395.

Mikko Tolonen, Eetu Mäkelä, Ali Ijaz, and Leo Lahti. Corpus linguistics and Eighteenth Century Collections Online (ECCO). *Research in Corpus Linguistics*, 9(1):19–34, April 2021. ISSN 2243-4712. doi: 10.32714/ricl.09.01.03. URL https://ricl.aelinco.es/index.php/ricl/article/view/161.

Myriam C. Traub, Jacco van Ossenbruggen, and Lynda Hardman. Impact Analysis of OCR Quality on Research Tasks in Digital Archives. In Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, editors, *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 252–263, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24592-8. doi: 10.1007/978-3-319-24592-8_19.

Oxana Vitman, Yevhen Kostiuk, Paul Plachinda, Alisa Zhila, Grigori Sidorov, and Alexander Gelbukh. Evaluating the Impact of OCR Quality on Short Texts Classification Task. In Obdulia Pichardo Lagunas, Juan Martínez-Miranda, and Bella Martínez Seis, editors, *Advances in Computational Intelligence*, Lecture Notes in Computer Science, pages 163–177, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19496-2. doi: 10.1007/978-3-031-19496-2_13.

A APPENDIX 1: 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: (al)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