Results from rough data? The large-scale study of early modern historiography with multi-dimensional register analysis

Aatu Liimatta¹, Yann Ryan¹, Tanja Säily¹ and Mikko Tolonen¹

¹University of Helsinki

Abstract
Multi-dimensional register analysis is a methodology which can be used to extract functional dimensions from a set of texts. These dimensions describe various functional differences between the set of texts. The differences can be due to various situational constraints related to the production of the text, or they can be related to differences in the author’s intent and communicative purpose. While this methodology has seen considerable use in contemporary linguistics, it has been less used in historical linguistics, and even less so in history, even though the ability to differentiate between various textual functions in historical data would be extremely useful and interesting from the point of view of a historian. In this paper, we perform a pilot study of multi-dimensional register analysis on a subset of texts from Eighteenth Century Collections Online (ECCO). In particular, our goal is to find out whether this kind of analysis is possible in the first place, or if it is hindered too much to be useful by the low quality of the ECCO data produced by optical character recognition (OCR). To do this, we first perform the analysis on ECCO data, after which we compare the results with results from running the same analysis on the same set of texts from ECCO-TCP, a manually cleaned subset of ECCO data. Our results show that not only are the results from the ECCO analysis interpretable, but they are also highly similar with the results from ECCO-TCP. Multi-dimensional register analysis appears to be a very promising and robust method which can work well even with low-quality data.

Keywords
register analysis, OCR issues, Scottish Enlightenment, historical writing, Eighteenth Century Collections Online (ECCO)

1. Introduction

Register analysis is a linguistic approach which examines language use in different situations and functions [1, pp. 6-7]. With methods such as the Multi-Dimensional Analysis, one can find dimensions of functional variation, which describe functional differences between texts in a dataset. However, studies of historical texts using these methods are rare. Such studies would provide valuable insight into historical linguistic variation.
While historians have long relied on context to interpret meaning and placed a particular emphasis on linguistic acts, the full extent of the linguistic context in larger corpora has often been overlooked in the study of the process of history writing (historiography). Register analysis can provide valuable insights into how authors construct and convey their accounts of the past, but historians have not yet utilized this tool to its full potential. The purpose of this paper is to address these shortcomings and bridge these gaps.

Our article seeks to explore the potential of using register analysis for Eighteenth Century Collections Online (ECCO), a large but noisy dataset of historical texts, and to systematically examine the convergence of register analysis and the analysis of early modern historical writing. This intersection has the potential to produce far-reaching impacts, particularly when history and sociolinguistic analysis are brought together. By analyzing language use in different contexts, register analysis can shed light on how historical narratives were constructed and conveyed by authors of the past. Thus, our article aims to highlight the potential benefits of incorporating register analysis into historical analysis, emphasizing the importance of taking a comprehensive and interdisciplinary approach to the study of the past.

2. Background

Register analysis is a field of linguistics which focuses on the use of language in different situational contexts and for different purposes. Traditionally, register analysis focused on differences between a priori register distinctions, such as “formal” and “casual” or “written” and “spoken” registers. However, since the end of 1980s, so-called Multi-Dimensional Analysis (MDA), originally developed by Douglas Biber [2], has been extremely influential in quantitative corpus-based register analysis. MDA rejects the idea of such a priori register distinctions. Instead, through computational and statistical analysis of texts, it is possible to extract multiple dimensions along which various functional linguistic features vary in the dataset in question. By analysing the sets of features associated with the dimensions and the texts in which they appear, it is possible to find functional distinctions which differentiate the texts in the dataset.\footnote{Other methods can also be used for register analysis. Most register analysis methods are based on the idea of comparing the occurrences of functional linguistic features in different texts. We have chosen to use MDA in our analysis since it likely is the best-known and most widely used of these methods in linguistics. However, the findings will also be relevant for studies using other similar methods of register analysis.}

Register analysis applied to large historical datasets has great potential, offering the possibility of studying historical language use at a much greater scale than usually done. Historians have not frequently leveraged the benefits of extensive linguistic analysis to establish a broader framework for their research pursuits. The subject of this analysis, ECCO, serves as a good example. ECCO contains over 30 million pages of text, from over 200,000 documents, making it much larger than most hand-curated historical corpora. [3] In 2004 Gale claimed that ECCO contained every significant work in the English language printed in the eighteenth century, plus thousands of other important works [3, p. 56], though studies have shown that its representation of the entirety of the eighteenth century publishing landscape is uneven [4]. Nevertheless, a linguistic analysis which could confidently be applied to this dataset would provide new opportunities to study the wide variety of texts within ECCO, including pamphlets, legal
documents and statutes, technical texts or instruction manuals, and non-elite writing. Register analysis, for instance, may allow us to understand more about the nature of writing about the past in the eighteenth century, by helping analyse changes in the style and communicative functions of historiography. History and other expository forms of writing moved towards professionalization throughout the century [5], and one hypothesis to be tested is that this may be reflected in the register dimensions.

However, while MDA has been used with great results for a long time (see [6][7]), only a handful of MDA studies have been conducted using historical data (e.g. [5][8][9]). Furthermore, these studies have used smaller, curated and hand-corrected corpora, meaning that the vast majority of historical text documents are excluded. ECCO and similar large-scale historical datasets are often not used in whole for linguistic or historical analyses, in part because of the low quality of the text data. ECCO text data has a large number of errors, for several reasons. First, the source texts themselves are of varying quality and suffer from printing issues such as bleed-through. Second, modern OCR engines are often not calibrated for the fonts used, and third, the OCR has been produced mostly from low-quality bitonal scans of microfilm. Hill and Hengchen [10] have shown that for a sample of ECCO, the mean token-level accuracy was 77%.

Our aim in this paper is to show that even in spite of these significant errors, certain types of linguistic analysis, such as MDA, can produce robust results which compare surprisingly well to a much cleaner, transcribed set of texts. We do this by comparing the above-described dataset (ECCO-OCR hereafter) to another, the ECCO Text Creation Partnership (ECCO-TCP). ECCO-TCP, released in 2011, is a collection of approximately 2,100 documents, covering a range of books from the eighteenth century. [11] ECCO-TCP is double-keyed, meaning that each document was transcribed twice, and a third transcriber resolved disputes. While it does have errors, it is produced in a similar manner and likely has a similar error rate to most linguistic corpora. Originally conceived as part of a larger project [3] (a sister project, EEBO-TCP, contains over 40,000 documents), ECCO-TCP is nevertheless useful for comparing a clean dataset with the OCR version of the same dataset, because it has full overlap with texts from ECCO-OCR.

ECCO-TCP and OCR have been used for similar comparisons before. Hill and Hengchen [10] compare ECCO-OCR and ECCO-TCP with respect to a number of linguistic tasks, including topic modeling, authorship attribution, collocation analysis and vector space modeling. They find that topic modeling works quite well in ECCO-OCR, while collocation analysis can be more problematic unless great care is taken in the selection of subcorpora and research questions. Their study indicates that an OCR accuracy of 80% (in terms of F1 Scores calculated at the page level) is sufficient for most tasks, whereas a level below 70–75% significantly degrades the quality of the results. The greatest number OCR errors are found in words containing the long-s or ligatures. However, while Hill and Hengchen consider various bag-of-words approaches, we instead focus on differences in results provided by a multi-stage computational and statistical analysis procedure which relies on the correct identification of longer linguistic constructions.

---

2https://textcreationpartnership.org/tcp-texts/ecco-tcp-eighteenth-century-collections-online/
3. Data and methods

3.1. Datasets

As our source of data, we use the ECCO-TCP dataset and the equivalent set of texts from ECCO-OCR. For this proof of concept pilot analysis, we use the ECCO-OCR subset as our primary dataset, since in the present paper our goal is to test the ability to use the MDA methodology for analyses of large-scale historical text data despite the OCR quality issues. The ECCO-TCP dataset is then used as a point of comparison, to contrast the results of the ECCO-OCR analysis with a clean baseline to gauge the robustness of the methodology on less-than-perfect datasets.

In this study, we focus primarily on functional register differences between genres. To do this, we needed an existing, externally-produced set of genre labels which allowed us to analyse some kind of categorical differences between the texts. We used a set of genre labels generated using a state-of-the-art method [12]. This taxonomy of genre labels was custom-created to reflect sensible eighteenth-century generic distinctions, such as religious texts, histories, ‘scientific’ works, and so forth.

3.2. Data processing

First, both datasets, the ECCO-OCR subset and ECCO-TCP, were tagged for part of speech. The datasets were then analyzed for the linguistic features of interest. The algorithms to identify the features are based on Biber [2]. Most of Biber’s original 67 features were included. However, a handful of the features are difficult to identify automatically, and so were not included in the analysis. Some of the features, viz. type-token ratio and mean word length, were excluded from the analysis because they are particularly vulnerable to low OCR quality. Furthermore, *first person pronouns* were split into singular and plural following some earlier studies (e.g. [13]), since the two features exhibit different functional behavior. In the end, the analysis included 58 features.

The number of occurrences of the features were then normalized. Normally, this would be done by dividing the number of occurrences of the feature by the number of running words or tokens in the text. However, because of the low quality of the OCR, the number of tokens extracted from the text by the part-of-speech tagging pipeline can be very far off from the “true” number. This is largely caused by the OCR process, which commonly splits individual words into multiple parts. Because of this, we instead normalized the feature counts by the number of *characters* in the text. The character count will also not be exactly correct, but it will be proportionally closer to the true value.

3.3. Multi-dimensional analysis

Since its inception, Multi-Dimensional Analysis (MDA) [2][14][1, p. 6] has been very influential in the field of quantitative corpus linguistics as a method for studying functional variation within language. The method is based on a simple idea: that linguistic features are functional, and therefore tend to be used more in the kinds of contexts for which they are particularly well-suited. Consequently, linguistic features well-suited for similar situational and functional
contexts can be expected to appear in the same texts, when the situation or function of the text calls for them, and similarly be absent at the same time, when they are not needed.

By analyzing the co-occurrence patterns of a large number of linguistic features, it is possible to find groups of co-occurring features, which tend to appear in the same texts and be absent from the same texts. In practice, this is typically done using factor analysis, though other methods are also possible [15][16][17]. The groups of co-occurring features, i.e. the factors, can then be interpreted as functional dimensions of variation based on the assumption that the features in the group are present or absent due to some set of underlying communicative functions or situational concerns. The functional tendencies of e.g. textual genres or other groups of texts can then be compared in terms of their positioning on these functional dimensions.

In this study, the factor analysis was run twice. After the first run, any feature which did not have an absolute loading higher than .35 on any of the eight extracted factors was removed from the dataset, in order to reduce noise caused by features which are not central to the patterns of variation observed in the data. The factor analysis was then run again on the reduced feature set. The number of factors was decided based on an inspection of factor solutions with different numbers of factors, from four to nine. In the end, seven factors were extracted, as this solution was found to be the most readily interpretable.

After this, dimension scores were calculated for each text on each of the dimensions. First, the feature frequencies were standardized to a mean of 0 and standard deviation of 1, to ensure that high-frequency features would not drown out low-frequency features in the analysis. Then, as is commonly done in MDA studies to calculate the dimension score for a text on a dimension [2], the standardized frequencies in that text were added together for every feature which had a loading equal to or greater than 0.3 on that dimension.

4. Analysis

The dimension scores for the texts were then plotted across genres in ECCO to aid in the functional interpretation of the dimensions, and to enable inter-genre comparisons of functional tendencies. The positioning of the genres along the extracted dimensions is shown in Figure 1.

4.1. Dimension 1: Past/narrative/literary vs. non-past/speech-like focus

Table 1
Features associated with the positive and negative poles of Dimension 1

| + past tense, agentless passives, by-passives, perfect aspect, amplifiers, total prepositional phrases, downtoners, hedges, pied-piping relative clauses |
| - contractions, present tense, discourse particles, other adverbial subordinators |

Table 1 shows the features associated with Dimension 1. The main dichotomy between the positive and negative poles of Dimension 1 appears to be the distinction between past tense and present tense, indicating a functional division between a past temporal focus and a present or future temporal focus. This division is also supported by the presence of the perfect aspect on
Figure 1: The positioning of the texts in the ECCO-OCR subset on the seven extracted dimensions by genre.

the positive pole of the dimension. In this manner, Dimension 1 is somewhat aligned with the narrative universal register dimension proposed by Biber [6].

At the same time, many of the other features on the dimension point towards a different kind of distinction. Features such as agentless passives, by-passives, and total prepositional phrases are associated with a more abstract, impersonal, “literary” style, whereas the complementary features such as contractions, discourse particles and other adverbial subordinators are more typical of more “oral” or speech-like registers.

In Figure 1 we can see that the genre which differs the most from the rest on Dimension 1 is arts. The “arts” genre mostly contains works such as plays and poetry, which indeed tend to use the present tense and at the same time contain many more speech-like features than the average genre. Literature is more spread out on this dimension, because in addition to speech-like dialogue sections, it also contains narrative sections making use of the features of the narrative pole of the dimension, such as past tense. The other genres are more strongly associated with the narrative, literary side of the dimension.
4.2. Dimension 2: Involved interpersonal focus

The features associated with Dimension 2 are shown in Table 2. The positive pole of the dimension is dominated by *first person singular pronouns* and *second person pronouns*, which point towards an involved, interpersonal function for this dimension. These functions are also supported by many of the other features on the dimension. For instance, *private verbs*, i.e. verbs which express internal mental states, such as *think*, *feel*, and *believe*, are often used in contexts where the speaker or writer is involved in the produced text. Similarly, *direct WH-questions* imply an interactive or interpersonal focus. Features such as *WH-clauses* and *time adverbials* help locate and refer to a specific referent or time.

On Biber’s well-known first dimension [2], this kind of involved register is placed in contrast with informationally dense registers. Dimension 2 in the present study also bears some signs of such a dichotomy. Both of the features on the negative pole of the dimension can be considered to contain a higher informational load; on the positive pole, instead, *pro-verb “do”* replaces a whole verb phrase, reducing the informational load of the text.

In Figure 1, arts again have a clearly different score on Dimension 2 compared to the other genres. Plays contain a high number of spoken lines, which tend to be very involved and interpersonal in nature, and poetry is also similar in many ways. The other genres are very different along this dimension, even literature, which is still the closest to arts with its higher number of characters’ lines but a large proportion of non-spoken description when compared to arts.

4.3. Dimension 3: Static statement

Table 3 lists the features associated with Dimension 3. The main feature on this dimension is “*be*” as *main verb*. The verb *be* is also closely linked with other features on the dimension. For example, *existential “there”*, i.e. the construction “there is/there are”, is formed using the verb *be*, and similarly *predicative adjectives* are adjectives occurring in a predicative position,
i.e. following the verb be, such as in the sentence “The house is big” (as opposed to attributive adjectives, such as in the phrase “the big house”).

In general, the positive pole of this dimension appears to describe texts which tend to use more predicative, or static, expressions. In other words, texts in the positive pole of the dimension express more than average the nature of something, that something is (like) something, or that there is something somewhere, as opposed to using other verbs to express what that something does, or what is done to it.

While the placement of the genres on this dimension is relatively level in Figure 1, there are still small differences in their tendencies. Arts, law, philosophy, politics and religion as well as scientific improvement appear to use this kind of description slightly more, whereas history and literature use it slightly less. This can be explained as history and literature having slightly more focus on active actions, whereas the other genres are slightly more interested in the nature of things. However, as can be seen in the figure, this difference is quite subtle.

4.4. Dimension 4: Expression of options and possibilities

| Table 4 |
| Features associated with Dimension 4 |
| + possibility modals, necessity modals, predictive modals, split auxiliaries, present tense, analytic negation: not, conditional adverbial subordinators: if & unless, infinitives, suasive verbs |

The main features associated with Dimension 4, listed in Table 4, are clearly the three types of modal verbs: possibility modals (such as can or may), necessity modals (such as must or should), and predictive modals (such as will or would). These features already clearly position texts scoring highly on this dimension as talking about various options, necessities and possibilities. The other features on the dimension support this function. For instance, analytic negation “not” is used to reverse the modality of the three modal verb classes to express that something can, must, or will not be done. Similarly, conditional adverbial subordinators “if” & “unless” are used to qualify and limit the scope of the modal content.

This dimension also contains all of the features included on Biber’s original Dimension 4 [2], which he labeled Overt expression of persuasion or Overt expression of argumentation [6]. However, Dimension 4 of the present study also includes features not included on Biber’s original dimension. Still, it is true that the dimension in this study, too, has a certain connection to persuasion and argumentation. We have tentatively given the dimension a wider label, Expression of options and possibilities.

In Figure 1, genres such as politics and religion are naturally placed high on this dimension, and even law, expressing the various consequences and means of dealing with different legal situations. On the other hand, history and literature have clearly lower scores on this dimension: they express how things were or what happened in the past or in the fictional world of a story, and so these genres do not need so many hypotheticals.
4.5. Dimension 5: Complex reference

The features associated with Dimension 5, shown in Table 5, are all complex clauses acting as subjects, objects, or modifiers. These features, such as WH relative clauses on object positions (e.g. “the question which was the main topic of discussion yesterday”), pied-piping relative clauses (e.g. “the manner in which it was done”), and “that” adjective complements (e.g. “I am glad that you think so”), all help to integrate large amounts of information and expressions of complex relationships between ideas, actors and objects into more condensed textual units.

In Figure 1, genres such as philosophy, politics, religion and scientific improvement are positioned relatively high on this dimension. These genres deal with topics which are often quite complex, and as such can make use of constructions which refer to complex referents. Other genres, such as history and law, have slightly lower scores, but also potentially a great deal of internal variation. Finally, literature and particularly arts do not typically use constructions like these that much.

4.6. Dimension 6: Evaluation and qualification of information

Table 6
Features associated with Dimension 6

+ total adverbs, emphatics, hedges, amplifiers, downtoners

All of the features on Dimension 6, listed in Table 6, are words which in some way evaluate or qualify information. Emphatics, hedges, amplifiers, and downtoners are used to either intensify or reduce the certainty or strength of the claim being made. Similarly, adverbs express the manner of performing an action, and in that way qualify or evaluate the content of the expression.

Figure 1 shows that most genres do not exhibit very large differences in terms of Dimension 6. However, scientific improvement and philosophy are on average slightly above the others. After all, evaluating and qualifying various ideas and claims is a central part of both fields.

4.7. Dimension 7: Third-person focus

Table 7
Features associated with Dimension 7

+ third person personal pronouns, public verbs, infinitives, suasive verbs, perfect aspect

Dimension 7 features, as shown in Table 7, are all associated with the focus being on a third-person actor. While third person personal pronouns are the clearest sign of this, other
features on this dimension also support the function of third-person focus. Most importantly, public verbs (i.e. verbs which express publicly observable actions, such as *say* or *explain*) and suasive verbs (i.e. verbs which imply intention to bring about change, such as *command* or *recommend*), clearly focus on the actions of a third-person actor.

Most genres do not differ that much on Dimension 7 in Figure 1. However, a couple of the genres have slight differences compared to the others. Most importantly, scientific improvement is below the others, focusing more on science and engineering than any person. On the other hand, both history and politics are placed slightly higher than the rest, since these genres often focus on historical characters or political actors and their deeds.

5. **OCR vs. TCP**

As shown in the previous section, the features associated with the seven dimensions form largely meaningful groups whose function and use is quite readily interpretable based on this analysis. This result is already very promising in terms of the robustness and usability of the MDA methodology even on less-than-perfect OCR data, and indicates that at least some useful results can be acquired from imperfect datasets.

However, based on these results only, it is unclear how similar the dimensions extracted from the imperfect OCR dataset are to dimensions which would be extracted from a clean version of the same dataset. In order to gauge the similarity of this result using the imperfect dataset to a result from a clean dataset, we performed the exact same analysis steps, including part-of-speech tagging, feature identification, and MDA with the same parameters, on the ECCO-TCP set of texts. Since the original analysis was based on the subset of texts from ECCO-OCR which matches the texts in ECCO-TCP, we can get a good idea of what the results would look like if the exact same analysis process was performed on a clean version of the same dataset.

The results from the analysis of the TCP dataset are very promising from the outset, when compared to the results of the OCR subset analysis. The dimensions extracted from the TCP dataset are in a slightly different order, reflecting small changes in their relative importance, but the dimensions comprise largely the same features as the OCR dimensions, making it trivial to map the TCP dimensions to the OCR dimensions one-to-one. Importantly, while there are small number of added or removed features, the features most strongly associated with the dimensions do not change; all changes in features on the dimensions take place among the features with weaker loadings on the corresponding factors.

But how much is the positioning of the texts and genres on the dimensions affected by these small differences in the feature makeup of the dimensions, the differences in the part-of-speech tagging, and the corresponding differences in the identification of the features in the texts? Figure 2 plots the results of the TCP analysis alongside the results of the OCR analysis in order to see how much the positioning of the genres differs on the seven dimensions. Based on this figure, it seems clear that the patterns are quite similar between the two datasets. There are small differences, but the dimensions clearly capture similar phenomena in the two datasets.

---

3 The TCP dimensions have been labeled here with the numbers of the equivalent OCR dimensions to simplify the comparison of the dimensions extracted from the two datasets.
A similar vertical positioning of a pair of boxes means that the analysis produces similar results for the two datasets on that dimension for that genre.

To further illustrate the similarity of the dimensions extracted from the OCR and TCP datasets, Table 8 shows the strong correlation of the dimension scores from the OCR and TCP analyses for all seven dimensions in all eight genres. That is, for any of the dimensions, a text scoring highly on that dimension in the OCR dataset is also very likely to score highly on the equivalent dimension in the TCP dataset, and vice versa.

MDA requires automated identification of complex linguistic features, and it would stand to reason that the low OCR quality of ECCO-OCR would be extremely detrimental to such analysis. However, it is not entirely unexpected that the method works even with this kind of data. A wide variety of MDA studies have used different sets of features. While the dimensions they have produced are not the exact same, the dimensions are nevertheless typically “compatible” with each other [18], meaning that they capture similar kinds of linguistic phenomena and functions even if their precise structure differs. For our purposes, if different sets of features produce similar results, it means that some results can still be expected from the analysis even if not all features in the set of features are identified correctly, such as because of OCR errors. MDA also mitigates the detrimental effect of OCR errors in another way: because the method works based on the correlations of the presence and absence of the features. Even if every
Table 8
P-values for the correlation between the dimension scores for ECCO OCR and TCP by genre

<table>
<thead>
<tr>
<th>Genre</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>D5</th>
<th>D6</th>
<th>D7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arts</td>
<td>0.923</td>
<td>0.958</td>
<td>0.941</td>
<td>0.883</td>
<td>0.877</td>
<td>0.739</td>
<td>0.686</td>
</tr>
<tr>
<td>History</td>
<td>0.702</td>
<td>0.887</td>
<td>0.948</td>
<td>0.932</td>
<td>0.767</td>
<td>0.763</td>
<td>0.898</td>
</tr>
<tr>
<td>Law</td>
<td>0.932</td>
<td>0.932</td>
<td>0.925</td>
<td>0.892</td>
<td>0.876</td>
<td>0.852</td>
<td>0.732</td>
</tr>
<tr>
<td>Literature</td>
<td>0.951</td>
<td>0.945</td>
<td>0.853</td>
<td>0.900</td>
<td>0.894</td>
<td>0.844</td>
<td>0.884</td>
</tr>
<tr>
<td>Philosophy</td>
<td>0.899</td>
<td>0.937</td>
<td>0.886</td>
<td>0.855</td>
<td>0.680</td>
<td>0.931</td>
<td>0.718</td>
</tr>
<tr>
<td>Politics</td>
<td>0.926</td>
<td>0.911</td>
<td>0.808</td>
<td>0.907</td>
<td>0.900</td>
<td>0.811</td>
<td>0.657</td>
</tr>
<tr>
<td>Religion</td>
<td>0.930</td>
<td>0.948</td>
<td>0.862</td>
<td>0.929</td>
<td>0.899</td>
<td>0.857</td>
<td>0.754</td>
</tr>
<tr>
<td>Scientific improvement</td>
<td>0.935</td>
<td>0.935</td>
<td>0.771</td>
<td>0.930</td>
<td>0.882</td>
<td>0.878</td>
<td>0.817</td>
</tr>
</tbody>
</table>

feature is only identified a small fraction of the times it actually appears, its correlation patterns should not change much: when the feature is identified correctly, it still appears in the same texts with other features which share its communicative functions.

6. Discussion

Historians have traditionally relied on context as their primary tool, but linguistic context has often been overlooked in their analytical toolkit. Register analysis involves examining language use in different contexts and identifying the features that characterize them. Despite its potential to provide valuable insights into the historical events and texts under examination, register analysis has rarely been utilized to its full potential by historians.

As our analysis shows, even when working with imperfect historical datasets, the multi-dimensional method of register analysis (MDA) produces dimensions “compatible” [18] with e.g. the register universals [6]. The method can therefore also be useful for providing insight into functional variation in historical texts for historians and linguists alike. By analyzing texts in terms of register dimensions and the features associated with them, we can find out much more about their various functions. For instance, when applied to historical narratives, register analysis can shed light on the ways in which authors use language to construct and convey their accounts of the past. In higher-level analyses, we can consider the dimensions and their uses in historical writing, and see to which degree the various functions appear in the texts we are interested in. But it is also possible to zoom in from this high-level picture, and analyze in detail the occurrences of the features associated with those dimensions in the texts.

For example, by analyzing texts which place high on Dimension 1, the past/narrative/literary vs. non-past/speech-like dimension, in terms of how they use the features associated with the dimension, such as past tense, agentless passives, and by-passives, we can gain insight into how authors attribute actions and events to particular actors or causes. In a similar manner, the other features on the dimension, as well as the other dimensions and their features, can also be indicative of an author’s stance towards the events they describe. Still other features, such as those associated with Dimension 5, can affect the flow and coherence of the narrative. By examining all of these different features and their distributions across historical texts, register analysis can help us better understand the ways in which language is used to construct historical
narratives. Many of the features of dimensions 1 and 5 are demonstrated (highlighted with bolding and italics, respectively) in the following passage from a text in the “history” genre:

He summoned a Parliament, to whom he made bitter complaints against the irruption of the Scotch, the absurd imposture which was countenanced by that nation, the cruel devastation which they had spread over the northern counties, and the complicated affront which had thus been offered both to the King and kingdom of England. (David Hume, 1759, *The history of England*)

Another important aspect of register analysis in historical narratives is the use of Dimension 3, the static statement dimension. This involves examining the ways in which language is used to make factual claims and convey information. For example, texts scoring high on this dimension may have a more declarative or authoritative tone, arising from simple statements of fact, e.g. using the verb be as a main verb as well as existential there and predicative adjectives, and the use of the pronoun it to emphasize the importance or significance of certain events or objects. Pro-verb do, indefinite and demonstrative pronouns, as well as causative and conditional adverbial subordinators, can all contribute to the overall clarity and specificity of the statement. Additionally, the use of other adverbial subordinators and attributive adjectives can further refine or contextualize the information being presented. By analyzing these linguistic features in the context of historical narratives, register analysis can help identify the ways in which authors use language to make claims and convey information in a persuasive and authoritative manner. Many of the features associated with Dimension 3 can be seen in the following two examples, from the “history” and “religion” genres respectively, using these features to build an authoritative tone and describing the nature of things.

There is no possible cafe, either of immorality or even inconvenience, but what is within the reach and correction of the COMMON LAW; for, *it is* a rule therein, that " *nothing* which is against REASON is lawfull:" and, fureiy, every thing that *is immoral is " against reason ;" and again, by another rule, " *nothing* ( that is inconvenient is lawful).§" (Granville Sharp, 1784, *An account of the ancient division of the English nation into hundreds and tithings*)

With refpet to what are called denominations of religion, if every one is left to judge of its own religion, *there is* no such thing as a religion that is wrong; but if they are to judge of each others religion, *there is* no such thing as a religion that is right; and therefore, all the world is right, or all the world is wrong. (Thomas Paine, 1791, *Rights of man*)

A crucial application of register analysis for historians would be the development of a large-scale automated system that could tag longer narrative sections in multiple texts on a similar topic and automatically detect instances where the register shifts to a more declarative or statement-like register, indicating a distinct type of activity within the framework of historical

---

1All examples in this paper have been copied directly from the OCR text files and include the original OCR errors.
writing. This would be an invaluable tool for historians seeking to analyze differences among early modern histories that follow common practices, enabling them to pinpoint those historians who break the narrative mold when discussing common events or themes. By taking the historian’s analysis to a new level, such an automated system could help uncover hidden insights and provide a more nuanced understanding of what particular authors were doing when writing about historical events and narratives.

Furthermore, the use of static statement in historical narratives can reveal a distinct aspect of history known as ‘universalism’, the tendency of authors to use history as a means of uncovering universal truths about human nature and progress, rather than simply recounting past events, for example the teleological approach taken by many eighteenth-century historians [19]. For instance, an author who makes a statement rather than providing a detailed explanation of a particular event or phenomenon may be attempting to convey a broader, more abstract idea. By using language to make categorical claims and assert general principles, authors can create a sense of universality that transcends individual historical contexts. Register analysis can give us insight into the ways in which authors use language to construct and convey these broader historical narratives, and the implications that this has for our understanding of the past.

Our next objective is to embark on large-scale analysis of historiography, starting by compiling a comprehensive and uniform corpus of relevant texts. Developing a method for tagging registers of interest, such as narrative and statement-like registers, will enable us to analyze the corpus at a larger scale. Our analysis will then focus on particular cases, allowing us to conduct a more nuanced examination of how specific early modern historians use registers. This endeavor will help advance the state of the art in the field and generate new knowledge about crucial differences between authors writing about British history during the Stuart era and beyond.

7. Conclusion

Our paper demonstrates that even when working with datasets suffering from significant problems with low-quality OCR, the multi-dimensional method of register analysis (MDA) can still be used to good effect. Through analysis of the extracted functional dimensions, we have shown that these dimensions are not only meaningful and interpretable in themselves, but also within the context of the genres present in the dataset. This is of particular relevance to the analysis of early modern historiography, which was a focus of our study. By highlighting the potential benefits of MDA in this context, our paper contributes to a growing body of literature demonstrating the value of interdisciplinary approaches to historical analysis. Furthermore, the comparison of these dimensions with the equivalent dimensions from the clean ECCO-TCP dataset shows that there are only minor differences between the dimensions extracted from the clean and dirty datasets and the positioning of the texts along those dimensions, particularly when the results are analyzed in aggregate. Overall, the two analyses capture similar phenomena, strongly supporting our hypothesis that MDA is a robust methodology which works well even with lower-quality data. Due to the many issues with the underlying data, the OCR quality of large datasets is unlikely to substantially improve in the near future. Consequently, any methods which can work reasonably well with such low-quality data are desirable.

The ultimate goal of our work with register analysis is to enable its application to full ECCO
datasets, and to utilize this approach in the analysis of early modern historiography from a historical perspective – an area which has yet to be fully explored. By demonstrating the value of this approach, we hope to encourage more researchers to incorporate register analysis into their historical analysis and contribute to a more nuanced understanding of the past.

Acknowledgments

This study is a part of the Academy of Finland funded project *Rise of Commercial Society and Eighteenth-century Publishing* (RiCEP; grant numbers 333716 and 333717).

References


