Shedding Light on Dickens' Style through Independent Component Analysis and Representativeness and Distinctiveness

Carmen Klaussner

European Masters Program in Language & Communication Technologies (LCT)

University of Groningen
Department of
Literature and Arts

Thesis Supervisors:
Prof. Dr. John Nerbonne
Dr. Çağri Çöltekin

University of Nancy 2
Department of
Cognitive Science

Thesis Supervisor:
Dr. Jean-Charles Lamirel

university of groningen

UNIVERSITÉ DE LORRAINE

Date: 31 July 2013

Acknowledgments

I would like to thank my supervisors in Groningen, John Nerbonne and Çağri Çöltekin, for their excellent guidance during this work, in particular for keeping an open mind for my interests and ideas, for having the oversight to terminate my listless wanderings into mathematical limbo, and finally for their understanding and patience in the face of frequent bafflement.
Also, I would like to sincerely thank my supervisor in Nancy, Jean-Charles Lamirel, for continued support from the distance in the joined endeavour of adapting *Independent Component Analysis* for style analysis.
I hope that this work will give some credit to their joined effort and that they would also regard it as such.

CONTENTS

“To them, I said, the truth would be literally nothing but the shadows of the images [. . .]

And if they were in the habit of conferring honours among themselves on those who were quickest to observe the passing shadows and to remark which of them went before, and which followed after, and which were together; and who were therefore best able to draw conclusions as to the future, do you think that he would care for such honours and glories, or envy the possessors of them?”

-Plato's 'The Republic', Book VII

1 INTRODUCTION

The concept of *style* is a characteristic that is somewhat difficult to define or measure distinctly and is thus far less tangible compared to other possible characteristics. The concept of an author's style, the *feel* of his writings, is reminiscent of the *feel* of a piece of music that we instinctively perceive to originate from a particular composer, such as Chopin or Debussy, without being quite able to name the exact reasons, because style is a composite feature, a sum of entwined parts.

Plato's *Allegory of the Cave* (Plato and Jowett 2011) describes some prisoners in a cave, who are chained so that they face the wall and are unable to turn their heads towards the light, which holds the truth. They can only glimpse at reality through the shadows projected at the wall in front of them, without knowing whether what they observe is in any way close to the truth. This allegory is often employed to express the sheer difficulty of any knowledge-seeking person at making deductions solely on the basis of some observations (shadows) without knowing their relationship to reality. Like the prisoners, we are reaching out for the truth, while not knowing which part of the shape reflecting reality is representative of the real object.

The associated predicament may be even be more fitting with respect to style analysis, where we are not only interested in a solid explanation of what we observe, but also in the explanation itself.

In our "cave" of style analysis, we imagine there to be two kinds of *prisoners*. The first is the expert or the close observer, who continues watching one or maybe a couple of particular shapes and is able to recognize details and spot one shape among many, even when a little distorted, but all others remain a puzzle to him. The second kind of prisoner tries to abstract and to generalize. He does not know any shape well, but has techniques that can tell him whether two shapes are similar and therefore finds those properties common to all shapes and those distinctive only for some. The first type of prisoner is very accurate, but lacks generalization ability, while the second type of prisoner is less specific, although potentially more impartial, as he may draw conclusions from his findings. Even if ever escaping from the cave is unlikely, one step closer towards the light might be achieved through combining beliefs and findings about style from both perspectives and fixing our vision on the shapes in front of us.

Thus, for this thesis, we are content to settle on a distortion of the truth, but hoping for some interesting insights into the style of an author. The following work is a tentative attempt at measuring what is generally conceived to be an author's *fingerprint*, in particular with respect to the author Charles Dickens, and all results should essentially be seen in this light, namely a modest attempt at quantifying something that is in fact very difficult to measure.

The remainder of this work is structured as follows: chapter 2 presents an insight into the diverse aspects of non-traditional style analysis, considering both past and present. Chapter 3 continues by building the statistical basis for this work. Chapter 4 explains experiments and the evaluation of the methods presented and chapter 5 closes with the conclusion to this study of Dickens' style and possible future continuation.

2 APPROACHES TO STYLE ANALYSIS

In this chapter, we introduce *Stylometry*, in particular in the realm of non-traditional authorship attribution. We begin by looking at the early beginning and tentative development of statistical methods to settle cases of disputed authorship. Stylometry, although set in the general field of text classification, differs considerably in regard to its underlying assumptions, which consequently place different requirements on the overall task. The present study is concerned with Dickens' style analysis and it therefore seems appropriate to consider related approaches that focus particularly on Dickens' style.

Thus, section 2.1 recounts early studies of authorship methods, that in part still form the basis for computationally more advanced approaches today. It continues with recent state-of-the-art techniques to solve questions of authorship and concludes with examples of where authorship attribution methods can be applied, which incidentally also form part of their motivation and charm. Section 2.2 deals with the specific characteristics of authorship attribution and how these affect common methodologies in the field. Finally, section 2.3 then concentrates on studies particularly relevant to the present task of analysing Dickens' style, both from the disciplines of statistics and machine learning, but also corpus linguistics.

2.1 EXPLORING THE USE OF STYLE ANALYSIS

Stylometry is an interdisciplinary research area combining literary stylistics, statistics and computer science (He and Rasheed 2004). It is an investigation into the *style* or *feel* of a piece of writing influenced by various parameters, such as genre, topic or the author. Stylometry for authorship attribution is not concerned with deciding on the topic of a document, but rather with unearthing features distinctive of its author that can be abstracted away from its source and taken as markers that will generally apply to the author's documents regardless of their individual topics.

Discriminatory features of an author (and a particular strata of his work) have to be considered with respect to the other authors he is to be distinguished from and the quality and general appropriateness of those features is subject to the authors' document collection as well as the reference that gave rise to it.

2.1.1 *First Attempts: Characteristic Curves of Composition*

The first pioneering attempts at authorship attribution were in 1887 by the American physicist Thomas C. Mendenhall, who investigated the difference between writers, such as Charles Dickens and William Thackeray by looking at word length histograms, extending English logician Augustus de Morgan's original suggestion, that average word length could be an indicator of authorship (Mendenhall 1887).

On the basis of these word length histograms, Mendenhall constructed *characteristic curves of compositions*, that revealed *persistent peculiarities* of an author seemingly impermeable to his influence. While two curves constructed on the basis of 1000 words showed irregularities for the same author, two 100,000 words-based curves were practically identical. Even when on one occasion, an author tried to actively manipulate his own writing in an attempt to simplify it for a different audience, his curves remained strikingly alike in their main feature.

Mendenhall concluded that, in order to show that the method was sound, it would need to be applied repeatedly and to different authors, i.e. for each author, several 100,000 word length curves needed to be compared. If these were found to be practically identical for one author, while being different for two different ones, the method could be reliably applied to problems of disputed authorship (Mendenhall 1887).

In 1901, Mendenhall conducted a second study, where he attempted to settle the question of Shakespeare's authorship, in particular the question of whether Francis Bacon had been author of his plays, poems or sonnets (Mendenhall 1901). An extensive study showed that Bacon's curve was quite dissimilar to the one of Shakespeare, but that the one constructed for Christopher Marlowe agreed with the one of Shakespeare as much as Shakespeare's curves agreed with themselves.

Although word length by itself may not be considered sufficient evidence to settle the question of disputed authorship, this early study already showed the benefit of focusing on unconscious stylistic features and also conveyed the need for enough data samples to support one's claim.

2.1.2 *Disputed Authorship in the Federalist Papers*

Among related statistical studies following this early attempt was the influential work by George K. Zipf in 1932 establishing *Zipf's law* on word frequency distributions in natural language corpora, stating that the frequency of any word is inversely proportional to its rank in the frequency table (Zipf 1932).

However, there was no considerable advancement in authorship attribution studies until well into the second half of the 20th century, which marked the emergence of what was to become one of the most famous and influential studies into disputed authorship. In 1964, the two American statisticians Frederick Mosteller and David L. Wallace set out to use word frequencies to investigate the mystery of the authorship of *The Federalist Papers* (Mosteller and Wallace 2008).

During the years of 1787-1788, both Alexander Hamilton and James Madison and John Jay wrote the *Federalist* in an endeavour to persuade the citizens of New York to ratify the constitution. The question of authorship arose because originally all articles had been published under the pseudonym of "Publius" and for 12 papers both Hamilton and Madison later put in a claim. Even considering additional factors and accounts could not settle the dispute satisfactorily.

Consequently, Mosteller and Wallace conducted an extensive study as to who wrote the 12 disputed papers, which to complicate matters all had to be attributed individually. Analysis using ordinary style characteristics, such as average sentence lengths did not yield suitable variables for discrimination between the two authors, which led them to word count analysis.

The authors preliminarily concluded that one single word or a few words would not provide a satisfactory basis for reliable authorship identification, but that many words in unison were needed to create an "overwhelming" evidence, that no clue on its own would be able to provide likewise (Mosteller and Wallace 2008, p. 10).

Preliminaries: Words and Their Distributions

They embarked on the laborious task of looking at word distributions in the search of choice of words with good discrimination power. High frequency words (mostly function words) seemed to provide better discriminators, being both frequent and less subjective to contextual influence. However, even words of high frequency had relatively small rates of usage, which led the authors to search for a more fitting distribution for the Bayesian study, settling on the *Poisson* and *negative binomial* distribution. In addition, stability and independence of the word distributions over time and context was also reasonably satisfied (Watson 1966).

Bayesian Study

The main study was concerned with the estimation of the final odds (log odds), which are the product of the initial odds and the likelihood ratio. The authors employed the Bayes theorem to obtain an approximation of the prior distributions that were needed to determine conditional/posterior probabilities. Given a vector of word frequencies with density of $f_1(x)$ for Hamilton and $f_2(x)$ for Madison, the likelihood ratio is (Watson 1966):

$$\frac{f_1(x)}{f_2(x)} \; and \; prior \; probabilities: \; \pi_1, \pi_2 \; \Rightarrow \; \frac{f_1(x)\pi_1}{f_2(x)\pi_2} \; (final \; odds) \tag{2.1.1}$$

A paper could then clearly be attributed to Hamilton, if $f_1(x)\pi_1 > f_2(x)\pi_2$ and to Madison if $f_1(x)\pi_1 < f_2(x)\pi_2$. Great pains were taken in the determination of the final odds to take into consideration a range of factors, so as to minimize the effects of variation in the choice of the underlying constants of the prior distributions (Khamis 1966).

After additional analyses, the authors were able to attribute all 12 papers to Madison and for each paper $\frac{f_2(x)}{f_1(x)}$ was so large as to render any conceivable $\frac{\pi_1}{\pi_2}$ insignificant (Mosteller and Wallace 2008).

Conclusion and Critical Acclaim

At the time, Mosteller and Wallace's work marked the departure point for non-traditional authorship attribution studies, as opposed to what had been a traditional human-expert-based methods domain (Stamatatos 2009). Apart from the authors' invaluable contribution to the advancement of authorship attribution studies, they were the first to give more credibility of the application of Bayes to practical problems. Although the assumption of independence of function words is technically not correct, conditional probabilities are difficult to estimate in practise (Malyutov 2005). Their verdict of authorship in favour of Madison was supported by more recent studies, e.g. (Bosch and Smith 1998) and (Fung et al. 2003) using support vector machines.

Considering the fast pace of research nowadays and the continued importance of *Inference and Disputed Authorship: The Federalist*, it can only be regarded as a remarkable achievement overall.

2.1.3 *Recent Approaches to Authorship Attribution*

During the time post-*Federalist papers* studies and until the late 1990s, research in authorship attribution experimented and proposed a variety of methods, including sentence length, word length, word frequencies, character frequencies, and vocabulary richness functions, although methods tended to be more computer-assisted than computer-based (Stamatatos 2009). This earlier period suffered from a lack of objective evaluation methods, as most methods were tested on disputed material and evaluation was mainly heuristic and intuition-driven.

The rise of the internet and the availability of electronic texts brought authorship attribution closer to the disciplines of information retrieval, machine learning and natural language processing (NLP) and saw the development of more sophisticated evaluation techniques allowing for inter-method evaluation and the blossoming of more advanced features, such as syntax-based features. This change also enabled the field to become more relevant to criminal law, computational forensics, as well as to more traditional applications of investigating authorship as in *Federalist* case (Mosteller and Wallace 2008). However, statistical or stylistic authorship attribution of literary pieces, hitherto the domain of literary scholars, is still not a widely accepted practise among literary experts (Mahlberg 2007).

Among the common methods developed and applied to authorship attribution are *Burrows Delta* (Burrows 2002), a simple measure of the difference between two texts and principal component analysis (PCA), which is reported to provide insightful clustering in literary stylometry (Burrows 1992), but is defeated by discriminant analysis, when the authors are non-literary and have a more similar background (Baayen et al. 2002).

Neural networks, an artificial intelligence method that models human brain behaviour, is less desirable for the task of authorship attribution regardless of performance. Given appropriate training data and a test sample, a neural network returns a decision without motivation, a property insufficient for application in e.g. forensic linguistics, where humanly understandable evidence is of the essence (Clark 2011).

2.1.4 *Applications of Authorship Attribution*

Authorship attribution has a variety of potential applications, as for instance plagiarism detection, email spam writer detection or in forensics. In the following, we consider some of these applications in more detail.

authorship verification An example of authorship verification already encountered was the *Federalist papers* case. Given a piece of disputed authorship and some *suspects* and examples of their writing, the task is to verify that a given target text was or was not written by this author (Koppel et al. 2009). The problem is complicated if authorship is not limited to a small set of possible candidates.

author profiling In the case where there is an anonymous text sample, but no candidate (set) at all, making comparisons impossible, profiling is concerned with the extraction of information e.g. gender, age, native language or neuroticism levels of the author of the anonymous text (Koppel et al. 2009). Thus, lacking training data, one opts to create a psychological profile. Neurotic personalities, for instance, tend to have an increased use of reflexive pronouns and pronouns for subjects.

plagiarism detection The availability of electronic texts has also facilitated the reuse of them, which in some cases results in unauthorized reuse, more commonly known as plagiarism. There are different kinds of this infringement on original ownership, some of which are easier to detect than others. Word-for-word plagiarism is a direct copy or a minimally rewritten equivalent of a source text without acknowledgement (Clough 2003). Other types include paraphrasing by changing the wording or syntax of the source.

Automatic plagiarism detection involves measuring similarities between two documents that would be unlikely to occur by chance or finding inconsistencies in the style of an author that would indicate borrowed passages adapted in wording or syntax and quite unlike the remainder of the text (Clough 2003).

authorship analysis in forensics Forensic stylometric authorship analysis (FSAA) is the authorship attribution equivalent relevant for scientific methodology for providing evidence in a courtroom situation (Clark 2011) and also sometimes used by the police even when evidence is too non-conclusive for the courtroom. Undoubtedly due to the severe repercussions of the acceptance of evidence, FSAA as a method is subject to the legal framework for admissibility of scientific evidence under the *Daubert Standard* (Clark 2011), namely before being admitted to provide evidence, a method has to fulfil the following criteria:

1. Testability or falsifiability
2. Peer review and publication
3. Known or potential error rate
4. General acceptance

Obviously the exact error rates are more significant in a setting where conviction might partly be based on a methods' results, and it is therefore vital to state with how much confidence these may be taken into account.

Cumulative sum charts (CUSUM) were accepted in court as expert evidence, despite them being criticized severely by the research community (Stamatatos 2009), who considered the method to be unreliable. These are "simply line graphs representing the relative frequencies of supposedly 'unconscious' and 'habitual' writing traits like sentence length or words that start with vowels (Clark 2011) and thus seem comparable to the technique put forward in Mendenhall 1887.

One of the issues with most statistical methods is that they are more suited to text analysis than forensic linguistics, where data is more scarce. Linguist expert opinion on matters of authorship is also scarcely used in court, which tends to rely on individuals close to the defendant (Clark 2011).

2.2 Characteristics of Style Analysis

In the realm of text classification, authorship attribution somewhat differs from the *normal* text classification strategies. The usual objective in information retrieval is to separate a text collection into subsets according to the topics by promoting content words not frequent over the whole collection and thus more indicative of certain topics. Function words are largely ignored, since most of them do not vary considerably across topically different documents and would therefore not assist separation (Koppel et al. 2009). Here, the documents themselves are the subject of interest, while their individual authors are given less consideration.

In contrast, for the task of authorship attribution, where the object is to reveal common characteristics of an author, one collects only examples of specific authors and the documents themselves may rather be considered as observations of a random variable, namely the author's individual style.

2.2.1 *Frequent Word Features*

The benefit of using the more frequent words in a language for the task was already identified from very early on. The reasons for their popularity are that they are frequent and thus provide more reliable evidence, more independent of context and less subject to conscious control by the writer (Mosteller and Wallace 2008). Nowadays, there exists a general consensus about the merit of function words in this particular application, since it has been shown repeatedly that the frequent words (mostly function words) in a text are better suited to the task.

However, the issue is far from being irrevocably settled and the notion is still occasionally addressed, e.g. recently in Vickers 2011, where it was claimed that rarer n-grams distinguish better than common word n-grams. This in turn was challenged by David L. Hoover (Hoover 2012), who argued that since there are so many rare n-grams, there will most certainly be some unique correlation found between an anonymous sample and a candidate author sample.

The Shape of Frequent Features

For the present study, we concentrate on frequent word features and therefore describe their properties more closely in the following. High frequency words in a language mostly consist of function words and the more frequent content words, that are less dependent on context, as for instance "innovation" (Mosteller and Wallace 2008) and in research often the 500 - 4,000 most frequent words are considered. Function words are supposed to be more representative of the somewhat inherent style of the author and their discriminatory power lies in the fact that the author may be less aware of the way and rate he uses them. Function words have the further advantage of being mainly a closed class group and thus less invariant over time, unlike content words, such as verbs or nouns that can freely admit new members.

As already indicated, stylometry is concerned with identifying distinctive markers of a particular author. In order to qualify as being discriminatory for an author, these features have to display a marked difference, in regular and consistent appearance or absence, when compared to appropriate other authors' texts. Thus, discriminators can be both positive and negative, where positive discriminators are noticeable by a marked or striking appearance, generally more than mere chance occurrence would suggest, given an appropriate reference, and correspondingly negative discriminators are conspicuous by a marked absence (Tabata 2012). Generally, frequent features come in different shapes, such as character features (character n-grams), word features or syntactic/semantic features, where the choice is also application-dependent, as well as language-specific (Stamatatos 2009).

Earlier approaches to feature selection included *average number of syllables/letters per word*, *average sentence length*, but these proved mostly inadequate for the task, while morphological features might be primarily relevant for languages rich in those features (Koppel et al. 2009). Usually for analysis, one item is created for all lexical items that share the same spelling, which leads to some ambiguity of the resulting combination. Depending on the language, for example in English, this means combining some nouns and verbs (if frequent), such as $the\ water_{noun}$ and $to\ water_{verb}$.

2.2.2 *Obstacles in Style Analysis*

Given the undoubtedly challenging task of finding discriminatory markers for an author seeing that the answer is unknown and evaluation more of a *relative quality* measure, there are certain additional complications rooted inherently in language and the nature of the task. We consider a setting, where we desire to find discriminatory words for two different authors and for want of imagination, we take Charles Dickens and Wilkie Collins (see Tabata 2012).

Choosing the Parameters

The task is to find characteristic terms for both Dickens and Collins separately, where the first step is to choose appropriate training data. Unfortunately, an author, assuming he wrote over a longer period of time, is bound to develop in his style and his writings might therefore display some differences depending on when the piece was written.

Thus, the inevitable question arises, which exact text samples are most representative of the author, although this might be reasonably approximated by the application. If we choose to look at whether Dickens' style changed over time, obviously both his early and late works should be present in the set. The right method for the task depends invariably on the classification task or the specific authors compared, as was also shown in (Mosteller and Wallace 2008), where average sentence length (even though successful elsewhere) turned out to be absolutely non-discriminatory for Hamilton and Madison.

Comparing only two authors, such as Dickens and Collins, might yield discriminators, that more or less only discriminate those two writer. These features need not be discriminatory for the two authors in general and a different reference set might return quite different results (see section 2.3.3). The final *Damocles-sword* question remains: are the markers identified really discriminatory overall or only appropriate for a specific application (He and Rasheed 2004).

Facing Feature Dilemmas

General decisions that have to be considered in regard to preprocessing steps are lemmatization, which may help to overcome dialogue vs. past tense narration style variation, but causes loss of stylistic variation of endings, as for instance -"ing" - a possible indicator of movement in Dickens. Often, personal pronouns, such as *he, she, I* (not possessive ones) are also removed from the word list, as narration style tends to exert influence over pronoun frequency (first person vs. third person), but distinguishing information may also be lost through this exclusion. Taking more words as discriminators tends to lessen the effects of small errors, although large lists are only appropriate for large texts and these are not always available.

In order to capture only variation in style, other confounding factors, such as genre or time period have to be eliminated at least in principle. For this reason, one usually resorts to comparing authors from the same time period, since language and general style undergo change over time and if one aims at detecting special characteristics of a particular author, one has to compare him to contemporaries, otherwise there is the risk of detecting elements characteristic of a certain time period rather than individual authors.

Ideally, comparisons should also be on the same text type, since one author's collection of poems opposed to another author's novels might show dissimilarities that would not have arisen if the genre had been the same, as genre distinctly influences the distribution of function and content words (Burrows 2007). Poems, for instance, respond less well to frequent word analysis and a change of topic distorts middle range word frequencies.

Independence of Discriminators

In the search for characteristic markers of two authors, ideally those markers are each primarily frequent for only one of the two writers in question. In the *Federalist* study (Mosteller and Wallace 2008), two markers were identified $while - whilst$ (quasi-synonymous), that each seem to be particularly close to one of Madison or Hamilton.

However, these *clear* cases are somewhat rare, since the use of function words is not completely arbitrary and their employment is subject to a language's grammar. One may also not always find real synonymous pairs, because language in general has the tendency to suppress redundancy and this will apply even more to function words than content words, which tend to have more different word senses. The ideal one might hope for is a good approximation to terms an author uses more frequently than he would normally *need to* and those he tends to avoid more than he would be *expected to* otherwise.

Thus, one possibility, as already noted above, is to not rely on a single word or a few words for reliable authorship identification, but many words in unison to create an "overwhelming" evidence, that no clue on its own would be able to provide likewise (Mosteller and Wallace 2008, p. 10).

2.3 DICKENS' STYLE ANALYSIS

Charles Dickens is perceived to have a somewhat unique style that sets his pieces apart from his contemporary authors (Mahlberg 2007). It also renders him a good candidate for style analysis, as there are likely to be features that distinguish him from his peers. Since the present study of authorship attribution is concerned specifically with Dickens's style, this section is devoted entirely to reviewing several independent studies of Dickens' style, not all of which are statistically motivated.

In section 2.3.1, we look at a *corpus stylistics* approach, that investigates meaningful word clusters. Section 2.3.2 describes the attribution of a disputed piece as Dickensian and section 2.3.3 relates a study into Dickens' style using Random Forests and which is incidentally the main work to which we are comparing in the present study.

2.3.1 *Corpus Linguistics' Approach to Dickens' Style*

Although, we are concentrating on statistical approaches to authorship attribution, the analysis is also centred around Dickens, a literary writer, and one can therefore draw on results of other disciplines and in this way place one's own results in a better perspective.

The application of corpus methodology to the study of literary texts is known as *corpus stylistics*, which investigates the relationship between meaning and form. The study presented in Mahlberg 2007 describes a work to augment the descriptive inventory of literary stylistics by employing corpus linguistics methods to extract key word clusters (sequences of words), that can be interpreted as pointers to more general functions. The study focuses on 23 texts by Dickens in comparison to a 19th century reference corpus, containing 29 texts by various authors and thus a sample of contemporary writing.

Similar to stylometry, there also exist positive and negative key clusters for an author in the sense that they occur either more or less frequent in Dickens than would have otherwise been expected by chance in comparison with the reference corpus of the 19th century. Focusing on 5-word clusters consisting mainly of function words, 5 local functions grouping word clusters are identified.

According to Mahlberg, Dickens shows a particular affinity for using *Body Part* clusters: e.g. "his hands in his pockets", which is an example of Dickens' individualisation of his characters. Although this use in general is not unusual for the time, his rate is significant, as Dickens, for instance, links a particular bodily action to a character more than average for the 19th century. The phrase 'his hands in his pockets", for instance, occurs ninety times and in twenty texts of Dickens, compared to thirteen times and eight texts in the 19th century reference corpus.

The *Body Part* function often simply adds contextual information, that embeds another activity more central to the story, which supports ongoing characterisation that will not strike the reader as unusual:

(1) "with his hand to his chin" $\rightarrow$ **thinking**

(2) "laying his hand upon his" [shoulder] $\rightarrow$ **supporting**

Mahlberg concludes, that the identification of *Body Part* clusters provides further evidence of the importance of body language in Dickens. As already noted in (Tabata 2012), if *Body*

Part clusters are more specific to Dickens, characteristic marker terms should also include body parts.

Thus, frequent clusters can be an indication of what function (/content) words are likely to be or not be among Dickens' discriminators, in this case, we would expect there to be examples of body parts, such as *face, eyes, hands*...

2.3.2 *Attributing Dickens' "Temperance"*

Recently, the issue of unattributed articles in periodicals under Dickens' editorship has been readdressed (Craig and Drew 2011). A small article, *Temperate Temperance*, published anonymously on 18 April 1863 in the weekly magazine All the Year Round (AYR) (1859-70) was assessed using computational stylistics in combination with internal clues. Contrary to other journals under Dickens' editorship, a complete record of author attribution for the individual articles in AYR has not survived and over two-third of the AYR articles are still unidentified.

The controversy in regard to this specific piece arose due to the negative verdict for Dickens' authorship by an early Dickensian scholar, acting on external evidence, which might not be completely reliable, especially in the light of several practical reasons that indicate this article to be one of Dickens (Craig and Drew 2011).

The authors use "Burrows method" (to identify the authorial signature) to investigate authorship of *Temperate Temperance* using a control group of likely candidates contributing to the journal or collaborating with Dickens on articles at that time, one among them is Wilkie Collins. Marker words are chosen for their ability to separate the training set and are then applied to the test set and the mystery article. When compared to each other author individually, *Temperate Temperance* clustered significantly with the Dickens segments rather than with the segments of the other author. However, in order to raise a substantial claim for Dickens authorship, it was felt that Dickens needed to be compared to a larger, more representative set. Cross-validation on the data shows, that Dickens test segments generally score higher on Dickens markers from the training set (84%), than non-Dickens markers.

The authors conclude that the method was able to distinguish a general Dickens' style and and on this basis classified the disputed article with the Dickens samples, although it remains a relative measure and in theory there could be a signature more fitting than that of Dickens. Unfortunately, the discriminatory markers are not listed in the study, which renders a direct comparison of results impossible. However, the sample might be used as a test piece for the final validity check of the model.

2.3.3 *Approaching Dickens' Style through Random Forests*

In regard to a particularly relevant application in terms of comparison, we consider Tabata 2012, where Tomoji Tabata applied the machine-learning technique *Random Forests (RF)* in order to extract stylistic markers of the author Charles Dickens that would be able to distinguish his work from both Wilkie Collins and a larger reference corpus.

Random Forests (RF) is a classification algorithm based on ensemble learning from a large number of classification trees randomly generated from a dataset with the advantage of being able to handle a high number of input variables. Tabata also reports a consistent high accuracy of the technique (96-100%), when applied to distinguish Dickens from a control set. RF identifies proximities between pairs of cases and also highlights those items contributing the most for classification.

The two authors Dickens and Collins were consequently analysed using RF and clusters were visualised by a multidimensional scaling diagram. Dickens' and Collins' texts were grouped in two distinct clusters, with two more unusual pieces (*Antonina (1850)* and *Rambles beyond Railways (1851))* appearing as outliers. RF found discriminatory terms that are consistently more frequent in one author than the other and are thus stylistic markers of Dickens when compared to Collins and vice versa. Table 2.3.1 and 2.3.2 show the discriminatory terms for respectively each author.

Table 2.3.1: Dickens' markers, when compared to Collins according to Tabata's work using Random Forests.

Dickens' markers
very, many, upon, being, much, and, so, with, a, such, indeed, air, off, but, would, down, great, there, up, or, were, head, they, into, better, quite, brought, said, returned, rather, good, who, came, having, never, always, ever, replied,boy, where this, sir, well, gone, looking, dear, himself, through, should, too, together, these, like, an, how, though, then, long, going, its

Table 2.3.2: Collins' markers, when compared to Dickens according to Tabata's work using Random Forests

Collins' markers
first, words, only, end, left, moment, room, last, letter, to, enough, back, answer, leave, still, place, since, heard, answered, time, looked, person, mind, on, woman, at, told, she, own, under, just, ask, once, speak, found, passed, her, which, had, me, felt, from, asked, after, can, side, present, turned, life, next, word, new, went, say, over, while, far, london, don't, your, tell, now, before

CONTRASTING DICKENS WITH A CONTEMPORARY REFERENCE CORPUS However, in order to arrive at some stylistic features of Dickens' in a wider perspective, the second part of the study compares the 24 Dickens' texts to a larger reference corpus consisting of 24 eighteenth-century texts and 31 nineteenth-century texts (a small subset of which is from Wilkie Collins). Apart from one outlier text, *A Child's History of England (1851)*, Dickens' texts again form one distinct cluster.

Table 2.3.3 shows the Dickensian markers, the positive and the negative ones. Tabata concludes that Dickens' markers show a predominance of words related to description of actions, in particular typical bodily actions, or postures of characters and lack terms denoting abstract concepts.

Table 2.3.3: Dickens' markers, when compared to the 18th/19th century reference corpus according to Tabata's work using Random Forests

Positive Dickens' markers
eyes, hands, again, are, these, under, right, yes, up, sir, child, looked, together, here, back, it, at, am, long, quite, day, better, mean, why, turned, where, do, face, new, there, dear, people, they, door, cried, in, you, very, way, man
Negative Dickens' markers
lady, poor, less, of, things, leave, love, not, from, should, can, last, saw, now, next, my, having, began, our, letter, had, I, money, tell, such, to, nothing, person, be, would, those, far, miss, life, called, found, wish, how, must, more, herself, well, did, but, much, make, other, whose, as, own, take, go, no, gave, shall, some, against, wife, since, first, them, word

A closer look at the results

Comparing the second set of markers to the first result, one can observe that certain characteristic markers for Dickens remained the same when compared to only Collins and to the complete reference corpus, also including other authors.[1] The markers for **Dickens** appearing in **both** sets given here, include:

(3) *these, up, sir, together, long, quite, better, where, dear, they, very*

Similarly, one can observe certain terms appearing **both** in **Collins set** and **Dickens' negative set**, which may also mark them as a bit more reliable as negative markers for Dickens:

(4) *leave, from, can, last, now, next, letter, had, tell, to, person, far, life, found, own, since, first, word*

However, the fact that these terms seem to be more consistent for Dickens may also be attributed to the possibility that they are less consistent in the reference set and vice versa. In contrast, when we look at the second analysis of Dickens' markers, there are terms that were not in the first set for Dickens, but are now in the second set as well as the first for Collins, when contrasted with Dickens on his own:

(5) *under, looked, back, at, turned, new*

Those terms seemed to be discriminatory for Collins, when comparing Dickens and Collins directly, but seem to be positive for Dickens when the reference set includes a larger set. There are also a couple of terms that appeared in the first (positive) analysis for Dickens, but also in the negative set in the second analysis:

(6) *should, having, such, would, how, well, but, much*

This slight display of arbitrariness of discriminatory terms in different analysis implies that at least to a certain extent, discriminatory negative and positive markers are influenced by the opposing set of documents. Since the second analysis was conducted against a more representative set, the stylistic markers obtained there are probably more reliable.

An interesting, but in the end rather futile question is, to what extent it would be possible to determine true Dickens' markers.

[1] Since we are not given the entire list of ranked discriminators, there obviously could be more terms that follow this scheme.

3 STATISTICAL ANALYSIS OF DICKENS' TEXTS

In this chapter, we explore two different statistical methods for characteristic term extraction and subsequent building of author profiles.

However, in section 3.1 we begin by describing the different data sets that form the basis for experiments and evaluation in this work, in particular by explaining preprocessing and the weighting scheme used to construct document-by-term matrices from the data sets. Then, in section 3.2, we introduce *Independent Component Analysis* in its native environment of blind source separation and then turn to its more specific interpretation in the field of text classification and particularly authorship attribution. Section 3.3 presents *Representativeness & Distinctiveness* feature selection in the area of *dialectrometry* and continues with its application to authorship attribution. Given these two statistical methods, section 3.4 defines three different models yielding characteristic terms for subsequent evaluation. The first two models consist of respectively *Independent Component Analysis* and *Representativeness & Distinctiveness* in isolation and the third model combines the two methods into one distinct model.

3.1 AUTHORSHIP DATA SETS

For all preliminary experiments as well as evaluation, we collected or were given data sets based on documents of Charles Dickens and Wilkie Collins or a larger reference set. Generally, for experiments and cross-validation evaluation, we consider three different term-by-document matrices that are described in more detail in the following part.

Section 3.1.1 gives an overview of the Dickens/Collins set also used in another previous work (Tabata 2012). Section 3.1.2 describes our own Dickens and Collins data set that differs slightly from the previous one and section 3.1.3 then turns to the Dickens vs. 18th/19th century comparison set. With the exception of the data set in section 3.1.1, all data was prepared and preprocessed according to the description in section 3.1.4. All data was collected from the *Gutenberg project*[1].

3.1.1 *Dickens and Collins Comparison 1*

In a previous study (Tabata 2012), the same search for discriminatory markers of Dickens has been conducted, comparing Dickens to his contemporary Wilkie Collins. For the purpose of comparing to this work, we consider the same input matrix build of the document sets of Dickens and Collins shown in table 3.1.1 and table 3.1.2.[2] The document-term matrix (47×4999) contains 47 documents (23 of Dickens and 24 of Collins) and is already preprocessed and weighted, so unlike the following sets, it is not subjected to the preprocessing and weighting described in section 3.1.4. The abbreviations shown in the tables are used as identifier for the exact document and full document labels are not used any more hereafter. In the following, we refer to this set as the *"DickensCollinsSet1"*.

[1] http://www.gutenberg.org/
[2] I would like to thank Tomoji Tabata for providing the input data and the description tables shown here.

Figure 3.1.1: Dickens' documents in Tabata's Dickens/Collins comparison as part of *DickensCollinsSet1*.

No.	Texts	Abbr.	Category	Date	Word-tokens
1	*Sketches by Boz*	(D33_SB)	Sketches	1833–6	187,474
2	*The Pickwick Papers*	(D36_PP)	Serial Fiction	1836–7	298,887
3	Other Early Papers	(D37a_OEP)	Sketches	1837–40	66,939
4	*Oliver Twist*	(D37b_OT)	Serial Fiction	1837–9	156,869
5	*Nicholas Nickleby*	(D38_NN)	Serial Fiction	1838–9	321,094
6	*Master Humphrey's Clock*	(D40a_MHC)	Miscellany	1840–1	45,831
7	*The Old Curiosity Shop*	(D40b_OCS)	Serial Fiction	1840–1	217,375
8	*Barnaby Rudge*	(D41_BR)	Serial Fiction	1841	253,979
9	*American Notes*	(D42_AN)	Sketches	1842	101,623
10	*Martin Chuzzlewit*	(D43_MC)	Serial Fiction	1843–4	335,462
11	*Christmas Books*	(D43b_CB)	Fiction	1843–8	154,410
12	*Pictures from Italy*	(D46a_PFI)	Sketches	1846	72,497
13	*Dombey and Son*	(D46b_DS)	Serial Fiction	1846–8	341,947
14	*David Copperfield*	(D49_DC)	Serial Fiction	1849–50	355,714
15	*A Child's History of England*	(D51_CHE)	History	1851–3	162,883
16	*Bleak House*	(D52_BH)	Serial Fiction	1852–3	354,061
17	*Hard Times*	(D54_HT)	Serial Fiction	1854	103,263
18	*Little Dorrit*	(D55_LD)	Serial Fiction	1855–7	338,076
19	*Reprinted Pieces*	(D56_RP)	Sketches	1850–6	91,468
20	*A Tale of Two Cities*	(D59_TTC)	Serial Fiction	1859	136,031
21	*The Uncommercial Traveller*	(D60a_UT)	Sketches	1860–9	142,773
22	*The Great Expectations*	(D60b_GE)	Serial Fiction	1860–1	184,776
23	*Our Mutual Friend*	(D64_OMF)	Serial Fiction	1864–5	324,891
24	*The Mystery of Edwin Drood*	(D70_ED)	Serial Fiction	1870	94,014
		Sum of word-tokens in the set of Dickens texts:			**4,842,337**

Figure 3.1.2: Collins' documents in Tabata's Dickens/Collins comparison as part of *DickensCollinsSet1*.

No.	Texts	Abbr.	Category	Date	Word-tokens
1	*Antonina, or the Fall of Rome*	(C50_Ant(onina))	Historical	1850	166,627
2	*Rambles Beyond Railways*	(C51_RBR)	Sketches	1851	61,290
3	*Basil*	(C52_Basil)	Fiction	1852	115,235
4	*Hide and Seek*	(C54_HS)	Fiction	1854	159,048
5	*After the Dark*	(C56_AD)	Short stories	1856	136,356
6	*A Rogue's Life*	(C57_ARL)	Serial Fiction	1856–7	47,639
7	*The Queen of Hearts*	(C59_QOH)	Fiction	1869	145,350
8	*The Woman in White*	(C60_WIW)	Serial Fiction	1860	246,916
9	*No Name*	(C62_NN)	Serial Fiction	1862	264,858
10	*Armadale*	(C66_Armadale)	Serial Fiction	1866	298,135
11	*The Moonstone*	(C68_MS)	Serial Fiction	1868	196,493
12	*Man and Wife*	(C70_MW)	Fiction	1870	229,376
13	*Poor Miss Finch*	(C72_PMF)	Serial Fiction	1872	162,989
14	*The New Magdalen*	(C73_TNM)	Serial Fiction	1873	101,967
15	*The Law and the Lady*	(C75_LL)	Serial Fiction	1875	140,788
16	*The Two Destinies*	(C76_TD)	Serial Fiction	1876	89,420
17	*The Haunted Hotel*	(C78_HH)	Serial Fiction	1878	62,662
18	*The Fallen Leaves*	(C79_FL)	Serial Fiction	1879	133,047
19	*Jezebel's Daughter*	(C80_JD)	Fiction	1880	101,815
20	*The Black Robe*	(C81_BR)	Fiction	1881	107,748
21	*I Say No*	(C84_ISN)	Fiction	1884	119,626
22	*The Evil Genius*	(C86_EG)	Fiction	1886	110,618
23	*Little Novels*	(C87_LN)	Fiction	1887	148,585
24	*The Legacy of Cain*	(C89_LOC)	Fiction	1888	119,568
		Sum of word-tokens in the set of Collins texts:			**3,466,156**

3.1.2 *Dickens and Collins: Augmented*

Despite the fact that we already have a data set for comparing Dickens and Collins, we created a new set for each author, as shown in table A.1.1 and table A.1.2. Both sets are based on the ones in Tabata's study presented in section 3.1.1 and additionally include some more unusual samples. Thus, Dickens's set also contains a collaboration between Dickens and Collins *(DC1423)* and two of a set of authors *(Dal...)*. In experiments, these documents were occasionally misclassified, so in terms of stylistic analysis, these might be interesting.

For correspondence to the previous set, we list the previous labels alongside our own identifiers. The set contains 45 documents of Dickens and 29 of Wilkie Collins. Constructing a combined matrix from this set yields a 74×51244 document-term matrix with 85% sparsity, that we reduce to 74×4870 with 15% sparsity. Hereafter, this set is referred to as the *"DickensCollinsSet2"*.

3.1.3 *Dickens vs. World set*

If author-pair comparisons have one disadvantage, it might be an overemphasis of the comparison between those two authors and especially using supervised methods, this will tend to pick out discriminatory features that help separating the two sets, but which are not necessarily the most representative of the author. For this purpose, it is sensible to test Dickens against a larger reference set comprised of various contemporary authors, so as to detect terms Dickens tends to use more or less than would be considered average for his time. In order to reconstruct a similar experiment to Tabata 2012, we collected the same reference set to oppose the 24 Dickens documents used in section 3.1.1. This reference set, rather than representing a single author serves as an example of that time period and in unison would correspond to something like the average writing style of that time.

Table A.2.1 and table A.2.2 show the 18th century and 19th century components of the world reference corpus to oppose Dickens. As already indicated, single authors' identity is disregarded here and all authors are collectively indexed by a "W" (for "World") in the beginning. The reference set consists of 55 documents and Dickens set contains 24 documents. These 79 documents combined yield a 79×77499 document-term matrix with a sparsity level of 87%. We reduce this to 79×4895 and a sparsity level of 18%. In the following, we refer to this set as the *"DickensWorldSet"*.

3.1.4 *Data Collection and Preparation*

The document sets described in the previous two sections, section 3.1.2, section 3.1.3 all originated from the *Gutenberg Project*. This requires some preparation to remove *Gutenberg*-specific entries in each file, that may otherwise create noise if left in the document. Thus, prior weighting, the following items were removed from each text file.[3]

ITEMS REMOVED FROM EACH TEXT FILE

- Gutenberg header and footer
- Table of contents
- Preface/introduction written by others
- Footnotes by editor/publisher
- Notes about names/locations of illustrations
- Limited markup employed by transcribers

[3] I would like to thank Çağri Çöltekin for providing the prepared data for Dickens and Collins.

Preprocessing and Term Weighting

Before applying our models to the data, it needs to be preprocessed and weighted appropriately. All documents collected for this study are preprocessed and weighted in the same way.

PREPROCESSING Before converting the data sets to document-term matrices, we remove all punctuation, numbers and convert all words to lowercase. This removes some finer distinctions, but one would assume that if there is a significant effect of some terms in the data this would show up nevertheless.

TERM WEIGHTING All of our data collected is weighted using relative frequency of the simple term frequencies. In addition, we use *Laplace* smoothing to assign some probability to terms not observed in a document (Jurafsky and Martin 2009, p. 132). In this setting, observed frequencies are assumed to be underestimates of the theoretical corpus size. Given an observed frequency for a term t in a document d_i the new weight $w(t)$ corresponds to eq. 3.1.1.

$$w(t) = \frac{obs.\ freq.(t) + 1}{1 \times\ |word\ types| + \sum_t obs.freq.} \tag{3.1.1}$$

3.2 INDEPENDENT COMPONENT ANALYSIS FOR CHARACTERISTIC TERM SELECTION

In this section, we consider *Independent Component Analysis* (ICA) in more detail. Since it was originally developed in the field of blind source separation, we begin by introducing it on its original ground and then shift to text classification and authorship analysis. To our knowledge, ICA has not been applied to the authorship attribution problem yet, although related feature extraction method principal component analysis (PCA) has had a long established tradition in authorship studies (Burrows 1992). Despite the fact that ICA partly relies on PCA for convergence (as discussed in section 3.2.2), the two methods make very different assumptions about the structure of the underlying data distribution. For this reason, we also consider an application of PCA to one of our datasets. Section 3.2.3 offers a deeper analysis of independent components with respect to text documents and section 3.2.4 presents the general model of ICA for extracting characteristic terms of an author.

3.2.1 *Independent Component Analysis*

Independent Component Analysis first put in an appearance in 1986 at a conference on Neural Networks for Computing. In their research paper *"Space or time adaptive signal processing by neural network models"* (Herault and Jutten 1986), Jeanny Herault and Christian Jutten claimed to have found a learning algorithm that was able to blindly separate mixtures of independent signals. The concept of independent components was presented more explicitly in 1994 by Pierre Comon, who also stated additional constraints with respect to the assumed underlying probability distribution of the components (Comon 1994).

Thus, the original motivation for *Independent Component Analysis* was blind source separation, as for instance the separation of speech signals, which is commonly known as the *cocktail-party problem*. Two microphones are located in different positions in a room and two different people are speaking simultaneously. The result of these two recorded signals are the mixed signals $x_1(t)$ and $x_2(t)$, which consist of x_1 and x_2 as amplitudes, and t, the time index specifying the time of recording (Hyvärinen and Oja 2000). Each recorded signal is a weighted sum of the original speech signals of the two speakers denoted by $s_1(t)$ and $s_2(t)$. At each point in time t, $s_1(t)$ and $s_2(t)$ are assumed to be statistically independent. The maximum number of sources that can be retrieved equals the number of samples, i.e. per mixed signal one can extract one independent component. The concept can be expressed in a linear equation, as shown in eq. 3.2.1 and eq. 3.2.2.

$$x_1(t) = a_{11}s_1 + a_{12}s_2 \tag{3.2.1}$$

$$x_2(t) = a_{21}s_1 + a_{22}s_2 \tag{3.2.2}$$

with a_{11}, a_{12}, a_{21}, and a_{22} as some parameters that depend on the distances of the microphones from the speakers (Hyvärinen and Oja 2000). Given only the recorded signals $x_1(t)$ and $x_2(t)$, it would be useful to be able to estimate the two original speech signals $s_1(t)$ and $s_2(t)$ based only on the assumption of mutual independence of the source signals.

ICA Model

For want of a more general definition of the ICA model, the time index t is dropped and it is assumed that each mixture x_j as well as each independent component s_k is a random variable instead of a proper time signal. The statistical *latent variables* model is defined as follows (Hyvärinen and Oja 2000): Assume that we observe n linear mixtures $x_1, \ldots, x_n$ of correspondingly n independent components, where the observed values $x_j(t)$ are a sample of this random variable.

$$x_j = a_{j1}s_1 + a_{j2}s_2 + ... + a_{jn}s_n, \; for\ all\ j \tag{3.2.3}$$

For clarity, these sums can be converted to a vector-matrix notation (with **x** and **s** being column vectors):

$$x = As \tag{3.2.4}$$

where $x = (x_1, x_2...x_n)^T$ is a vector of observed random variables and $s = (s_1, s_2...s_n)^T$ the vector of the latent variables (the *independent components*). A is the unknown constant matrix, the $'mixing\, matrix'$ A. Both the mixture variables and the independent components are assumed to have zero mean. In order to retrieve the original sources or independent components **s**, the ICA algorithm tries to estimate the inverse W of the mixing matrix A, as in eq. 3.2.5.

$$s = Wx = A^{-1}x \tag{3.2.5}$$

AMBIGUITIES OF ICA Due to the fact that both the *mixing matrix* A and the source signals s are unknown, there are certain ambiguities related to the ICA model in eq. 3.2.4. Neither the variances (energies) of the independent components nor their order can be determined (Hyvärinen and Oja 2000). Since both A and s are unknown, the variances cannot be resolved as any multiple scalar of one of the sources s_i could be cancelled by dividing the corresponding column in A by the same scalar, so often components are assumed to have unit variance: $E\{s_i^2 = 1\}$. The ambiguity of the sign remains: a component can be multiplied by -1 without affecting the model, which is fortunately insignificant in most applications. For the same reason, i.e. A and s being unknown, the order of the components is arbitrary, since the terms in the sum in eq. 3.2.6 can be changed freely and any can be the "first" component.

$$x = \sum_{i=1}^{n} a_i s_i \tag{3.2.6}$$

ICA Algorithm

In order to estimate the independent components, ICA relies on the assumption of pairwise statistical independence between all components. Conceptually, statistical independence of two random variables y_1, y_2 implies that their joint *probability density function (pdf)* is factorisable and thus the probability of both variables occurring together equals multiplying their single probabilities.

$$p(y_1, y_2) = p(y_1)p(y_2). \tag{3.2.7}$$

Another basic assumption is non-gaussianity of the independent components and if in fact more than one component is gaussian, the mixing matrix A cannot be estimated (Hyvärinen and Oja 2000). According to the *Central Limit Theorem*, the distribution of a sum of independent random variables tends towards a gaussian distribution and thus usually has a distribution that is closer to gaussian than any of the two original random variables. Practically, non-gaussianity can be estimated by higher-order statistics, such as *kurtosis*, *negentropy* or *minimization of mutual information*.

Before applying ICA, the variables are decorrelated or whitened to help convergence using a second-order technique, such as principal component analysis or singular value decomposition (SVD) (see section 3.2.2). After the whitening of the data, ICA simply adds a rotation to achieve statistical independence. The *unmixing* matrix $W = A^{-1}$ and *mixing* matrix A can be estimated all at once (symmetric approach) or one at a time (deflation approach), where after each iteration, with W's weights usually being initialised randomly, the newly-estimated row vector (for the later creation of one component) has to be decorrelated with the previously estimated weight vectors to ensure that it does not converge to any of the previous ones.[4] The independent components are then obtained by multiplying the mixed signal matrix x by W, as shown in eq. 3.2.8.

$$\begin{aligned} s &= W \times x \\ s &= W \times A \times s, \text{ where } W = A^{-1} \\ s &= I \times s, \text{ with I = Identity matrix} \end{aligned} \tag{3.2.8}$$

ICA uses higher-order statistics and is in this respect superior to other feature extraction methods, such as principal component analysis that only remove second-order correlations (Väyrynen et al. 2007). However, ICA relies on PCA/SVD as a preprocessing step and for this reason we discuss this in more detail.

3.2.2 *Preprocessing in Independent Component Analysis*

Whitening of the data is a preprocessing step that helps ICA to converge, and if dimensionality reduction is desired, it can also be performed at this step. Both principal component analysis (PCA) and singular value decomposition (SVD) can be used to perform whitening and in the following, we describe how their respective application to a document-term

[4] Examples of ICA Implementations are: *FASTICA.*, *Infomax*, *JADE*.

matrix yields a new data representation of mutually decorrelated variables. Since text classification is the topic under discussion, we aim at defining and interpreting formulas with respect to terms and documents.

Preliminaries: Mean & Variance

For the following calculations, we need the concepts of mean over a variable x, as defined in eq. 3.2.9 and variance within one variable x_k, as defined eq. 3.2.10.

$$\mu = \frac{1}{n}\sum_{k=1}^{n} x_k, \text{ with } n \text{ being the number of samples} \tag{3.2.9}$$

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \mu)^2, \text{ with } \mu \text{ being the mean over all } n \text{ samples} \tag{3.2.10}$$

Further, we employ covariance between two different variables x and y as in eq. 3.2.11, where x_k and y_k are the k_{th} samples of two different variables with μ_x and μ_y as their respective variable means.

$$\sigma_{xy} = \frac{1}{n}\sum_{k=1}^{n} (x_k - \mu_x)(y_k - \mu_y) \tag{3.2.11}$$

The sample covariance matrix given a term-by-document matrix x is shown in matrix 3.2.12 Elements along the diagonal show variances within each term and the elements off-diagonal display the covariance between different terms. Since covariance between two variables is symmetric the elements off-diagonal are mirrored over the diagonal.

$$\sigma_x = \begin{array}{c} \\ term_1 \\ term_2 \\ \vdots \\ term_n \end{array} \begin{array}{c} \begin{array}{cccc} term_1 & term_2 & \dots & term_n \end{array} \\ \begin{pmatrix} \sigma^2_{term_1} & \sigma_{term_1,term_2} & \dots & \dots \\ \sigma_{term_1,term_2} & \sigma^2_{term_2} & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{term_n,term_1} & \dots & \dots & \sigma^2_{term_n} \end{pmatrix} \end{array} \tag{3.2.12}$$

Decorrelation of terms results in a joined covariance matrix that is diagonal, having only entries on the diagonal for the variance within a term and zero covariance between the terms (off - diagonal), as shown in matrix 3.2.13.

$$\sigma^2 = \begin{array}{c} \\ term_1 \\ term_2 \\ \vdots \\ term_n \end{array} \begin{array}{c} \begin{array}{cccc} term_1 & term_2 & \dots & term_n \end{array} \\ \begin{pmatrix} \sigma^2_{term_1} & 0 & \dots & 0 \\ 0 & \sigma^2_{term_2} & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \sigma^2_{term_n} \end{pmatrix} \end{array} \tag{3.2.13}$$

PCA ALGORITHM SVD and PCA are related provided SVD is done on mean-normalized data, meaning that before applying either technique, the matrix x has to be centred or mean-normalized by calculating the mean of each term μ_i and subtracting it for each document, as shown in matrix 3.2.14.

$$x_{cent} = \begin{array}{c} \\ term_1 \\ term_2 \\ \vdots \\ term_i \\ term_n \end{array} \begin{array}{c} \begin{array}{cccc} doc_1 & doc_2 & \dots & doc_n \end{array} \\ \begin{pmatrix} \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ x_i^1 - \mu_i & x_i^2 - \mu_i & \dots & x_i^n - \mu_i \\ \dots & \dots & \dots & \dots \end{pmatrix} \end{array} \tag{3.2.14}$$

Classic principal component analysis is performed via eigenvalue-eigenvector decomposition (EVD). An *eigenvector* is a non-zero vector $\vec{v}$ that satisfies $A\vec{v} = \lambda\vec{v}$, where A is a square matrix and λ a scalar and also the *eigenvalue* (Baker 2005). Eigenvectors are vectors of a matrix that are projected on a multiple of themselves without changing direction. The eigenvalue belonging to an eigenvector affects the projection and is also a measure of the vector's magnitude. For PCA, the eigenvectors and eigenvalues are used to evaluate the principal directions and dynamics of the data. Eigenvectors and eigenvalues are extracted from the covariance matrix of the term-by-document matrix. In the present case, where the data is already centred, the formulas for variance and covariance reduce to eq. 3.2.15 and eq. 3.2.16 respectively.

$$\sigma^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k)^2 \tag{3.2.15}$$

$$\sigma_{xy} = \frac{1}{n} \sum_{k=1}^{n} (x_k)(y_k) \tag{3.2.16}$$

The next step is the decomposition of covariance matrix σ_x, which is now a square $k \times k$ matrix, we call A, with $A \equiv xx^T$. A can be decomposed into $A = EDE^T$, where D is a diagonal matrix containing the eigenvalues of A, and E being the matrix of eigenvectors arranged as columns (Shlens 2005).

Matrix Z, the whitening matrix in eq. 3.2.17, should be ordered according to the largest eigenvalues. The largest eigenvalue corresponds to the vector along which direction the data set has maximum variance and thus for dimensionality reduction, one can discard those eigenvectors with the correspondingly lowest eigenvalues. As a last step, we need to project the data matrix x along these new dimensions to obtain the decorrelated new representation or the whitened matrix x̃, as shown in eq. 3.2.18.

$$Z = D^{-\frac{1}{2}} E^T \tag{3.2.17}$$

$$\tilde{x} = Zx \tag{3.2.18}$$

SVD ALGORITHM Singular value decomposition is a more stable solution to obtain the eigenvectors and can be performed on any matrix, be it square, non-square, non-singular or even singular, which also makes it a more powerful decomposition technique. The decomposition of a matrix A (*document x term* matrix x) is defined in eq. 3.2.19 (Shlens 2005). The matrix A decomposes into USV^T, where S is the diagonal matrix of singular values of $n \times m$ matrix A, U contains the left singular orthogonal eigenvectors and V contains the right singular orthogonal eigenvectors.

$$
\begin{aligned}
A_{mn} &= U_{mm} S_{mn} V_{nn}^T \\
U &= A \times A^T \\
V &= A^T \times A
\end{aligned} \tag{3.2.19}
$$

The connection to the previous eigendecomposition is by multiplying the matrix A by A^T as shown in eq. 3.2.20, where the columns of U contain the eigenvectors AA^T and the eigenvalues of AA^T are the squares of S.

$$
\begin{aligned}
AA^T &= USV^T \mid \times A^T \\
&= USV^T (USV^T)^T \\
&= USV^T (VSU^T), \text{ where } V^T V = I \text{ (Identity matrix)} \\
&= US^2 U^T
\end{aligned} \tag{3.2.20}
$$

Dimensionality can be reduced by selecting the largest values in S and their corresponding values in V. Similar to before, the whitening matrix Z for x is retrieved as in eq. 3.2.21 and the whitened matrix $\tilde{x}$ in eq. 3.2.22.

$$Z = S^{-\frac{1}{2}} V^T \tag{3.2.21}$$

$$\tilde{x} = Zx \tag{3.2.22}$$

In conclusion, finding the principal components amounts to finding an orthonormal basis that spans the column space of the data matrix A (Shlens 2005). Singular value decomposition is a more powerful method of deriving the required values, as eigenvectors and eigenvalues are directly and accurately estimated from A, instead of extracting them from the covariance matrix.

Differences PCA and ICA

Principal Component Analysis and *Independent Component Analysis* are deeply related, as became already apparent through the fact that ICA relies on PCA for preprocessing. However, PCA and ICA make opposite assumptions about the underlying data distribution of their to-be-retrieved components. PCA assumes a gaussian distribution and uses the measures of μ and σ^2 to estimate the new directions of maximal variance in the data. Thus, the method is only able to remove second-order correlations, whereas ICA resorts to higher-order statistics, such as *negentropy* or *kurtosis* to achieve statistical independence (Väyrynen et al. 2007). While *statistical independence* implies *uncorrelatedness*, the reverse condition is not necessarily true: *uncorrelatedness* does not imply *statistical independence* (Hyvärinen and Oja 2000). Another difference concerns orthogonality of the components, which is a necessary condition for PCA, but not for ICA, where components can be orthogonal, but do not need to be.

However, although superior in some respects, *Independent Component Analysis* is not always the better choice, especially when a gaussian distribution assumption is more suitable for the data, as for instance in (Baek et al. 2002), where PCA outperformed ICA on a face recognition task. Since PCA is the "simpler" of the two techniques, one should ensure that PCA is not suited to the task, before applying a computationally more expensive algorithm, such as ICA.

APPLYING PCA TO AUTHORSHIP DATA For the purpose of testing *Principal Component Analysis* on our data, we consider an example application to a two-author dataset. The input data is a 74×4870 document-by-term matrix, where are 45 documents by Dickens and 29 by Collins, weighted with relative frequencies using *Laplace* smoothing, as described in more detail in section 3.1). The principal components are computed with pre-centering of the data.[5] The results provide information about each component-proportion of variance ratio, i.e. to what extent a component explains the variance in the data as well as the partitioning of terms into the new components, i.e. which original features are joined into new feature combinations.

Table 3.2.1 shows the proportion of variance of the first six principal components representing the new decorrelated features. For dimensionality reduction, one usually aims at retaining about 70% of the variance. The first two principal components pc1 and pc2 account for about 60% of the variance, while the remainder is spread out over the other 72 components. In this case, choosing pc1 to pc4 would account for about 70%, although pc1 and pc2's contribution to explanation of variance is far more substantial than the other two components.

Table 3.2.1: Proportion of variance of first principal components when applied to Dickens-Collins dataset.

Principal component no.	pc1	pc2	pc3	pc4	pc5	pc6	pc ...
Proportion of variance	0.32	0.29	0.06	0.05	0.05	0.03	...

Table 3.2.2 and table 3.2.3 show the highest positively and negatively associated terms for the first two components respectively. If a term is positive for pc1, such as *and*, *but* and *that*, it means that if a document is positively associated with that component those terms, are also positive for it. Conversely, if there is a negative association between a component and a term, e.g. *her* or *she* for pc1, these are also negative for a positively associated document. Generally, there appears to be a complementary distribution for the terms and the components, i.e. if a term is positively linked to pc1, it has a negative association with pc2, however a term can also be associated with the same sign and two different components, such as the term *the*, which is linked negatively with both first principal components. Considering the type of terms with a high weight, one can observe that these are almost exclusively function words and also seem to somewhat correspond to the terms with the highest relative frequency in the input *document* $\times$ *term* matrix. There are only few content words or verbs among the highest associated terms.

Table 3.2.2: Highest negatively and positively associated terms for principal component 1.

Term	and	but	that	upon	very	the	her	she	you
Weight in pc1	0.57	0.11	0.09	0.08	0.67	-0.22	-0.19	-0.19	-0.09

Table 3.2.3: Highest negatively and positively associated terms for principal component 2.

Term	you	her	she	said	what	the	and	their	they
Weight in pc2	0.33	0.31	0.25	0.09	0.08	-0.650	-0.420	-0.083	-0.070

Figure 3.2.1 shows the new projection of the documents onto the first two principal components, listing the highest associated terms for each component at each axis. The

[5] The principal components are computed in R: *prcomp(document-term_matrix, center= TRUE)*, which uses singular value decomposition for estimation.

document sets of the two authors do intersect to some extent and fail to form distinct clusters or associate clearly with a negative or positive part of a component. Figure 3.2.2 shows the same projection of the documents onto the components with additionally indicating the term projections onto the components. Most terms are hidden in the cloud in the middle, since their connection with both components is rather low and thus their association with documents strongly linked to a component is also low.

From this example, we conclude that although this experiment is no guarantee for successful application of ICA to the data, it is at least worthwhile investigating whether the higher-order method is able to capture more interesting latent variables indicative of more conclusive links for authorship analysis.

Figure 3.2.1: Projection of the Dickens (D)/Collins (C) documents on first 2 principal components showing the most positive(+) and most negative(-) terms on each axis.

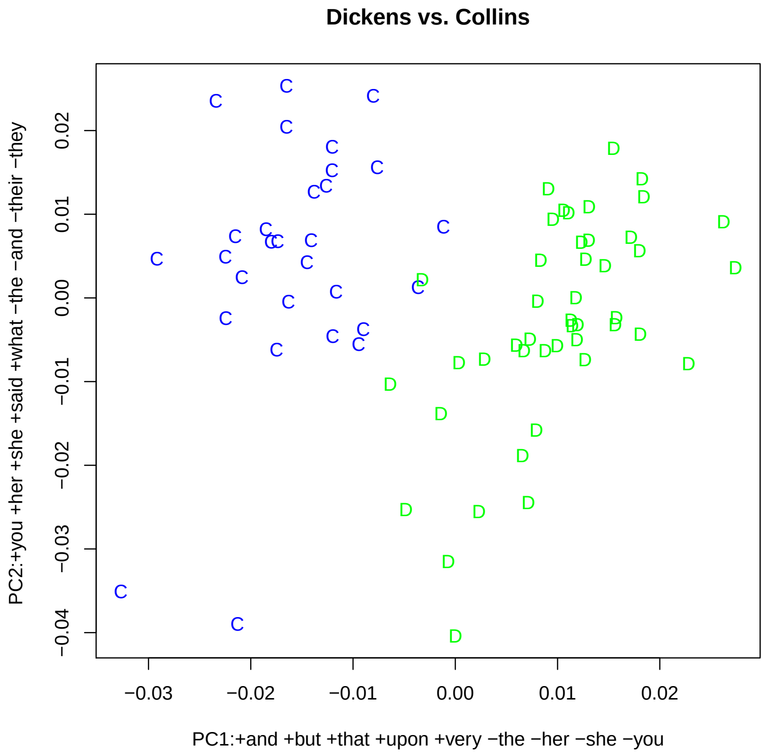
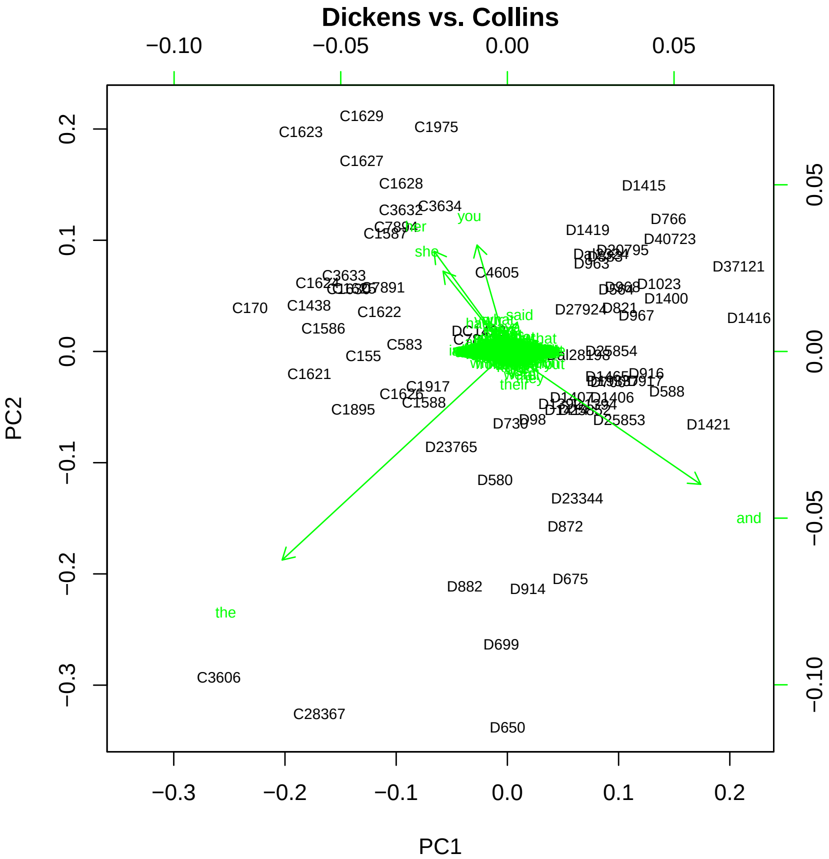


Figure 3.2.2: Projection of the Dickens/Collins documents on first 2 principal components with arrows showing term projections onto components.



3.2.3 *Independent Component Analysis in Text Classification*

Independent component analysis has been applied to text classification on numerous occasions, such as in Honkela and Hyvärinen 2004, where ICA was applied to a term-context matrix with the result of ICA identifying components relating to distinct syntactic concepts, such as adjectives or nouns. Using a more restricted context allowed for more detailed and condensed components.

For the purpose of authorship attribution, we consider term-by-document matrices of two joined author sets. Given the ICA model: $\mathbf{x} = \mathbf{As}$, where the input matrix is assumed to separate into independent components $\mathbf{s}$ and mixing matrix $\mathbf{A}$, with neither A nor s known, there are certain ambiguities related to the output components that have to be interpreted in relation to the input features.

Component Interpretation

Given an input matrix, such as a term-by-document matrix, using ICA feature extraction, there are potentially two dimensions of separation, i.e. one can try to separate both:

1. terms into latent concepts (term-by-document input)

$$X_{term \times document} = A_{term \times concept} \times S_{concept \times document} \tag{3.2.23}$$

2. documents (document-by-term input)

$$X_{document \times term} = A_{document \times concept} \times S_{concept \times term} \tag{3.2.24}$$

In the case where we are attempting to separate authors, both directions are possible, although they may differ in results. The interpretation of the first is, that documents are mixtures of latent concepts grouping terms and conversely for the second approach, terms are mixtures of latent concepts grouping documents. The first option emphasizes a hierarchical structure, whereby $terms \subset concepts \subset documents$. Given a set of term-document relations, we are looking for a new data representation that groups terms into latent concepts and assigns a weight of those concepts in the documents. Consequently, the importance of the $term_j$ with respect to a $document_i$ is reflected by the weight of all $concept_c$ in $document_i$ (with $term_j \in concept_c$) and the weight of $term_j$ in $concept_c$.

Using a document-by-term input to ICA focuses on estimating the weights of latent concepts in terms, and documents then encode the concepts. Overall, it seems that the two approaches differ slightly in regard to their implications. For this study, we concentrate on the first approach of using a term-by-document input.

Looking for Characteristic Deviations

The interpretation of the output matrices A and s is dependent on the weighting of the original input features. [6] We take a term-by-document matrix x (Figure 3.2.3) with relative frequency weighting and center the values are as part of the preprocessing, which leaves only the standard deviation of the relative frequency for each term in each document. Given the output matrices, $A_{term \times concept}$ (Figure 3.2.4) and $S_{concept \times document}$ (Figure 3.2.5), the mixing matrix A encodes a set of concepts consisting of characteristic joint deviations from the mean frequencies. $S_{concept \times document}$ then assigns a measure of how relevant a concept is given a particular document.

Figure 3.2.3: Term-document matrix x, the input to the ICA algorithm.

$$x = \begin{array}{l} \\ able \\ about \\ above \\ abroad \\ absence \\ absolutely \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} D1023 & D1392 & D1394 & D1400 & D1406 & \dots \end{array} \\ \begin{pmatrix} 0.0004 & 0.0004 & 0.0002 & 0.0002 & 0.0005 & \dots \\ 0.0032 & 0.0015 & 0.0034 & 0.0002 & 0.0003 & \dots \\ 0.0002 & 0.0004 & 0.0005 & 0.0002 & 0.0004 & \dots \\ 0.0001 & 0.0003 & 0.0001 & 0.0000 & 0.0001 & \dots \\ 0.0001 & 0.0001 & 0.0001 & 0.0001 & 0.0001 & \dots \\ 0.0000 & 0.0001 & 0.0001 & 0.0001 & 0.0001 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

Weights can be both positive and negative and have to be interpreted depending on the polarity of the input weight. In the case, that the input is positive, a positive weight for $term_j$ in $concept_c$ should be interpreted as a positive association. Generally, higher weights regardless of sign indicate more relevance. In the present example, where the input is positive, e.g. a high negative weight for a component in a document indicates increased

[6] I would like to thank Jason Palmer from the University of California, San Diego for his insightful explanations of ICA with respect to weight interpretation.

Figure 3.2.4: Term-concept matrix A, the *mixing* matrix returned by ICA.

$$A = \begin{array}{l} \\ able \\ about \\ above \\ abroad \\ absence \\ absolutely \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} c1 & c2 & c3 & c4 & c5 & \dots \end{array} \\ \begin{pmatrix} 0.0761 & -0.0146 & -0.1073 & -0.0915 & -0.0712 & \dots \\ 0.1689 & -0.0131 & -0.0494 & -0.0454 & -0.0676 & \dots \\ -0.0878 & -0.0259 & 0.0522 & -0.0023 & -0.0181 & \dots \\ -0.1415 & 0.0204 & 0.0775 & -0.0401 & 0.0003 & \dots \\ -0.0024 & -0.1343 & 0.0498 & 0.1406 & -0.1017 & \dots \\ 0.0028 & -0.0041 & -0.0183 & 0.0477 & -0.0910 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

Figure 3.2.5: Concept-document matrix S, the *source* matrix returned by ICA.

$$s = \begin{array}{l} \\ c1 \\ c2 \\ c3 \\ c4 \\ c5 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} D1023 & D1392 & D1394 & D1400 & D1406 & \dots \end{array} \\ \begin{pmatrix} 1.0000 & -1.0000 & 1.0000 & 1.0000 & 1.0000 & \dots \\ -1.0688 & -1.0269 & 0.9187 & -1.0688 & -1.0688 & \dots \\ -1.0007 & -0.9531 & 0.9558 & 1.0025 & -1.0006 & \dots \\ -1.0386 & -0.8975 & 0.8958 & -1.1518 & 0.9906 & \dots \\ -0.9303 & 0.9171 & -0.9577 & 1.1641 & 0.1081 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array}$$

negative correlation. Values near zero whether positively or negatively signed hint at overall insignificance. Since, the overall aim is to extract discriminatory or characteristic terms for each author also the highly negatively weighted terms for each author could be used for analysis. Consequently, based on the $A_{term \times concept}$ matrix, for each concept, we build two term lists: one for positive keywords and one for negative ones.

There are two possible ways a term is listed among the positive(/negative) keywords for a given document. The first and principal way is if $term_j$ has a positive weight in $concept_c$, which again has a positive weight in $document_i$. For illustration, we can look at the term *able* in the A matrix above, that is weighted with 0.076 for concept c_1. In matrix s, c_1 itself is weighted with 1.00 for document *D1023*. In this setting, this is an entirely positive association for *able*. The second possibility might be through negative association. Again, looking at the term *able* which has a negative association (-0.11) with concept c_3. When considering matrix s and the same document as before (*D1023*), c_3 has a high negative association (-1.00). Since *able* has a negative association with c_3, that has a negative association with *D1023*, this should somehow also positively contribute to its overall weight in *D1023*.

Mixed cases are those where a term has a negative weight in a concept that itself has a positive weight in a document: $-term \in (+concept \in document)$ or the reverse: $+term \in (-concept \in document)$, which always results in a negative association for the term in a document overall.

In-depth Component Interpretation

In order to understand the *concepts* formed out of terms by independent component analysis, we explore the distributions of terms within components. For this purpose, we consider a 74×1500 document-by-term matrix, containing 45 documents by Charles Dickens and 29 by Wilkie Collins and weighted using relative frequency (this is described in more detail in section 3.1).

On this basis, we computed and extracted 73 independent components, which all have different weights for each document. For analysis, we retain those components for a document with a weight above their own average weight over all documents. Similarly, we retain those terms for a component that also lie above their average weight for that specific component (These thresholds are explained in more detail in the next section 3.2.4).

There do not appear to be "author-exclusive" components, i.e. components only strong for one specific author, so the differences appear to be more subtle. After having thus discarded of components less important for each individual document, we search for components that seem to predominate for an author's documents. We identify two components, component three appears in a large number of Dickens' documents and component 20 is frequent for Collins' documents.

Table 3.2.4 shows the 30 highest negatively and positively associated terms for those two components ordered according to importance. A concept is formed through the positive terms in the component, e.g. in component 3: *wrong*, *ignorant trial*, *doubt*, and a marked absence of the negative terms, e.g. *over*, *empty* and *wall*. Although, an obvious concept may fail to be found here, component 3 contains some terms relating to a court situation, e.g. *trial*, *justice*, *written*, *pen* and possibly associated emotions, such as *doubt*, *weakness*, *confidence*, *promise* and *hesitation*. The negative terms for this component are more prosaic, such as *wall*, *sofa* and *gentlemen* This component ranks high in documents, such as *David Copperfield*, *Great Expectations* and *The Perils of English Prisoners*.

Component 20 contains some terms related to possibly the theatre and travelling: *stage*, *play*, *parts*, *post*, *town* and *country* and a marked absence of emotions and senses, such as *distrust*, *afraid* and *look*, or *heard*. The component does not rank very high in Collins' documents, but features in a large number of them, it is, however, more prevalent for *Man and Wife* and *After Dark*.

We tentatively conclude, that there do not seem to be exclusive components for authors, but that their weight in a specific document is rather sensitive to the topic within. The answer with respect to characteristic terms for an author does therefore not seem to have a straightforward answer, but may be approximated through a more cumulative effect of collecting the positive and negative terms in important components over an author's documents in the hope that characteristic ones will also show a high overall effect. Thus, we continue by discussing the thresholding methods to discard less important components for documents and the less important terms for concepts, as well as the final weight combination of terms in documents and terms for author profiles.

3.2.4 *ICA General Model*

Since every term in a component and every component in document receives a weight even if it is very close to zero, some pre-selection may be appropriate to select those terms and components more relevant for separation.

TERM-COMPONENT THRESHOLD For this purpose, we use an unsupervised approach to term selection and define individual term thresholds based on all weights of a term over all components. In this way, we capture the mean activity for a term and on this basis can decide in which components it is more active than others. This activity may be positive or negative and in order to ascertain the mean activity of a term we take absolute values of term weights x_j in all n components to compute the mean for the term, which then becomes the threshold for that term δ_{term} (eq. 3.2.25). Thus, if the absolute weight $|w_{ij}|$ for

Table 3.2.4: 30 highest negatively and positively associated terms for two components, C3 and C20. Each one being important for separately Dickens or Collins by appearing a number of their respective documents.

Terms	*Predominant Dickens Component: C3*		*Predominant Collins Component: C20*	
1	wrong	-over	two	-look
2	ignorant	-empty	stage	-distrust
3	page	-lips	description	-wonder
4	certainly	-wall	forth	-answered
5	besides	-previous	fine	-loss
6	easily	-sofa	post	-daughter
7	written	-shoes	george	-sorry
8	too	-does	months	-face
9	useful	-gentlemen	round	-remember
10	months	-strength	order	-forgive
11	remembrance	-wind	november	-expression
12	trial	-beyond	play	-yes
13	doubt	-period	parts	-upstairs
14	weakness	-light	town	-person
15	cross	-man	forward	-try
16	promise	-bed	scene	-how
17	friend	-nephew	large	-thought
18	confidence	-uncle	appears	-heard
19	justice	-outside	north	-poor
20	pen	-inn	party	-ask
22	happiness	-looks	off	-sense
23	help	-dark	country	-perhaps
24	hesitation	-mouth	year	-innocent
25	beginning	-sir	monday	-remarked
26	fear	-society	three	-suggested
27	betrayed	-without	forty	-returned
28	caused	-eyes	morrow	-forget
29	distrust	-top	twelve	-creature
30	need	-rooms	land	-afraid

term t_j in component c_i is below its individual threshold δ_{term}, term t_j does not belong to the set of prevalent terms for component c_i.

$$\delta_{term} = \frac{1}{n} \sum_{j=1}^{n} |x_j| \tag{3.2.25}$$

COMPONENT-DOCUMENT WEIGHTING Similarly, in order to select discriminatory components for each document, we define individual thresholds for each component, since the overall activity of components may differ. For each component $comp$, we compute its activity threshold δ_{comp} from its mean activity over all n documents. As usual, we are interested in both highly relevant positive and negative components, so we take absolute values of all weights of the component in all document to compute δ_{comp} (eq. 3.2.26). If then the absolute weight $|w_{ij}|$ for component c_i in document d_j is below its δ_{comp}, component c_i does not belong to the set of prevalent components for document d_j.

$$\delta_{comp} = \frac{1}{n} \sum_{j=1}^{n} |c_j| \tag{3.2.26}$$

Another option is the selection of components according to class labels, as for instance with the *Representativeness & Distinctiveness* feature selection method (see section 3.3). Choosing appropriate components is then on the level of the complete author set, i.e. we retain components more consistent in the set and discard those more inconsistent with respect to the complete set. Employing *Representativeness & Distinctiveness* feature selection would involve choosing components with similar weights over all of, for instance Dickens' documents, but which are at the same time less consistent or very different in weight over all of Collins' document set. The representative and distinctive components are usually chosen by taking the mean over all corresponding values, although this is also dependent on the number of original components.

Instead of having an individual component set for each document, as in the simple average threshold, all documents of one author have the same set of components with only the respective weight being different for each single document. Thus, there is the component set cs_1 for the author-document set ds_1 and component set cs_2 for author-document set ds_2. Both methods of using a threshold or selecting components for documents on the basis of class result in a selected set of components for each document in an authors' set.

weight combination into term-document representations In order to obtain a single term weight for a document from the above separated representations, all weights for one term are joined to form an overall weight in a document. Each term may feature in more than one component, which is the reason why we collect all weights for a term if they are above its δ_{term}.

We start from the matrix $S_{concept \times document}$ and consider all components that were retained either through thresholding or specific component selection by taking all components cs_1 for author 1 ($/cs_2$ for author 2). Important is the weight for the component c_i in a document d_j, which we denote as α_{ij}. We now iterate over all components c_i and collect all terms t_k for each component in the set, whose weight is above their individual term threshold δ_{term} for that component. This we denote as β_{ik} referring to the the weight of a term t_k in a component c_i.

The final weight w_{t_k,d_j} of term t_k in document d_j is then computed by taking the weight of the corresponding component, α_{ij}, and multiplying it by β_{ik} the weight of the term in the component. For each term t_k for each document d_j, we iterate over its n remaining components after pre-selection and sum over the weights (eq. 3.2.27).

$$w_{t_k,d_j} = \sum_{i=1}^{n} \alpha_{ij} * \beta_{ik} \tag{3.2.27}$$

Thus, if the term appears in more than one component for document j with a value above its threshold δ_{term}, the corresponding weights are added. According to the definitions above, a positive α_{ij} and a positive β_{ik} give a positive weight for the term in the document. A negative α_{ij} and a negative β_{ik} also give a positive weight. Any combination of negative and positive α_{ij} and β_{ik} returns a negative overall weight for that term in a document.

from term-document weights to author profiles For the final abstraction from document-term collections of one author to a list of positive and negative discriminators, we can take the mean value of a term t over all of the author's documents n, as its representative profile P weight (eq. 3.2.28) as the representative value for the term in the authors' document set and select the highest absolute terms as positive and negative discriminators.

$$P(t) = \frac{1}{n} \sum_{j=1}^{n} x_j \tag{3.2.28}$$

The combination of individual term-document weights into a weight for an author profile is another approximation, since a term is sometimes positive for a document of an author and sometimes negative. By taking taking the average over all documents, we take into account all information, Since this is rather an exploratory study than presenting a final and ideal solution to weight thresholding and combination, this aspect is not discussed here any further and left for future work.

The final terms for the author profiles are chosen by computing a threshold $\delta_{profile}$ computed from the mean and standard deviation over all terms in the profile and adding them, as shown in eq. 3.2.29. By including the standard deviation, we counterbalance the effect of very small weights on the mean. Additionally, we multiply by some scalar α depending on the number of input terms.

$$\delta_{profile} = (\overline{terms} + sd(terms)) \times \alpha \tag{3.2.29}$$

INTERSECTING AUTHOR PROFILES Having completed the arduous task of combination of term-concept and concept-document weights into the final combination into two distinct author profiles, there are some curious properties to be observed. Using the same example input as earlier in this section, we derived two author profiles from 45 documents by Dickens and 29 by Collins. For the final profiles, we only retain the highest weighted terms for each profile. Dickens' profile contains 137 terms and Collins' profile 139 terms. If we intersect the two profiles in the search for common terms of both authors and consider the corresponding weight each of the common term has in each profile in table 3.2.4, we can observe, that for the terms both authors share, if Dickens has a positive weight for a term, Collins has a negative weight for the same term and vice versa. This is surprising, since all terms for components and components for documents were chosen in an unsupervised fashion with no relation to the class labels. Thus ICA might be in fact detecting properties relating to authorship through cumulative effect over terms in concept and concepts in documents.

However, with respect to this example, there is of course a bias, since only the two authors are compared and the method would strive to detect characteristic deviations. If Dickens and Collins do not differ substantially for the usage of some terms this is also not likely to reflect in components. Theoretically, Dickens and Collins could share unusual properties with respect to usage of terms, and even if these might be rendered a little less discriminatory through this sharing, they should not be regarded as completely invalid, since they might still be contributing to overall discrimination for the author.

Table 3.2.5: 32 common Terms of Dickens' profile and Collins' profile showing opposite signed weights yielded by ICA analysis.

Common Term	*Weight in Collins Profile*	*Weight in Dickens' Profile*
enough	0.64	-0.53
words	0.67	-0.49
feel	0.61	-0.46
produced	0.64	-0.48
later	0.58	-0.61
discovered	0.61	-0.57
wait	0.61	-0.67
met	0.68	-0.63
since	0.63	-0.47
interval	0.51	-0.45
advice	0.58	-0.49
speak	0.66	-0.51
motive	0.54	-0.52
answered	0.67	-0.43
meet	0.53	-0.43
absence	0.58	-0.52
speaking	0.71	-0.44
heard	0.56	-0.57
asked	0.56	-0.48
though	-0.74	0.47
heaven	-0.56	0.44
deal	-0.69	0.50
always	-0.60	0.45
down	-0.85	0.51
person	0.54	-0.49
such	-0.57	0.53
off	-0.60	0.46
upon	-0.84	0.57
many	-0.71	0.48
great	-0.52	0.50
much	-0.65	0.63
head	-0.61	0.46
being	-0.69	0.45

3.3 REPRESENTATIVENESS AND DISTINCTIVENESS

In this section, we introduce *Representativeness & Distinctiveness* by first considering its original application in the field of dialectrometry and then interpreting its application to authorship attribution and the purpose of building author profiles.

Section 3.3.1 contains the general introduction to representative and distinctive features and explains its application to dialect data. In section 3.3.2, we transfer to representative and distinctive terms for authorship attribution and section 3.3.3 describes the general model that is used in this work and the motivation for concentrating on certain representative and distinctive features for different types of evaluation.

3.3.1 *Representativeness and Distinctiveness for Dialectrometry*

Representativeness & Distinctiveness (Prokić et al. 2012) was originally applied in the realm of dialectrometry, a study of dialect differences between different sites within a language area with respect to a choice of lexical items. The degree of difference between two sites is characterised by the aggregate differences of comparisons of all lexical items collected at each site. In the context of dialectrometry, *Representativeness & Distinctiveness* is a measure to detect characteristic features (lexical items), that differ little within a group of sites and considerably more outside that group. Characteristic features are chosen with respect to one group g of sites $|g|$ within a larger group of interest G, where $|G|$ includes the sites s both within and outside g (Prokić et al. 2012).

Representativeness

The degree of *Representativeness* of feature f of the group g is then defined as the mean difference of all site comparisons $d_f(s, s')$ (using an appropriate distance function for the comparison between two sites. Consequently, $\overline{d_f^g} \to 0$, as the values of the features approach a constant value for all $s \in g$ as defined in 3.3.1.

$$\overline{d_f^g} = \frac{2}{|g|^2 - |g|} \sum_{s,s' \in g, s \neq s'} d_f(s, s') \tag{3.3.1}$$

Thus, for each feature, the *Representativeness* measure compares all values within the group $|g|$ and collects the pairwise differences, which are then normalized by the number of comparisons. In this way, the less the value for that feature varies within the group, the smaller $\overline{d_f^g}$ becomes, which indicates that the feature is more representative of the whole group.

Distinctiveness

Similarly, *Distinctiveness* of a feature measures the mean difference between the group and elements outside the group. $\overline{d_f^g} \to \infty$, $s \in g, s' \notin g$ as the feature f become more distinctive for group g, as defined in 3.3.2.

$$\overline{d_f^{g'}} = \frac{1}{|g|(|G| - |g|)} \sum_{s \in g, s' \notin g} d_f(s, s') \tag{3.3.2}$$

The comparison is performed for each feature with respect to the elements outside the group $|g|$, but within the larger group of interest $|G|$. For each feature, the values of that feature within $|g|$ are compared to those outside. In contrast to *Representativeness*, if the values are ranging greatly, the feature is more distinct or different for both groups. For *Distinctiveness*, we prefer features that have very different values in each of the two sets. Characteristic features are those with relatively large differences between $\overline{d_f^{g'}}$ and $\overline{d_f^g}$. To overcome comparability difficulties in regard to missing features or different distributions, $\overline{d_f^g}$ and $\overline{d_f^{g'}}$ are standardized and compared based on these z-scores. Standardization is calculated for every feature f separately, with d_f referring to all accumulated distance values with respect to feature f (eq. 3.3.3).

$$\frac{\overline{d_f^{g'}} - \overline{d_f}}{sd(d_f)} - \frac{\overline{d_f^g} - \overline{d_f}}{sd(d_f)} \tag{3.3.3}$$

3.3.2 *Representative & Distinctive Terms for Authorship Attribution*

In the following, we interpret *Representativeness & Distinctiveness (RD)* for detection of characteristic features of an author, given some of his document samples and samples by a different source. The group D (g) comprises all of his documents d (*the sites s*) and DS is the union of all documents $d \in D$ and the documents by other authors. The distance function in this case is the absolute difference between the logarithm of the relative frequencies of f with respect to two documents d and d'. The usual input are relative frequencies of the original term frequency weighting, which provide a better picture between the ratio of term frequency and document size. The logarithm lessens the effect of rather high frequencies.

Thus, the distance d_f between document d and d' with respect to feature f, is set as the absolute difference between the logarithm of the relative frequency of their respective input values (eq. 3.3.4)

$$d_f(d,d') = |log(relFreq(f) - log(relFreq(f')| \tag{3.3.4}$$

Representativeness of a feature f for document set D is then defined in eq. 3.3.5

$$\overline{d_f^D} = \frac{2}{|D|^2 - |D|} \sum_{d,d' \in D, d \neq d'} d_f(d,d') \tag{3.3.5}$$

The *Distinctiveness* measure for comparing to outside documents corresponds to eq. 3.3.6

$$\overline{d_f^{D'}} = \frac{1}{|D|(|DS| - |D|)} \sum_{d \in D, d' \notin D} d_f(d,d') \tag{3.3.6}$$

$\overline{d_f^{D'}}$ and $\overline{d_f^D}$ are standardized by using all distance values calculated for feature f to yield the degree of representativeness and distinctiveness for term dt in D with respect to DS as defined in eq. 3.3.7.

$$dt = \frac{\overline{d_f^{D'}} - \overline{d_f}}{sd(d_f)} - \frac{\overline{d_f^D} - \overline{d_f}}{sd(d_f)} \tag{3.3.7}$$

Having performed this process for all features yields an ordered dt list, where the highest values are the most representative and distinctive and thus desirable for separating the two sets.

3.3.3 *The Representativeness-Distinctiveness' General Model*

Given that the above process has been performed, the highest terms of those lists to be included in the respective author profile still have to be chosen. For selecting the highest rated features of the standardized features, we define threshold δ_{dt} as α times the mean over all characteristic features plus their standard deviation, as in eq. 3.3.8. Depending on the number of input features, the profiles tend to admit more terms than a simple mean threshold could restrict. Additionally, the mean is also lowered considerably if less representative and distinctive items are admitted. Adding the standard deviation should account for some large differences in values and α can be adjusted according to input term size. Generally, δ_{dt} still remains subject to individual experimentation given specific input to yield enough but not an interminable number of terms.

$$\delta_{dt} = (\frac{1}{n} \sum_{k=1}^{n} dt^k + sd(dt)) \times \alpha \ (with\ \alpha > 0) \tag{3.3.8}$$

Although *Distinctiveness* is a comparative method and symmetric for the author's set and the comparison set, *Representativeness* relies solely on the author's set D and probably differs for the other set. Consequently, the differences between representative and distinctive terms might be different as well. For this reason, we compute characteristic features from both perspectives, the author's and the comparison set's. In the following, we explain the different subsets of the chosen discriminatory terms that we use for evaluation.

Comparing Authors on the basis of Representative Features

In order to evaluate how well the chosen terms do separate the two authors, we motivate the choice of only selecting representative features from both author profiles. The issue in connection with using all discriminatory terms lies in the calculation of the *Distinctiveness* measure. If we calculate representative and distinctive features for an author, we can be sure that the values for those terms are consistently similar for that author, while being different for the outside set. There are consequently two different scenarios with respect to a term *being different* in the other author's set.

1. The term t_i is consistent in set D with a high frequency. The same term t_i is consistent in the opposing author's set nD (for *nonDickens*) with a low frequency. Thus, the term is representative and distinctive for both sets, even though we did not consider the *Representativeness* for set nD. Obviously, the converse could also be true: a consistently low frequency for set D and a consistently high frequency for the set nD. This first case does not produce any issues for measuring similarity, since on the basis of these features there is reliable similarity within sets and accentuated differences between the sets.

2. The second possibility is the one that may cause problems. Assuming a representative and distinctive term for set D, with a frequency of either high or low. However, the same term is not representative for set nD and values may fluctuate from high to low. Although this term is not representative for nD, it is distinctive from D to nD, because it is constant in D while not being so in nD. Clustering on the dataset on the basis of these terms may create noise, since it will not show similarities for documents within nD and may have occasional rather similar values to the ones in D that rate it closer to documents in D.

We consider an example: based on a 85 x 500 most-frequent-features matrix (relative frequency weighting), we perform the RD method for both author sets, in this case Dickens (54 documents) and Collins (31 documents) (see section 3.1). Two representative and distinctive term lists are returned (dt_D and dt_{nD}), one for Dickens and one for Collins, which are then subjected to pre-selection at a level of $\alpha = 1.2$ (see eq. 3.3.8).

Tables 3.3.1 and 3.3.2 show the representative and distinctive terms for Dickens and Collins respectively. Table 3.3.3 is the intersection of $dt_D \cap dt_{nD}$ and thus the terms representative and distinctive for both Dickens and Collins. Table 3.3.4 shows those terms $t_i \subset (dt_D \cup dt_{nD}) - dt_D \cap dt_{nD}$. They are the residual of the intersection $dt_D \cap dt_{nD}$ and consequently those features that are only representative for one set and either not representative or not highly representative for the other.

Table 3.3.1: 127 ordered Dickens' representative and distinctive terms when compared to Collins on 500 most frequent terms.

Dickens' markers
upon, but, though, much, many, indeed, and, only, often, several, down, being, off, great, nor, pretty, left, very, fire, first, then, deal, towards, pleasant, person, all, always, afterwards, company, fact, still, however, therefore, none, because, rather, wind, enough, youll, coming, letter, such, times, suit, within, boy, question, high, heard, where, they, sometimes, return, leave, moment, there, every, shaking, lord, own, words, eye, side, life, glad, couldnt, bright, change, answer, along, across, power, fellow, already, short, everything, husband, about, necessary, asked, dead, street, excuse, full, speak, mind, woman, back, sitting, returned, kind, long, end, head, whole, who, sense, things, case, ever, shop, was, spoke, each, small, course, three, herself, room, other, she, men, like, those, good, better, less, understand, feel, wouldnt, for, arms, whom, dare, whether, town, conversation

In the following, we show that the set in table 3.3.3 is more suitable for a clustering comparison than the set in table 3.3.4. We compute the dissimilarity between documents on the basis of the list of term values, using the *complete link* measure (For details on the clustering method and evaluation of clustering, see section 4.1.2.). The example used here is an exaggerated case, since a list of representative and distinctive features for one

Table 3.3.2: 167 ordered Collins' representative and distinctive terms when compared to Dickens on 500 most frequent terms.

Collins' markers

upon, only, left, many, very, return, but, under, first, much, words, and, leave, down, letter, already, answer, being, since, though, returned, they, heard, indeed, feel, great, speak, enough, shaking, ask, full, air, end, still, brought, fire, place, has, were, observed, nor, her, moment, such, often, back, passed, spoke, question, stopped, looked, times, she, appearance, asked, off, interest, sat, next, bright, lost, told, their, object, its, circumstances,room, where, change, remember, new, then, mind, necessary, time, would, never, had, last, mentioned, herself, own, let, side, your, there, little, course, pleasant, which, open, letters, once, boy, turned, deal, can, always, felt, with, just, present, youll, couldnt, will, looking, received, several, one, like, sometimes, who, person, far, these, old, hands, into, without, itself, stood, cries, here, subject, cried, the, began, again, word, because, some, cold, within, come, darling, replied, house, use, woman, bring, having, large, corner, fine, friends, each, excuse, may, well, between, beautiful, was, why, secret, door, this, arms, hear, other, shall, part, truth, came, another, seems, duty, goes

Table 3.3.3: 73 representative and distinctive terms for both Collins and Dickens when compared on 500 most frequent terms.

Both Collins' and Dickens' markers

and, was, but, she, there, very, they, who, upon, much, down, like, such, then, being, great, where, only, other, back, own, first, mind, still, though, woman, room, always, many, off, returned, left, because, heard, indeed, boy, enough, fire, course, asked, moment, speak, times, letter, words, person, often, leave, herself, side, arms, full, question, within, sometimes,end, deal, nor, each, bright, answer, necessary,spoke, feel, pleasant, several, shaking, youll, change, couldnt, excuse, return, already

Table 3.3.4: 148 joined individual representative and distinctive terms for both Collins (94 terms) and Dickens (54 terms) when compared on 500 most frequent terms, but not including the the terms in table 3.3.3.

Both Collins' and Dickens' separate markers

under, since, ask, air, brought, place, has, were, observed, her, passed, stopped, looked, appearance, interest, sat, next, lost, told, their, object, its, circumstances,remember, new, time, would, never, had, last, mentioned, let, your, little, which, open, letters, once, turned, can, felt, with, just, present, will, looking, received, one, far, these, old, hands, into, without, itself, stood, cries, here, subject, cried, the, began, again, word, some, cold, come, darling, replied, house, use, bring, having, large, corner, fine, friends, may, well, between, beautiful, why, secret, door, this, hear, shall, part, truth, came, another, seems, duty, goes, pretty, towards, all, afterwards, company, fact, however, therefore, none, rather, wind, coming, suit, high, every, lord, eye, life, glad, along, across, power, fellow, short, everything, husband, about, dead, street, sitting, kind, long, head, whole, sense, things, case, ever, shop, small, three, men, those, good, better, less, understand, wouldnt, for, whom, dare, whether, town, conversation,

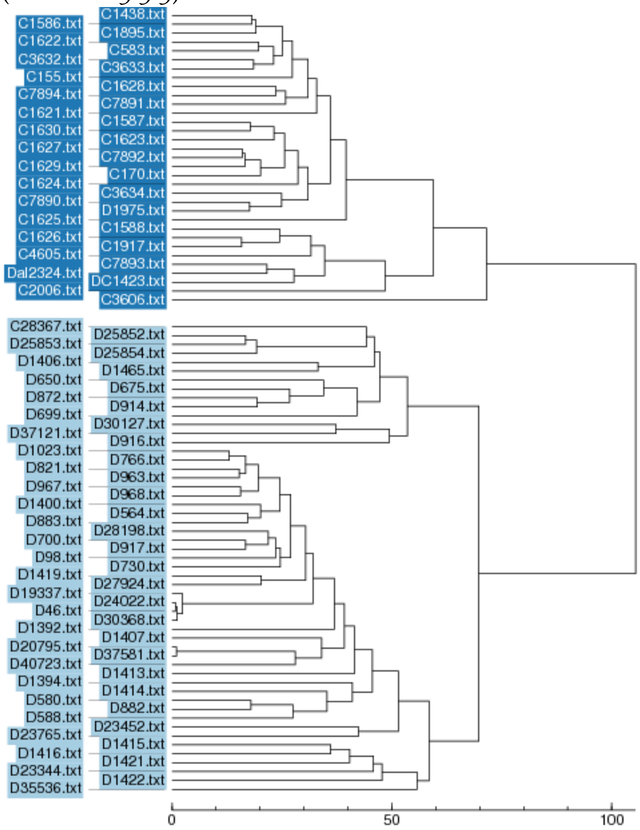
author set is to some extent also representative for the outside author set. However, we want to emphasize that those remaining non-shared features are less suited for clustering comparison and may introduce noise. The validity of their *Representativeness* for the individual set is not diminished and they are still regarded as coherent terms for that author.

The dendrograms in figure 3.3.1 and figure 3.3.2 show the clustering for both sets, the intersection of representative and distinctive features and the non-intersection set respectively. [7] In Figure 3.3.1, we can observe 3 misclassifications, one for Collins and two for Dickens.[8] The corresponding *adjusted Rand Index* (see section 4.1.2) given the ideal

[7] All illustrative figures were created in Gabmap: http://www.gabmap.nl/.

[8] All documents starting with a *D* refer to Dickens, all starting with a *C* refer to Collins' documents. Variations on this indicate collaborations between authors.

Figure 3.3.1: Dendrogram 'complete link' of dissimilarity Matrix on the basis of 73 both Dickens and Collins representative and distinctive terms for 500 most frequent input terms (see table 3.3.3).



separation is 0.82. Figure 3.3.2 shows 8 misclassifications for Dickens and 3 for Collins and the corresponding *adjusted Rand Index* is 0.018 and thus quite low.

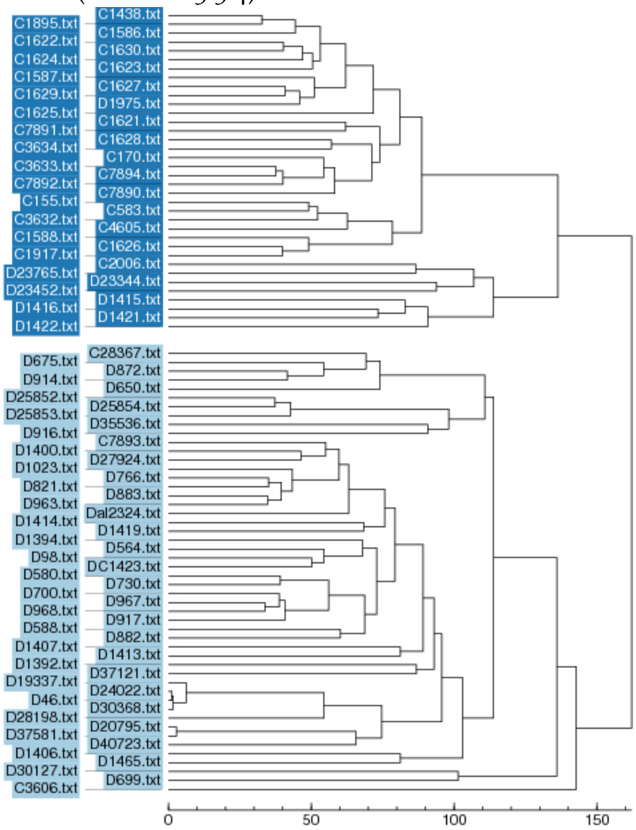
Naturally, the more terms lie in the intersection of both representative and distinctive terms for both sets, the higher their degree of *Representativeness & Distinctiveness* and the better the individual author's lists would perform, because the terms only representative for one author will have less influence in comparison. Since it is difficult to be sure of the exact distribution and also for the sake of consistency, for comparison and evaluation of discrimination ability, we choose only features representative and distinctive of both authors.

Selecting Frequent Discriminatory Terms for Author Profiles

Another particularity that needs to be addressed is the meaning of the discriminatory term list that holds distinguishing terms for each author based on comparison to another set. The terms within could be either consistently frequent or consistently infrequent for that author, depending on the comparison context. If we aim at comparing histogram differences of an author's profile (see section 4.1.1) and an unseen document on the basis of the representative and distinctive features, the consistently frequent features for an author provide a better basis for this comparison, since the representative and distinctive score rewards consistent features of an author regardless of their frequency.

Thus, we imagine a world where a term is either frequent for an author, such as Dickens or infrequent, which makes it frequent for his opponent, e.g. Collins. In order to select

Figure 3.3.2: Dendrogram 'complete link' of dissimilarity Matrix on the basis of 148 separate representative and distinctive terms for Dickens and Collins for 500 most frequent input terms (see table 3.3.4).



terms that are more indicative for one author than the other, we refer back to the input document-by-term matrix, weighted with relative frequencies. For each author, for every term we sum over the frequency of that term in all his documents divided by n number of documents. For one author set, for each term t_i, we compute its overall document frequency d_{freq} as defined in eq. 3.3.9.

$$d_{freq}(t_i) = \frac{1}{n}\sum_{j=1}^{n} relFreq(t_j) \tag{3.3.9}$$

This returns a d_{freq} list of average weights for each term in the overall set of the author. Having computed a list for each author, we now compare the values for term t_i for both authors and assign the term to the author in whose d_{freq} list it was more frequent. Those then remain in the frequency list d_{freq} for that author. There are no shared terms and merging the two document frequency lists gives us all original terms in the input matrix. Under this scheme, terms that are only slightly more frequent on average for one author are also assigned to his list. However, since this is merely a reference list and those terms are not likely to be selected for discriminatory terms, this should not have too negative an effect.

Then, given the representative and distinctive term list dt for the author, we know, that each term is either more frequent/absent for him compared to the other author's set. We compare to the frequency list for that author (d_{freq}) and only retain terms that are in the intersection of $d_{freq} \cap dt$.

We consider an example for the two-author set of Dickens and Collins. Having collected all mean term frequencies for each author, we observe that some terms are more frequent for one author than the other. There is one list for Dickens d_{Dfreq} and one for Collins d_{Cfreq}, while $d_{Dfreq} \cap d_{Cfreq} = \emptyset$, which is a necessary condition. Given a *dt* list for Dickens of 74 terms (after having applied the feature selection), we intersect those terms with his d_{Dfreq}.

Table 3.3.5 shows those markers that remain, when we retain only those of his *dt* features where he is likely to have more frequent scores than Collins. For the remaining 26 terms that are not in this list, Collins seemed to have a higher mean frequencies in his input documents, which is why those terms were allocated to Collins' list (d_{Cfreq}).

Although, this solution is somewhat heuristic, it nevertheless seems a reasonable approximation to identifying frequent/infrequent markers of an author. *Representativeness & Distinctiveness* alone primarily identifies characteristic terms that are consistently *different* for two authors without telling us which author consistently avoided or frequented certain terms. In this case, one possibility for selecting the negatively-associated terms for Dickens would be to extract those terms representative and distinctive for both authors, but only *frequent* for Collins.

Table 3.3.5: 48 representative & likely to be frequent features for Dickens.

Dickens' frequent representative and distinctive markers
and, was, but, all, there, very, they, who, upon, about, much, down, like, good, such, then, being, great, where, head, ever, long, though, always, many, off, every, returned, those, because, indeed, boy, whole, fire, three, things, coming, rather, kind, times, towards, everything,often, high, pretty, full, eye, short

3.4 MODEL SCENARIOS FOR CHARACTERISTIC TERM SELECTION

In the following section, we present three different models for creating author profiles, where section 3.4.1 describes the first model: separate *Representativeness & Distinctiveness*. Section 3.4.2 discusses the simple ICA model and section 3.4.3 considers the combined model of both ICA and *Representativeness & Distinctiveness*. Given two document sets of both Dickens and another author or reference set, each model is meant to yield an author profile consisting of both positively associated as well as negatively associated terms with some weight as an indication of how consistent the term is for that author. For each model, we exemplify the selection process and for all examples we use Charles Dickens and as opposing author Wilkie Collins. Input to all models is a *document* $\times$ *term* matrix constructed from a document set of Dickens and Collins and weighted with relative frequency.

3.4.1 *Model 1: Separate Representativeness - Distinctiveness*

The first model considers characteristic term selection using the *Representativeness & Distinctiveness* measure in isolation. Since this has been described in detail in section 3.3, we only give an overview of the process here.

CHARACTERISTIC TERM SELECTION FOR DICKENS AND COLLINS Figure 3.4.1 shows the the general selection of representative and distinctive features for both authors. Characteristic terms are then obtained by retaining all terms with weights over the mean plus the standard deviation of the complete weight list.

$$
\begin{array}{c}
document \times term\ matrix \\
\Downarrow \\
representative/distinctive\ terms : t_{rd} \\
\Downarrow \\
\text{take mean } \mu \text{ and } \sigma \text{ of all values:} t_{rd} \\
\Downarrow \\
retain\ t_{rd} > \alpha \times (\mu + \sigma) \\
\Downarrow \\
Dickens'\ Profile :\ P_D
\end{array}
\qquad
\begin{array}{c}
document \times term\ matrix \\
\Downarrow \\
representative/distinctive\ terms : t_{rd} \\
\Downarrow \\
\text{take mean } \mu \text{ and } \sigma \text{ of all values:} t_{rd} \\
\Downarrow \\
retain\ t_{rd} > \alpha \times (\mu + \sigma) \\
\Downarrow \\
Collins'\ Profile :\ P_C
\end{array}
$$

Figure 3.4.1: Characteristic Term Selection Process in *Representativeness & Distinctiveness* for Dickens and Collins

REPRESENTATIVE FEATURES FOR CLUSTERING The terms we retain for clustering evaluation consist of the intersection of the two author profiles, P_D and P_C. Thus, we obtain terms that are consistent for both author sets.

(1) $Dickens_{rep} \& Collins_{rep} = P_D \cap P_C$

SELECTING FREQUENT FEATURES FOR AUTHOR PROFILES In order to select only the most frequent features of an author for histogram comparison, we compute mean term frequencies on the basis of the *document* $\times$ *term* matrix for all terms for each author and divide terms into d_{Dfreq} and d_{Cfreq} for Dickens' frequent items and Collins' frequent items respectively, so that $d_{Dfreq} \cap d_{Cfreq} = \emptyset$. The respective profiles are then compared to the frequency lists and we retain only those terms for which an author is likely to have been more frequent than the opposing author.

(2) $P_{Dfreq} = d_{Dfreq} \cap P_D$

(3) $P_{Cfreq} = d_{Cfreq} \cap P_C$

3.4.2 *Model 2: Separate Independent Component Analysis*

Similarly to before, since the simple model of ICA for characteristic term selection is described in detail in section 3.2.4, we only briefly depict the general process here.

1. From the *document* $\times$ *term* $\Rightarrow$ estimation of $A_{document \times component}$ and $S_{component \times term}$
2. A: for each term t_k, compute its threshold δ_{t_k} given all its values and retain its presence in components, where $w(t_k) > \delta_{t_k}$
3. S: for each component c_i, compute its threshold δ_{c_i} given all its weights and retain it for documents, where $w(c_i) > \delta_{c_i}$
4. Reconstruct a reduced *document* $\times$ *term* matrix by combining term-component weights and component-document weights:

$$
w_{t_k, d_j} = \sum_{i=1}^{n} \alpha_{ij} * \beta_{ik}
$$

5. *Dickens' Profile*: for each term in Dickens' documents, compute μ over all weights and keep terms with highest absolute weight, as this yields positive and negative terms

6. *Collins' Profile*: for each term in Collins' documents, compute μ over all weights and keep terms with highest absolute weight

3.4.3 *Model 3: ICA & Representative and Distinctive Components*

The third model combines both techniques by first using ICA to compute A and S and then reducing S by selecting the most representative and distinct components for each author's set.

1. From the *document* $\times$ *term* $\Rightarrow$ estimation of $A_{document \times component}$ and $S_{component \times term}$.
2. A: for each term t_k, compute its threshold δ_{t_k} given all its values and retain its presents in components, where $w(t_k) > \delta_{t_k}$
3. S: using *Representativeness & Distinctiveness* select components most representative and distinctive for each author $\Rightarrow$ cs_1, cs_2: set of components for Dickens and Collins respectively. This is a supervised selection of components according to their discrimination ability of the two author sets.
4. Reconstruct a reduced *document* $\times$ *term* matrix by combining term-component weights and component-document weights, according to cs_1, cs_2:

$$w_{t_k,d_j} = \sum_{i=1}^{n} \alpha_{ij} * \beta_{ik}$$

5. *Dickens' Profile*: for each term in Dickens' documents, compute μ over all weights and retain terms with highest absolute weight
6. *Collins' Profile*: for each term in Collins' documents, compute μ over all weights and retain terms with highest absolute weight

4 EVALUATING DICKENS' CHARACTERISTIC TERMS

In this chapter, we evaluate the contribution of this work, in particular the appropriateness of the characteristic term selection models proposed in the previous chapter. The quality of each model is evaluated with respect to the discrimination ability and consistency of its choice of characteristic terms for an author's document set in comparison to another author or reference set comprising different authors. Since there is no gold standard that defines the relative importance of a given term for an author, evaluation of a ranked characteristic term list and consequently also a specific model that produced this list is based mainly on the ability to identify unseen documents of the author, on degree of clustering ability and consistency in term selection given different training sets.

Thus, in section 4.1, we describe general evaluation methods that should help determine the validity of the chosen characteristic terms as well as the corresponding model. All methods are generally applicable to all model scenarios with some adjustments allowing for basic differences between *Representativeness & Distinctiveness* and ICA weighting. This should also allow us to compare between models and determine whether the two separate methods are to be preferred to the combined approach.

In section 4.2, we evaluate the results of the different models on the data sets and compare between the different models as well as to results of the previous studies of Dickens' style.

4.1 EVALUATION METHODS

In this section, we explain how characteristic terms are evaluated according to different criteria, such as relative closeness of an author profile to an unseen document, consistency of term selection when different subsets of the training corpus are chosen and separation ability in clustering. For all experiments, we consider different measures of correctness, given certain desirable characteristics of the results as stated in the following. Considering a discriminatory term list of a set of *Dickens'* documents as opposed to a set of document not by *Dickens*, referred to here as *nonDickens*, the following criteria should be met:

1. Cross-validation: performance of discriminatory term lists / author profile
    - Unseen Dickens histograms should be closer to $Dickens'$ profile histograms than the $nonDickens'$ profile histograms
    - Unseen $nonDickens$ histograms should not be close
2. Clustering based on characteristic terms should discriminate
3. Consistency of term lists/profiles over different iterations

Each of these three criteria is addressed in one separate section and thus section 4.1.1 explains how cross-validation is performed and how on the basis of a set of ranked characteristic terms, one can obtain an author profile and test a profile's closeness to an unseen document. Section 4.1.2 explains and exemplifies clustering on the basis of characteristic terms of two author sets and section 4.1.3 addresses the consistency of term selection for different subsets of an author's document collection.

4.1.1 *Relative Histogram Differences of Author Profiles*

For the first requirement of estimating author profile closeness, we test for the average distance between an unseen document (Dickens or other) and an author's profile based on the terms in that profile. Generally, an unseen document is always compared to both the *Dickens* and the *nonDickens* profile. A good author's profile should always have a lower distance for unseen documents belonging to that author than for those belonging to the *opposing* set.

Having applied a model for characteristic term selection on a training document-by-term matrix x, we obtain profiles P_D and P_{nD}, (for the Dickens and non-Dickens set respectively) each containing a set of terms that are considered discriminatory for the individual author set. For testing generalization ability, these profiles have to be evaluated against documents in the set of $\cup_{Test_D}$ or $\cup_{Test_{nD}}$ documents, but not in the training set $\cup_{Train_D}$ or $\cup_{Train_{nD}}$ that formed the basis for the document-by-term matrix x used in training.

Given an author profile P containing a set of chosen discriminatory terms t, we would like to know the relative importance of each term in the profile in relation to all other terms in the profile. For this purpose, we compute the relative frequency histograms over the profiles P_D and P_{nD}, where e.g. for any profile P, the histogram value rt_j of the weight for term $w(t_j)$ in profile P is defined by eq. 4.1.1.

$$rt_j = \frac{w(t_j)}{\sum_{i=1}^{n} w(t_n)}, \textit{where n = no. of terms in the profile} \tag{4.1.1}$$

For comparison, we choose an unseen document d_{test} that was not part of the training set and compare each profile separately to the unseen document. Each profile P is compared to d_{test} on the basis of the relative frequency distribution over the terms in P, which follows the assumption that if an unseen document belongs to a certain author and the terms in the author's profile are representative for that author, the distributions over those terms in both profile and unseen document should be very similar and more similar than when comparing to another author's profile. For creating a document vector of d_{test}, the same preprocessing and weighting as for the training set has to be used to make comparisons valid.

Thus, for both profiles P_D and P_{nD}, the following steps are performed separately: given a profile P, the d_{test} vector is reduced to only the terms t in the profile. The relative frequency histograms are computed for both the profile P and reduced d_{test} as described above. In order to determine how much the two histograms differ, the difference between rt_j of all terms t_j in d_{test} and P is compared using the *Manhattan* distance or absolute distance. Consequently, $dist(P, d_{test}, rt_j)$ refers to comparing P and d_{test} with respect to rt_j as in eq. 4.1.2

$$dist(P, d_{test}, rt_j) = |P_{rt_j} - d_{rt_j}| \tag{4.1.2}$$

To obtain the mean difference between a profile P and the test document d_{test}, we take the mean over all distances, as defined in eq. 4.1.3. This also accounts for differences in profile length of the two author profiles compared.

$$mdist(P, d_{test}) = \frac{1}{n} \sum_{j=1}^{n} dist(P, d_{test}, rt_j) \tag{4.1.3}$$

After the above steps have been performed for both profiles P_D and P_{nD}, the mean distances $mdist(P_D, d_{test})$ and $mdist(P_{nD}, d_{test})$ are compared. If the document has been one of $\cup_{Test_D}$, a discriminatory profile P_D should have a lower value for $mdist(P_D, d_{test})$ than profile P_{nD} for $mdist(P_{nD}, d_{test})$ and conversely, if the document has been in $\cup_{Test_{nD}}$,

$mdist(P_{nD}, d_{test})$ should be lower. Documents in $\cup_{Test_{nD}}$ can be tested for a negative match and reveal general issues with the profile selection method, although a good profile for Dickens should rather be chosen on the basis of similarity to an unseen Dickens document.

CROSS-VALIDATION: CHOOSING THE BEST KEYWORD LIST In order to choose the best profile, we can use cross-validation on the training set using the method proposed in this section for comparison. For document vector d_i in $\cup_{Train_D, Train_{nD}}$ (for i ∈ 1...n documents), we remove d_i from the training set, train the remaining $n-1$ document vectors and test each resulting P_D/P_{nD} on d_i. The best model profile for Dickens has the smallest distance for an unseen Dickens document and respectively for *nonDickens*, we choose the smallest distance for an unseen *nonDickens* document. Cross-validation can be done with leaving out only one document, which uses all resources (*Leave-one-out* validation), but is computationally expensive or by leaving out more than one (e.g. *Leave-five-out* validation). For all of our experiments, we use the latter option of removing five new documents on each iteration.

DISTRIBUTIONS OF AUTHOR PROFILE DISTANCES Thus, after each iteration in cross-validation, we measure the distance of an author profile to an unseen document based on the relative frequency histogram distribution of the terms in the given author profile. Generally, we would like unseen Dickens documents to be closer to the *Dickens* profile than the *nonDickens* profile. Additionally, it would also be preferable if the individual distances of the two author profiles to the unseen document are likely to originate from two different distributions, meaning that their distribution mean is not likely to be the same.

For this purpose, we also consider an *Independent Sample T-Test* on the two distributions of individual distances between author profile and unseen document. Essentially, this is testing how confident we are that a single document belongs to a certain author. If the mean difference between the two samples is high and significant, the current model is more confident about its choice. For this, we take the distribution consisting of the list of individual distances between Dickens' profile and the unseen document: $dist(P_D, d_{test}, rt_j)$ for all t_j in profile P_D and the distribution consisting of the list of individual distances for *nonDickens*: $dist(P_{nD}, d_{test}, rt_j)$ for all t_j in profile P_{nD}.

For illustration, we consider an example of evaluating two different author profiles using *Leave-five-out* cross-validation. The t-test was computed using *Welch's* test with different sample sizes (two term lists seldom have the same length) at an α significance level of 0.05.[1] Depending on the type of test document, we assume one group mean to be greater than the other, e.g. if a Dickens' document is tested, the assumption is that the *nonDickens* profile has a larger mean distance to the test document.

Table 4.1.1 shows the results for testing *Dickens* and *Collins* profiles on 23 Dickens test documents. The profiles were computed using the separate ICA model on a 47 Dickens/Collins document set, which is described in more detail in section 3.1.1.

The mean distances between the *Dickens*/*Collins* profile and the respective test document are displayed in column two and three under *Dist.D.* and *Dist.C.* respectively. Column four computes the difference between *Dickens* and *Collins* distances, i.e. how much closer Dickens' profile is to the document. Since here we are testing for unseen Dickens documents, the assumption is, that Dickens should always have a lower distance to the unseen document. Consequently, for each iteration the distance for *Dickens* is deducted from the one of *Collins* and deducting a smaller value from a larger one should always be positive.

Further, column five shows the p-value given the alternative hypothesis that the sample means of *Collins* has a greater mean than that of Dickens. Since in all cases shown here, p

[1] This was computed in R using: t.test($dist(doc - prof_D)$, $dist(doc - prof_C)$, alternative="greater")

< 0.05, we can reject the null hypothesis that the sample means are equal, which means that there does seem to be a significant difference between the two samples for all of Dickens' test documents. At a confidence level of 95%, it is assumed that the difference between the sample means lies in the interval displayed in column six. All intervals are positive, meaning that the mean difference is unlikely to ever be zero. With respect to our author profiles, this indicates that the profiles constructed for Dickens are appropriate in so far as to seemingly recognize Dickensian test documents.

Table 4.1.1: ICA on *DickensCollinsSet1*. Results of evaluating distances for profiles P_D and P_C to test closeness to Dickens' documents also showing t-test results for hypothesis assuming greater mean for *Collins* profile to the test document.

Author Profile Comparison						
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)
1	D33_SB	0.0107	0.0157	0.0050	0.00	0.0027 ... Inf
	D36_PP	0.0107	0.0160	0.0054	0.02	0.0011 ... Inf
	D37a_OEP	0.0099	0.0155	0.0056	0.00	0.0023 ... Inf
	D37b_OT	0.0086	0.0150	0.0064	0.00	0.0038 ... Inf
	D38_NN	0.0076	0.0158	0.0083	0.00	0.0058 ... Inf
2	D40a_MHC	0.0086	0.0166	0.0079	0.00	0.0051 ... Inf
	D40b_OCS	0.0066	0.0168	0.0102	0.00	0.0074 ... Inf
	D41_BR	0.0065	0.0170	0.0105	0.00	0.0081 ... Inf
	D42_AN	0.0083	0.0160	0.0076	0.00	0.0052 ... Inf
	D43_MC	0.0064	0.0175	0.0111	0.00	0.0078 ... Inf
3	D46a_PFI	0.0089	0.0161	0.0072	0.00	0.0046 ... Inf
	D46b_DS	0.0061	0.0166	0.0105	0.00	0.0082 ... Inf
	D49_DC	0.0077	0.0154	0.0077	0.00	0.0056 ... Inf
	D51_CHE	0.0107	0.0158	0.0051	0.02	0.0010 ... Inf
	D52_BH	0.0076	0.0165	0.0089	0.00	0.0064 ... Inf
4	D54_HT	0.0090	0.0161	0.0070	0.00	0.0038 ... Inf
	D55_LD	0.0072	0.0164	0.0092	0.00	0.0070 ... Inf
	D56_RP	0.0085	0.0170	0.0085	0.00	0.0064 ... Inf
	D59_TTC	0.0095	0.0152	0.0057	0.00	0.0034 ... Inf
	D60a_UT	0.0084	0.0168	0.0084	0.00	0.0064 ... Inf
5	D60b_GE	0.0110	0.0153	0.0044	0.00	0.0018 ... Inf
	D64_OMF	0.0108	0.0151	0.0043	0.00	0.0016 ... Inf
	D70_ED	0.0100	0.0161	0.0061	0.00	0.0035 ... Inf
	mean	0.0087	0.0161	0.0074		
	sd	0.0015	0.0007	0.0020		
	SE	0.0003	0.0001	0.0004		

Additionally, a paired t-test can be performed on the mean distances of all test instances in cross-validation, which compares the samples $mdist(P_D, d_{test})$ and $mdist(P_C, d_{test})$ over all test documents. Globally, this corresponds to model evaluation, i.e. how well a profile recognizes the correct test documents. For this purpose, we construct two samples: one containing all mean distances for Dickens (column one in table 4.1.1) and the second sample containing all mean distances for Collins (column two in table 4.1.1). The paired t-test confirms model validity with a p-value of 0.02 and a positive confidence interval of 0.002 to Inf, meaning that Dickensian documents are reliably classified as such ones. A second trial using Collins' test documents is then also performed to ensure Collins' profiles validity and the ability of Dickens' profiles to also reject foreign author profiles.

4.1.2 *Clustering Dissimilarity of Author Sets*

Given a list of discriminatory terms for two different author sets, we would like to ascertain to what extent the collection of terms is able to highlight differences between the sets and identify distinct clusters grouping the documents of different authors. As has been shown before, the terms used for discrimination ability should be selected according to separation ability for both author sets. Ideally, frequencies with respect to all terms should be consistent and fairly complementary between two author sets, e.g. Dickens uses *upon* consistently and frequently and Collins uses the term consistently and infrequently. In order to test discrimination ability of a discriminatory term list for two authors, we build a dissimilarity matrix comparing all documents in the complete training set.

Dissimilarity Matrix

A dissimilarity matrix (or distance matrix) D_M describes pairwise distances for M objects, which results in a square symmetrical MxM matrix, where the ij_{th} entry is equal to the value of a chosen measure of distinction d between the i_{th} and the j_{th} object. The diagonal elements, comparing an object to itself are not considered or are usually equal to zero. A sample dissimilarity matrix is shown in matrix 4.1.4. Thus, in our case each document pair in $Dickens \cup nonDickens$ is compared based on the differences of $term_i$ in a given term list. A common measure of distinction d would be *Manhattan* or *Euclidean* distance.

$$D_M = \begin{pmatrix} 0 & d_{12} & \dots & d_{1j} \\ d_{21} & 0 & & \vdots \\ \vdots & & \ddots & \vdots \\ d_{j1} & \dots & & 0 \end{pmatrix} \tag{4.1.4}$$

Clustering on the basis of dissimilarity between objects, in this case documents can be done via hierarchical clustering. Agglomerative hierarchical clustering, for instance is an iterative clustering process, whereby cluster objects are joined together based on a distance measure between the elements within the clusters. All elements begin in their own clusters and and are joined until the desired number of output clusters has been reached. A common distance measure for joining clusters together is the *complete link* method, which assesses closeness on the basis of the most distant elements in two clusters X and Y, in order to avoid the merging of two clusters based on only two single elements from each set being close. The distance $D(X,Y)$ between clusters X and Y is defined in eq. 4.1.5, where X and Y are two sets of elements or clusters and $d(x,y)$ is the distance between elements $x \in X$ and $y \in Y$.

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y) \tag{4.1.5}$$

Adjusted Rand Index for Evaluation of Clustering

In addition to visual clustering that gives more of an intuition of separation between two sets, a clustering result can be evaluated by comparing two different partitions of a finite set of objects, namely the clustering obtained and the ideal clustering. For this purpose, we can employ the *adjusted Rand Index* (Hubert and Arabie 1985), which is the corrected-for-chance version of the *Rand Index*. Given a set S of n elements, and two clusterings of these points, U and V, defined as $U = \{U_1, U_2, \dots, U_r\}$ and $V = \{V_1, V_2, \dots, V_s\}$ with a_i and b_i as the number of objects in cluster U_i and V_i respectively. The overlap between U and V can be summarized in a contingency table 4.1.6. where each entry n_{ij} denotes the number of objects in common between U_i *and* $V_j : n_{ij} = |U_i \cap V_j|$.

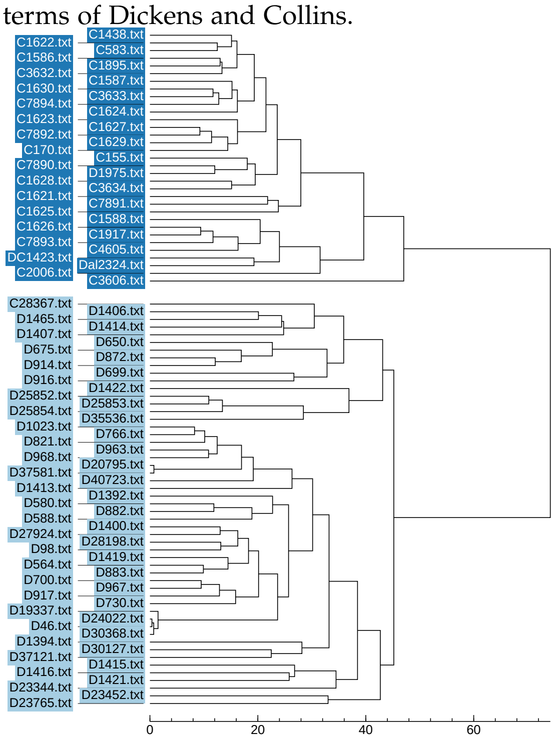
$$
[n_{ij}] = \begin{array}{c} U\ V \\ U_1 \\ U_2 \\ \vdots \\ U_r \\ Sums \end{array}
\begin{pmatrix} V_1 & V_2 & \dots & V_s & Sums \\ n_{11} & n_{12} & \dots & n_{1s} & a_1 \\ n_{21} & n_{22} & \dots & n_{2s} & a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ n_{r1} & n_{r2} & \dots & n_{rs} & a_r \\ b_1 & b_2 & \dots & b_s & \end{pmatrix} \tag{4.1.6}
$$

The adjusted form of the *Rand Index* is defined in eq. 4.1.7 and more specifically given the contingency table 4.1.6 in eq. 4.1.8, where n_{ij}, a_i, b_j are values from the contingency table.

$$
AdjustedIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex} \tag{4.1.7}
$$

$$
ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{4.1.8}
$$

Figure 4.1.1: Dendrogram 'complete link' of dissimilarity Matrix on the basis of 300 input terms of Dickens and Collins.



The index is bounded between [-1,1], with 0 being the expected value and 1 the highest positive correlation between two different clusterings. For illustration of using the two methods presented above, we consider an example of pairwise comparison of documents of a dataset of *Dickens* $\cup$ *Collins*, with 55 documents belonging to Dickens and 31 to Collins. This yields a 86 x 86 dissimilarity matrix containing all pairwise comparisons of documents in the set. Figure 4.1.1 depicts an example dendrogram showing clustering based on a dissimilarity matrix with distances computed using the *complete link* measure. The *adjusted Rand Index* corresponding to the clustering in figure 4.1.1 is 0.82, so very close to the ideal separation, which is also confirmed, when we consider the small number of misclassifications (3 for Dickens and 1 for Collins).

4.1.3 *Profile Consistency*

Since characteristic terms of an author should be fairly consistent and independent of the exact sample of his documents, we expect a sound method for characteristic term selection to be able to return a substantial overlap of terms on each iteration of cross-validation. For this purpose, we monitor individual cross-validation profiles, i.e. given that a particular document is removed from the training corpus, we keep the exact list of terms identified on the basis of the new training set that is basis for that iterations' author profile. The assumption here is that leaving out different documents of an author should not lead to a vast difference in the term list if the method under investigation is in fact able to detect distinct stylistic elements of an author.

If the terms selected after each iteration differed due to choice of a different author sample, the method would not detect distinct terms of an author and the process would be rather unstable and validity of the characteristic terms in general would be questionable. In order to test for consistency of term choice between iterations, we keep track of all author's profiles and compute their intersection. Despite our previous pre-selection, as for instance in the case of *representative* and *distinctive* terms, here we use the complete set for each author, since consistency is solely based on the intersection of all profiles given different iterations in cross-validation. Thus, we compute the overall intersection of all profiles given n number of cross-validation iterations as defined in eq. 4.1.9.

However, using this method, we are not measuring the degree of agreement with respect to profile length and number of intersections. The more terms each list holds and the more individual lists are intersected, the less likely becomes a high degree of agreement. Although should there be in fact a high consensus among individual profiles, even a large number of intersections should retain a substantial number of common terms.

$$\cap Profile = Profile_1 \cap Profile_2 \cap ... \cap Profile_n \tag{4.1.9}$$

Table 4.1.2 shows an example of intersecting individual profiles of Dickens for 13 iterations. Column two lists the simple length of each list regardless of the number of terms intersecting and column three then shows how many common terms are left after the intersection with the previous iterations' lists. Since single documents do differ with regard to the frequency of terms, slight deviations are to be expected. In this example, the length varies about 1-2 terms per list. Considering the high number of terms left after each intersection, the changes are very slight.

Table 4.1.2: Testing profile consistency by intersecting each new profile of an author by the intersection of the previous ones.

iteration	Doc. removed	Profile length	Terms after intersection
1	D1023	52	52
2	D1392	53	51
3	D1394	50	50
4	D1400	51	50
5	D1406	53	50
6	D1407	52	50
7	D1413	52	50
8	D1414	49	49
9	D1415	55	49
10	D1416	53	49
11	D1419	52	49
12	D1421	52	49
13	D1422	52	49
mean		52	49
sd		1.5	1.0

4.2 EVALUATION OF DICKENS' TERMS

In this section, we evaluate the three different models proposed earlier on our datasets. In order to ensure comparability between models, we aim to take the same input terms, when possible.

Section 4.2.1 recounts some findings of earlier experiments that influence our choice in parameters for the main evaluation, further in section 4.2.2 we explain differences in evaluation of author profiles of the two methods employed. In section 4.2.3, we consider the first dataset, *DickensCollinsSet1*, and the second dataset, *DickensCollinsSet2*, for all three models. In section 4.2.4, we then attend to the third dataset of *DickenWorldSet*. Finally, in section 4.3, we discuss our results in comparison to the previous work on Dickens' style and close with a general overview of contribution and future work.

4.2.1 *Characteristic Term Experiments*

In order to better comprehend how the different parameters, such as term input size or number of terms in profiles affect the model performance and results, we consider these influences on our methods and with respect to our different models. Generally, for each profile, we would like a substantial number of discriminatory terms, e.g. about 50-100 to make estimation reliable for later interpretation.

However, since all our models return ranked profiles, the more terms we extract the less discriminating the profile becomes, so that results may deteriorate. This dilemma is especially relevant for the first model of *Representativeness & Distinctiveness* with respect to comparing profiles, as we generally need to extract more terms to ensure that we obtain sufficiently frequent ones.

Factors Influencing Representative and Distinctive Terms

Judging from previous experiments, the joined value representing the degree of *Representativeness & Distinctiveness* mainly relies on the sample of documents for the author and the opposing set. After having estimated the representativeness of a term, highly representative terms are chosen from this pool depending on the distinctiveness of the term, which depends on the comparison set. Given that the document set is large enough, leaving out some documents should not have a considerable effect on the highest ranked terms, since the measure places a lot of emphasis on consistency.

In contrast, increasing the number of input terms does not change the degree of *Representativeness & Distinctiveness* of individual terms, but has an impact on the ranking, as previously higher ranked terms are occasionally shifted downwards and elements are inserted if they are more representative and distinctive. Increasing the input size should improve the terms in the profile regarding discrimination ability, as more terms are tested for potential suitability.

However, this presents an issue with respect to testing author profiles on the basis of frequency, since the more infrequent terms are considered, the more infrequent terms are likely to be included in the final profile.

Factors Influencing Characteristic Terms of ICA Models

The effects of the sample of input terms to the ICA models is more subtle, as ICA tries to find common characteristic deviations over terms in the different documents and then build concepts accordingly. In addition, ICA is restricted by the term-document ratio and the number of document samples directly determines the number of possible input terms. If the ratio grows too large the matrix becomes singular and components cannot longer be estimated.

Previous experiments have shown that estimation of components given a document set size of 47 to 80 documents performs best with about 1500 input terms. Additional experiments on different frequency strata indicated that removing the 50-70 most frequent terms of the 1500 most frequent terms positively influences both estimation and later profile performance. The exact number of terms to be removed should be subject to closer experiments, since obviously we would like to retain as many very frequent terms as possible. We attribute the better performance to the fact that for the input terms to ICA models, we use only relative frequency weighting as opposed to smoothing over larger frequencies using logarithm. Were we to use logarithm additionally, interpretation of the ICA results would be considerably complicated.

Also, we found that keeping the number of to-be-extracted components at about 47-55 components even with a larger document size substantially improves performance, which might be attributed to the fact that on increasing document size the number of theoretical concepts stays relatively stable. Due to time constraints, we have not exhausted all experiments with respect to combinations of factors and how they influence the final result. The following evaluation is based on approximations satisfying most of the current criteria.

4.2.2 *Differences in Evaluation of Representativeness & Distinctiveness vs. ICA*

The two methods employed here, *Representativeness & Distinctiveness* and *Independent Component Analysis* differ in one basis characteristic, being supervised and using class labels for term selection opposed to being unsupervised without reference to information about class membership. This property causes differences in the way we perform the evaluation of author profile histograms of each method.

EVALUATION OF TERMS IN ICA Since *Independent Component Analysis* is unsupervised and does not take class labels into account for determining the distribution of terms over components, evaluation can be done on ICA weights directly. For cross-validation, we first compute the independent components for the complete set. Then, depending on the type of cross-validation, for each iteration we remove one or more documents from the document-component matrix, compute author profiles on the basis of the training set and then evaluate on the test documents. This has the advantage, that we compare ICA weights directly to ICA weights and profile distance evaluation should be valid.

EVALUATION OF REPRESENTATIVE AND DISTINCTIVE TERMS The choice of representative and distinctive terms is on the basis of the class labels, which renders an approach like the previous one for ICA impossible. Instead, we use the ranked list of representative and distinctive values for terms as basis for relative histogram comparison. This solution is far from being ideal, because these scores do not necessarily correspond to the relative frequencies of the respective terms, but rather consistency in term weights over a document set. A term could have a lower frequency, but still be awarded a high representative and distinctive value, simply because it is constant over many documents, while being inconsistent or consistent with a different frequency for the comparison set.

To some extent, this issue is alleviated by our excluding infrequent terms for an author, but there still might be irregularities resulting from this issue and evaluation for *Representativeness & Distinctiveness* on the basis of profile distances should be regarded with caution. Another aspect is that by removing the infrequent items, we may also remove highly representative and distinctive items, which are replaced by terms less discriminatory but more frequent.

4.2.3 *Characteristic Terms of Dickens and Collins (1) and (2)*

For the first experiment, we consider the *DickensCollinsSet1*, a 47 x 4999 document-by-term matrix that was introduced in section 3.1.1.

In order to be able to compare directly to the results in Tabata 2012, we consider the same input and use all terms as input when possible. For the first model of *Representativeness & Distinctiveness*, we use the full input feature set.

Since both ICA-based models are restricted by the document-term ratio, we cannot use all input features, but only a subset of 1500 terms to make estimation of components still valid. Thus, for strict comparison based on the sample of input features, only the first model is directly comparable to the results in Tabata 2012. The type of cross-validation is leave-five-out, so at each iteration, we remove five new documents from the set for testing the profiles.

In addition, we also briefly report on the results using the *DickensCollinsSet2*, that differs with respect to the data set and weighting scheme. Since results were not substantially different and also for reason of succinctness, we do not report on them in the same level of detail. However, all results are referenced alongside the ones conducted on the *DickensCollinsSet1*. For all experiments on the *DickensCollinsSet2*, the number of input terms was reduced to the 1500 most frequent terms, while removing the 70 most frequent terms.

Representative & Distinctive Terms of Dickens vs. Collins

In order to compute the representative and distinctive terms for the *DickensCollinsSet1* matrix, we choose all 4999 input terms and set α to 1.3. In experiments, this threshold was found to yield a sufficient number of terms in each profile. Thus, we discard all terms with scores $< 1.3 \times (\mu + sd)$, where μ is the mean over all term values and sd is the standard deviation over the term values. For the second set of *DickensCollinsSet2*, due to using fewer input terms, α is also lowered to 1.1.

Table 4.2.1 shows the results for testing both Dickens' and Collins' profile on Dickens' documents. Each iteration relates to a different profile for each author based on the current training set, given that five test documents of Dickens have been removed. *Dist.D.* and

Dist.C. respectively show the mean differences for the profiles of Dickens and Collins to the current test document on the basis of the terms contained in their individual profiles.

Consequently, all documents in one iteration are compared on the basis of same profile. Further, *Dist.C-Dist.D* computes the difference between the two mean differences given one test document. In this case, where test documents are Dickensian, we would like the difference to be positive based on the assumption that all Collins profiles should have a larger distance to an unseen Dickens document. The remaining values are the p-values from the t-test over the individual histogram differences from each profile to the test document and corresponding confidence intervals. Ideally, we would like the two histograms of profile-test differences to originate from two different distributions as a consolidation of choosing the correct closest profile.

With respect to the present results, these are not favourable, since the Collins profiles are consistently rated closer to all of Dickens' test documents. Correspondingly, all p-values are not significant, so there is no significant difference in mean.

Additionally, we compute a paired, one-tailed t-test over all mean differences from each profile for all test documents, i.e. comparing all values in *Dist.D.* to the ones in *Dist.C.*. The t-test over the mean differences yields a p-value of 1 (with a confidence interval of -0.0134 to *Inf*), thus given the null hypothesis that sample means are equal, obtaining a mean of the differences of -0.012 or greater is almost certain. These results in isolation strongly indicate a very unsuitable selection of terms in the profiles.

Our original assumption with respect to the profile comparisons was that we select the most characteristic terms for each author and on the basis of his best terms determine an unseen document's closeness. However, in this case one might suspect that closeness is not in fact accurately measured, which is the reason why we take a closer look at the comparisons. We argue that the terms in Dickens' profile are appropriate, but that their representative and distinctive value is simply less close to the relative frequency distribution than the terms in Collins' profile.

For this purpose, we compare the unseen Dickensian document to both Dickens and Collins as before, but do two comparisons using the same terms for both authors each time. Thus, we compare the unseen document to Dickens and Collins both on the basis of Collins' highest terms and on the basis of Dickens' highest terms, although on the basis of their respective representative and distinctive values for those terms. Comparing the mean distances for both authors to the unseen Dickens document on the basis of Collins' terms returns mean distances of 0.015 and 0.015 and similarly, the mean distances on the basis of Collins' terms is 0.029 and 0.029.

If Collins' profile was truly closer to the unseen Dickens' document, Dickens should still not be competitive using the same terms. This strongly indicates, that successful terms are rather influenced by a representative and distinctive value close to the general relative frequency rather than true similarity. It leads us to believe that the current evaluation scheme is not measuring authorship accurately.

With regard to the Collins' test documents, the results are slightly more favourable, as shown in table 4.2.2, which is also supporting our previous analysis. All test documents are rated closer to the Collins' profiles, which is confirmed by significant p-values indicating a true difference in mean between all profile-test instances in favour of Collins. The model evaluation using a paired t-test comparing profile mean differences to the test document is also significant with a p-value less than 0.001. The evaluation on the basis of the *DickensCollinsSet2* in table B.1.1 and table B.1.2 shows a very similar result with respect to author profile comparisons.

However, if we consider the last column showing the *adjusted Rand Index*, the results for all iterations are rather positive and thus inconsistent with the earlier findings of unsuitable characteristic markers of Dickens. The clustering is performed on the basis of representative

Table 4.2.1: Representativeness & Distinctiveness on *DickensCollinsSet1*. Results of evaluating distances for profiles P_D and P_C to test closeness to Dickens' documents also showing t-test results for hypothesis assuming greater mean for *Collins* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis of shared representative and distinctive terms of both profiles.

				Author Profile Comparison			Clustering
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
1	D33_SB	0.0291	0.0145	-0.0145	0.94	-0.0300 … Inf	0.83
	D36_PP	0.0282	0.0145	-0.0137	0.94	-0.0280 … Inf	
	D37a_OEP	0.0281	0.0150	-0.0131	0.93	-0.0280 … Inf	
	D37b_OT	0.0273	0.0142	-0.0131	0.94	-0.0271 … Inf	
	D38_NN	0.0273	0.0145	-0.0129	0.93	-0.0269 … Inf	
2	D40a_MHC	0.0266	0.0136	-0.0130	0.94	-0.0268 … Inf	0.83
	D40b_OCS	0.0267	0.0135	-0.0133	0.94	-0.0271 … Inf	
	D41_BR	0.0263	0.0135	-0.0128	0.94	-0.0266 … Inf	
	D42_AN	0.0267	0.0142	-0.0125	0.92	-0.0268 … Inf	
	D43_MC	0.0263	0.0138	-0.0125	0.95	-0.0252 … Inf	
3	D46a_PFI	0.0207	0.0177	-0.0030	0.66	-0.0148 … Inf	0.83
	D46b_DS	0.0200	0.0168	-0.0032	0.68	-0.0146 … Inf	
	D49_DC	0.0201	0.0168	-0.0033	0.69	-0.0145 … Inf	
	D51_CHE	0.0206	0.0171	-0.0034	0.69	-0.0151 … Inf	
	D52_BH	0.0202	0.0169	-0.0033	0.68	-0.0148 … Inf	
4	D54_HT	0.0270	0.0145	-0.0125	0.94	-0.0256 … Inf	0.91
	D55_LD	0.0273	0.0144	-0.0128	0.94	-0.0260 … Inf	
	D56_RP	0.0279	0.0147	-0.0133	0.93	-0.0277 … Inf	
	D59_TTC	0.0282	0.0142	-0.0140	0.95	-0.0280 … Inf	
	D60a_UT	0.0280	0.0149	-0.0131	0.93	-0.0275 … Inf	0.83
5	D60b_GE	0.0305	0.0119	-0.0186	0.98	-0.0332 … Inf	
	D64_OMF	0.0301	0.0118	-0.0183	0.99	-0.0321 … Inf	
	D70_ED	0.0298	0.0119	-0.0179	0.98	-0.0318 … Inf	
	mean	0.0262	0.0145	-0.0117			
	sd	0.0034	0.0016	0.0049			
	SE	0.0007	0.0003	0.0010			

and distinctive terms of both author profiles, but based on the original relative frequency input matrix, which makes all weights comparable. Since clustering also indicates that separation ability of the two author sets on the basis of terms in both profiles is high, one may consider, whether the current evaluation actually correctly measures author profile distances.

In figure 4.2.1, we can observe the dendrogram computed on the basis of representative and distinctive terms of the 4th iteration for both authors, which shows no misclassifications and figure B.1.1 shows the dendrogram for *DickensCollinsSet2* also showing only one misclassified Collins document.

As a third measure, we consider profile consistency of both profiles over all iterations to evaluate how much agreement exists between different profiles of one author's set. Table 4.2.3 indicate a fair agreement for Dickens with a mean profile length of 354 terms and a profile intersection of 244 terms over 9 iterations. Table 4.2.4 reports a mean length of Collins of 336 terms, while profile agreement is on 209 terms.

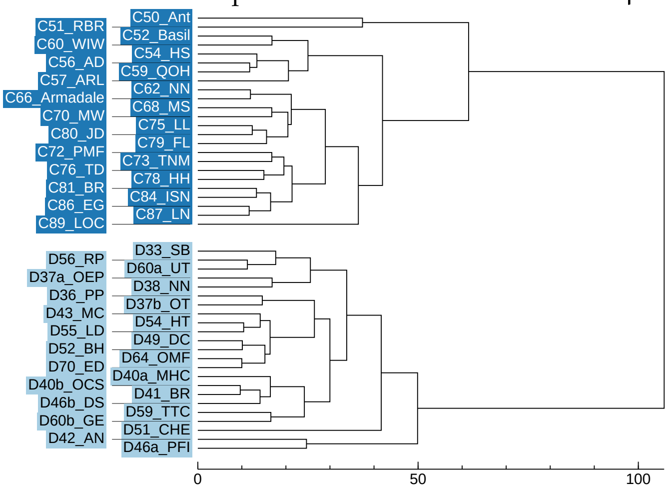
Profile consistency of the *DickensCollinsSet2* is shown in tables 4.2.5 and 4.2.6 and also here, agreement over both author profiles seems fair. The intersecting terms over all profiles are representative and and distinctive for the respective author, but not all of them are also frequent. This can also be observed in the fact, that both authors share a number of terms, such as *upon*, *first*, *such*, where there are probably large differences, either high frequency for a term in one author and low frequency for the other or vice versa.

If we consider the representative and distinctive and frequent terms for each author that form the basis for the profile evaluation, we observe a discrepancy with respect to the

Table 4.2.2: Representativeness & Distinctiveness on *DickensCollinsSet1*. Results of evaluating distances for profiles P_D and P_C to test closeness to Collins' documents also showing t-test results for hypothesis assuming greater mean for *Dickens* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis of shared representative and distinctive terms of both profiles.

Author Profile Comparison							Clustering
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
	C50_Ant	0.0332	0.0118	-0.0214	0.02	0.0046 ... Inf	
	C51_RBR	0.0320	0.0122	-0.0198	0.02	0.0037 ... Inf	
6	C52_Basil	0.0284	0.0110	-0.0174	0.02	0.0034 ... Inf	0.83
	C54_HS	0.0282	0.0114	-0.0168	0.02	0.0033 ... Inf	
	C56_AD	0.0285	0.0110	-0.0175	0.02	0.0035 ... Inf	
	C57_ARL	0.0289	0.0113	-0.0176	0.02	0.0035 ... Inf	
	C59_QOH	0.0287	0.0110	-0.0177	0.02	0.0037 ... Inf	
7	C60_WIW	0.0260	0.0138	-0.0122	0.06	-0.0010 ... Inf	0.91
	C62_NN	0.0266	0.0136	-0.0130	0.05	-0.0001 ... Inf	
	C66_Armadale	0.0265	0.0138	-0.0128	0.06	-0.0004 ... Inf	
	C68_MS	0.0254	0.0143	-0.0110	0.07	-0.0016 ... Inf	
	C70_MW	0.0265	0.0142	-0.0123	0.06	-0.0009 ... Inf	
8	C72_PMF	0.0362	0.0144	-0.0218	0.02	0.0038 ... Inf	0.83
	C73_TNM	0.0375	0.0143	-0.0232	0.02	0.0046 ... Inf	
	C75_LL	0.0368	0.0145	-0.0223	0.03	0.0036 ... Inf	
	C76_TD	0.0378	0.0140	-0.0237	0.02	0.0046 ... Inf	
	C78_HH	0.0369	0.0144	-0.0225	0.02	0.0043 ... Inf	
9	C79_FL	0.0359	0.0154	-0.0205	0.03	0.0022 ... Inf	0.83
	C80_JD	0.0359	0.0153	-0.0206	0.03	0.0024 ... Inf	
	C81_BR	0.0372	0.0150	-0.0223	0.02	0.0040 ... Inf	
	C84_ISN	0.0368	0.0152	-0.0216	0.03	0.0031 ... Inf	
	C86_EG	0.0366	0.0153	-0.0213	0.03	0.0024 ... Inf	
	mean	0.0321	0.0135	-0.0186			
	sd	0.0047	0.0016	0.0040			
	SE	0.0010	0.0003	0.0009			

Figure 4.2.1: *DickensCollinsSet1*. Clustering on representative and distinctive terms of 4th iteration with "complete link" method based on the 4th iteration.



separation of the terms. Both authors have *upon* as a high ranking term in their profiles, where for Collins the weight is 1.8 and for Dickens it is 1.5.

Even though *upon* is more representative for Collins it is allocated to Dickens' profile, because the term is generally more frequent for Dickens and we presume that a number of words follow this scheme. Additionally, the number of base terms in each profile is quite

Table 4.2.3: Profile consistency over nine iterations and 56 of the 244 intersecting representative and distinctive terms for Dickens based on the *DickensCollinsSet1*.

Dickens' most prominent markers
+upon, being, but, much, so, though, such, and
-first, discovered, produced, only,
left, resolution, future,
letter, words, attempt,
return, end, serious,
followed, wait, events, suddenly,
later, news, lines, advice,
absence, chance, written,
position, happened, placed,
enough, second, failed,
waited, hesitated, opened,
patience, questions, met, risk,
moment, result, offered, conduct,
inquiries, heard, entirely, speaking,
waiting, useless, discover

Iteration	Profile length	Terms after intersection
1	350	350
2	355	281
3	353	264
4	347	256
5	346	252
6	362	250
7	354	246
8	350	244
9	365	244
mean	354	265
std.	6	34
SE	2	11

Table 4.2.4: Profile consistency over nine iterations and 56 of the 249 intersecting representative and distinctive terms for Collins based on the *DickensCollinsSet1*.

Collins' most prominent markers
+first, only, discovered, left, produced, followed,
placed, return, words, resolution, end, second, to,
enough, attempt, suddenly
-upon, so, being,
such, very, much, many,
though, air, presently,
difficult, fire, leaning, but,
shaking, indeed, and, returned,
a, looking, indifferent,
would, busy, particularly,
brought, greater, beside, down,
pair, or, with,
bless, great, strong,
grown, usually, pretty, carried,
observed, like

Iteration	Profile length	Terms after intersection
1	343	343
2	338	289
3	341	265
4	344	254
5	351	232
6	330	222
7	322	218
8	329	213
9	328	209
mean	336	249
std.	9	44
SE	3	15

large, because only a few will be rated as frequent for an author compared to the opposing set and for a reliable profile evaluation a list of at least 40 terms is advisable. Therefore, we have to retrieve enough general terms to have sufficient terms for profile histogram evaluation, as otherwise it might not be a reliable result.

In this case, unfortunately, the lower ranked but more frequent terms will be given preference and ascend in the ranking, but these have not the same validity as more representative and distinctive though more infrequent terms. For this reason, the evaluation on the basis of representative and distinctive scores is unlikely to be an accurate measure of the *Representativeness & Distinctiveness* method and alternatives should be explored.

ICA on Dickens vs. Collins

Previous experiments showed that ICA performed better and more consistently, if the most frequent terms were excluded from the input terms, which motivated us to remove

Table 4.2.5: Profile consistency over 14 iterations and 56 of the 139 intersecting representative and distinctive terms for Dickens on *DickensCollinsSet*2.

Dickens' prominent markers
+much, upon, many, down
-wait, position, met, answer, answered, suddenly, waiting, interests, already, question, asked, person, view, speaking, resolution, future, enough, later, moment, experience, attempt, questions, discovered, waited, opened, spoke, produced, discovery, advice, leave, words, result, interval, servant, heard, hesitated, events, influence, absence, motive, speak, plainly, advanced, useless, discover, still, informed, failed, mind, woman, leaving, sudden

Iteration	Profile length	Terms after intersection
1	180	180
2	185	160
3	192	152
4	194	150
5	171	148
6	172	145
7	169	145
8	171	145
9	167	144
10	182	143
11	170	139
12	187	139
13	169	139
14	180	139
mean	178	148
sd	9	11
SE	2	3

Table 4.2.6: Profile consistency over 14 iterations and 56 of the 127 intersecting representative and distinctive terms for Collins on *DickensCollinsSet*2.

Collins' prominent markers
+followed, return, wait, under, attempt produced, answer, since, longer, discovered, leave, place, heard, already, hesitated, follow, possessed, placed, words, moment
-upon, many, much, very, down, indeed, such, being, great, though, heaven, deal, off, often, bless, like, always, fire, times, brought, air, company, , returned, carried, looking, full, wear, glad, hot, shoes, fact, observed, nor, although, bright, comfort

Iteration	Profile length	Terms after intersection
1	209	209
2	199	184
3	207	179
4	207	167
5	221	162
6	205	155
7	212	154
8	209	154
9	209	154
10	200	143
11	201	134
12	197	130
13	211	128
14	205	127
mean	207	156
sd	6	24
SE	2	6

the first 70 terms and compute the profiles on the remaining 1430 terms. The number of to-be-estimated components is set to 47 for the *DickensCollinsSet1*.

Components for documents are chosen by computing the mean activity of a component over all documents and retaining its contribution only in those documents where its activity is above its own mean. Similarly, we retain a term for a component when its activity in that component is higher than its mean activity over all components. The weights are combined and terms are retained for a profile, when they lie above $1.1 \times (\mu + sd)$, where mean and standard deviation is over the unrestricted profile. For the *DickensCollinsSet2*, we set the number of to-be-estimated components to 50 and discard terms in the original profile at a level of 1.0.

Tables B.2.1 and B.2.2 show the results for testing on Dickens' and on Collins' documents respectively. All of Dickens' documents are consistently rated closer to the Dickens' profile and significant p-values indicating a difference in individual profile means additionally support these findings. The paired t-test confirms model validity with a p-value of 0.02 and a positive confidence interval of 0.002 to *Inf*, that reliably recognizes Dickensian documents. Regarding Collins' test documents, except for two cases, all are correctly identified as belonging to Collins. In three cases, there is no significant difference in mean of the two individual profile distributions. Interestingly, the two misclassified cases $C50_{Ant}$ and $C51_{RBR}$ are exactly those that seemed to be outliers in the previous study of comparing Dickens and Collins (Tabata 2012). The overall model evaluation is favourable with a p-value of 0.02 and confidence interval of 0.0027 to *Inf*.

Table B.2.3 and table B.2.4 show the results for testing on the second dataset of Collins and Dickens, where testing on Dickens yields correct classifications with overall significant differences to the Collins profile in all except three cases, *D23344*, *D23765* and *DC1423*, which are two maybe slightly more unusual pieces of Dickens, namely *The Magic Fishbone* and *Captain Boldheart & the Latin-Grammar Master* and the shared Dickens and Collins piece, *No Thoroughfare*. Testing on unseen Collins pieces returns the same two misclassifications, *C238367* (*Rambles Beyond Railways*) and *C3606* (*Antonina*), but the paired t-test on the overall model performance yields significant values for both types of test documents.

Profile consistency over the different profiles are shown in table 4.2.7 and table 4.2.8, where Dickens' profiles have a mean length of 108 and agree on 55 terms over the nine iterations, while Collins' profiles have a mean length of 111 with 62 common terms.

Table 4.2.7: ICA. Profile consistency over nine iterations, showing 55 negative and positive features shared by all of Dickens profiles on *DickensCollinsSet*1.

Dickens 'positive and negative markers
+upon, its, down, great, much, being, come, such, though, like, then, many, old, where, says, never, returned, head, always, off, here, well, indeed
–question, mind, position, us, end, place, herself, looked, enough, way, before, still, tell, can, doctor, heard, answered, last, words, moment, next., asked, back, lucilla,oscar, own, left, letter, room, only, emily, first

Iteration	Profile length	Terms after intersection
1	109	109
2	122	87
3	112	75
4	108	65
5	100	55
6	104	55
7	104	55
8	104	55
9	106	55
mean	108	68
std.	6	19

Table 4.2.9 and table 4.2.10 show profile consistency for Dickens and Collins on the *DickensCollinsSet2*. With respect to the previous analysis of representative and distinctive terms, it is notable that certain terms continue to crop up in the two author profiles in both analyses, such as $upon$, $first$ and $such$. $Upon$, for instance is also rated frequent for Dickens and infrequent for Collins, as it has been done previously in the analyses of representative and distinctive terms.

Clustering is consistently at 0.83 for all iterations regardless of test document type. Figure 4.2.2 shows the dendrogram for clustering, which has two misclassifications and most fittingly these are the two Collins documents already conspicuous during profile evaluation. Figure B.2.1 shows the dendrogram for clustering on the basis of shared terms of iteration one for the second set with two misclassifications *C238367* and *DC1423*, that also appeared in profile evaluation. Clustering for the second set is less stable, but has

Table 4.2.8: ICA. Profile consistency over nine iterations and showing 62 negative and positive features shared by all Collins profiles on *DickensCollinsSet*1.

Collins 'positive and negative markers
+first, letter, only, asked, woman, room, looked, words, own, back, answered,left, still, moment, tell, myself, enough, can, husband, again, wife, door, mind, life, toward, spoke, heard, let, speak, answer, leave, marriage
–any, martin, returned,people, indeed, pecksniff, quite, good, off, every, where, like, tom, never, many, though, then, young, some, these, gentleman, such, much, its, sir, being, down, old, great, upon

Iteration	Profile length	Terms after intersection
1	115	115
2	110	108
3	109	103
4	108	98
5	104	87
6	102	74
7	115	68
8	120	64
9	114	62
mean	111	87
std.	6	20

Table 4.2.9: ICA. Profile consistency over 14 iterations, showing 56 of 103 negative and positive features shared by all of Dickens profiles on *DickensCollinsSet*2.

Dickens 'positive and negative markers
+many, upon, often, though, such, very, indeed, down, ago, much, round, heaven, bless, deal, off, bear, warm
–surprise, longer, possession, marriage, life, return, inquiries, turned, decided, excuse, entered, once, sudden, placed, serious, informed, failed, offered, anxiety, absence, feeling, servant, sense, impression, followed, feel, address, interval, interview, influence, explanation, visit, suspicion, result, person, plainly, spoke, hesitated, addressed

Iteration	Profile length	Terms after intersection
1	234	234
2	257	187
3	253	149
4	270	130
5	260	118
6	250	108
7	258	106
8	254	104
9	255	103
10	249	103
11	245	103
12	254	103
13	248	103
14	248	103
mean	252	125
sd	8	40

occasionally almost perfect separation according to the *adjusted Rand Index*. Thus, all tests were favourable with respect to the appropriateness of the identified characteristic markers.

ICA with Representative and Distinctive Components on Dickens vs. Collins

For the third model of ICA combined with *Representativeness & Distinctiveness*, we choose the same number of input terms, as for the previous ICA experiment, namely the 70 to 1500 most frequent features of the input matrix. Term selection for components also remains the same by selecting a term for a component if it lies above its individual mean activity over all components.

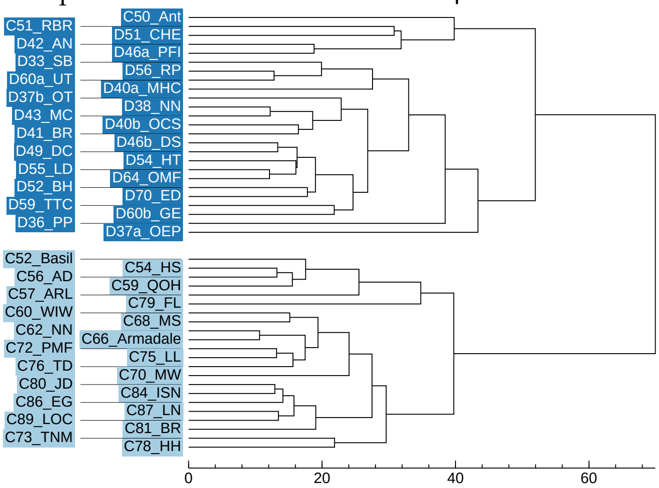
The selection of components for documents is different for this model, as we retain those components that are representative and distinctive for an author's set. The number of components are chosen according to a threshold multiplied by the mean over all representative and distinctive values over all components of a document set. This threshold is also dependent on the number of components, where extracting fewer components also

Table 4.2.10: ICA. Profile consistency over 14 iterations, showing 56 of 140 negative and positive features shared by all of Collins profiles on *DickensCollinsSet2*.

Collins 'positive and negative markers
+suddenly, wait, discovered, position, met, attempt, already, words, waiting, produced, enough, view, speaking, spoke, leave, answer, heard, moment, resolution,led, events, longer, later, motive, future, speak, absence, still, waited, silence, return, useless, suspicion, followed, possession, interests, stopped, discovery, placed, failed, influence
–times, heaven, better, glad, though, great, off, indeed, such, down, being, very, much, many, upon

Iteration	Profile length	Terms after intersection
1	242	242
2	246	239
3	240	231
4	243	230
5	246	230
6	243	228
7	247	228
8	248	227
9	247	227
10	247	207
11	227	177
12	240	158
13	246	154
14	240	140
mean	243	208
sd	5	35

Figure 4.2.2: *DickensCollinsSet1*. Clustering characteristic terms returned by ICA with "complete link" method based on the 4th iteration.



requires lowering the threshold. In this case, where we specify 47 components, we set the threshold to 1.0, which then effectively corresponds to the mean over all values.

For the second set, we chose 48 components, as previous tests showed this to yield good results and components are retained for each author set by at a level of 0.4× the mean over the representative and distinctive values for the components of an author's set. For both datasets, we retain terms for each profile by taking the mean plus the standard deviation over the complete profile. Similar to before weights are combined to form unique profiles over the document sets and terms in profiles are retained if their absolute weight is above the mean over absolute activity plus the standard deviation over the profile multiplied by scalar 1.3.

Table B.3.1 and table B.3.2 show the results for testing on unseen Dickens documents and Collins documents respectively. For all Dickens test documents, all differences of Dickens' profile to the Collins' profile are significant, thus the Dickens' profile is comfortably winning on all its documents. The t-test over mean differences is also highly significant with $p < 0.00001$ with a positive confidence interval of 0.0095 to *Inf*.

Regarding Collins' test documents in table B.3.2, we observe a similar development as in the previous ICA experiment. Two out of the three misclassified Collins documents are the documents, $C50_{Ant}$ and $C51_{RBR}$. In these three cases, p-values are obviously not significant with respect to the greater mean assumption for the Dickens' profile. The t-test for model evaluation is still significant with a confidence interval of 0.0067 to *Inf*.

Table B.3.3 and table B.3.4 show the results for the second set, where results are quite similar as for using separate ICA on the same input matrix, where also *D23344*, *D23765* and *DC1423* were misclassified. For unseen Collins pieces, again *C238367* and *C3606* are misclassified, while all others are consistently closer to the correct profile. Model evaluation for both types of unseen documents using a paired t-test yield significant differences with $p < 0.00001$ and a confidence interval of 0.0029 to *Inf* for unseen Dickens mean differences and $p < 0.0001$ with an 0.0037 to *Inf* interval for unseen Collins mean differences.

Table 4.2.11: Combined ICA & RD. Profile consistency over nine iterations, showing 22 negative and positive features shared by all of Dickens profiles on *DickensCollinsSet*1.

Dickens 'positive and negative markers
+upon, much, says, great, being, down, then, boffin, off
–moment, door, words, young, left, first, own, woman, only, letter, looked, room, answered

Iteration	Profile length	Terms after intersection
1	88	88
2	89	53
3	69	38
4	89	29
5	84	23
6	92	23
7	80	22
8	82	22
9	87	22
mean	84	36
std.	7	22

Table 4.2.12: Combined ICA & RD. Profile consistency over nine iterations, showing 33 negative and positive features shared by all of Collins' profiles on *DickensCollinsSet*1.

Collins 'positive and negative markers
+first, room, letter, own, left, only, looked, words, again, still, back, moment, spoke, let, door, heard, speak
–pickwick, like, every, off, its, where, such, many, being, some, gentleman, down, much, young, great, upon

Iteration	Profile length	Terms after intersection
1	75	75
2	80	65
3	70	54
4	84	52
5	83	50
6	85	46
7	84	40
8	83	36
9	81	33
mean	81	50
std.	5	14

Profile consistency for the profiles over different iterations are shown in table 4.2.11 and table 4.2.12. In this experiment, consistency is less for Dickens compared to the previous experiments, where of the mean profile length of 84 only 22 terms intersect on all profiles. Collins fares slightly better with a mean length of 81 and 50 intersecting terms. Tables 4.2.13

and 4.2.14 show profile consistency and intersecting terms for Dickens and Collins profile for the second set, where consistency is a little better.

Table 4.2.13: Combined ICA & RD. Profile consistency over 14 iterations, and showing 56 of 126 negative and positive features shared by all Dickens profiles on *DickensCollinsSet*2.

Dickens 'positive and negative markers
+upon, many, such, much, being,
deal, though, down, fact, nor,
great, times, rather, half, short,
glad, never, indeed, having, off,
less, heaven
–sense, marriage, feel,
silence, surprise, simply, approached, absence,
proved, addressed, life, spoken, servant,
possession, sadly, information,placed, serious,
living, hesitation, seriously, second, customary,
curiosity, interval, event, entered, inquiries,
promise, control, opened, speak, decided,
discover

Iteration	Profile length	Terms after intersection
1	243	243
2	252	192
3	245	169
4	258	144
5	246	142
6	243	133
7	258	130
8	250	128
9	247	127
10	235	127
11	240	127
12	246	127
13	252	127
14	245	126
mean	247	146
sd	6	34

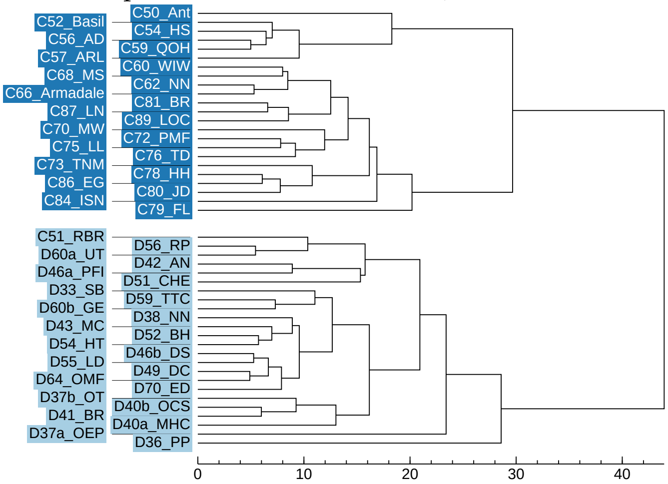
Table 4.2.14: Combined ICA & RD. Profile consistency over 14 iterations, showing 57 of 117 negative and positive features shared by all of Collins profiles on *DickensCollinsSet*1.

Collins 'positive and negative markers
+position, wait, second, answered, leave,
leaving, advice, view, questions, met,
words, discovered, reached, later, question,
answer, enough, offered, return, produced,
waiting, experience, asked, attempt, person,
heard, speak, waited, language, end,
suddenly, chance, risk, led, customary
– deal, like, bless, great, off,
times, fact, whole, where, going,
always, rather, indeed, down, never,
though, being, glad, such, many
upon, much

Iteration	Profile length	Terms after intersection
1	250	250
2	253	237
3	238	189
4	249	189
5	246	180
6	253	180
7	252	180
8	249	180
9	249	180
10	243	149
11	242	129
12	243	127
13	251	120
14	246	117
mean	247	172
sd	5	41

Notably, among the most characteristic reoccurring terms are again *upon*, *first such*, and *many*, which appeared also in previous experiments and thus seem to be good separators for the two authors. Clustering on the whole is occasionally slightly better than with using ICA separately. Figure 4.2.3 shows the dendrogram based on common terms of the 4th iteration, where one document of Collins is misclassified and ends up in the Dickens cluster. Figure B.3.1 shows the result of clustering on the 11th iteration for the second set, which shows one misclassified Dickensian document as belonging to Collins' documents, namely their shared piece, *DC1423*.

Figure 4.2.3: Clustering on combined ICA & RD characteristic terms on *DickensCollinsSet*1 with "complete link" method on the 4th iteration.



As a preliminary conclusion to the Dickens vs. Collins comparison, we note that using different datasets and different models, there are certain terms that consistently appear with respect to comparing Dickens' and Collins' documents. The evaluation on profile distances for *Representativeness & Distinctiveness* was not successful and unlikely to be suitable in this particular setting.

With respect to our investigation into style, in particular whether there exists something distinctly measurable corresponding to a unique *fingerprint*, our results are encouraging, as different methods show considerable overlap in discriminatory terms. Our shared Dickens and Collins piece was conspicuous twice in clustering which might indicate overlaps in style that condemn it to reside at the border of the two author sets.

However, as we shall see in the following comparison using a larger reference set to oppose Dickens' set, the previous rather salient markers in comparing to Collins somewhat disappear, which confirms earlier assumptions, that the reference set exerts a considerable amount of influence over the characteristic markers that are chosen for an author.

4.2.4 *Characteristic Terms of Dickens vs. World*

In this part, we evaluate our models with respect to a wider document set and contrast Dickens with a reference corpus comprised of 18th and 19th century texts by different contemporary authors of that time. For all experiments, we take a subset of the original 79×4895 matrix, namely again the first 1500 terms, leaving out the 70 most frequent ones.

Representative and Distinctive Terms of DickensWorldSet

For selecting representative and distinctive profiles using this larger document set, we choose representative and distinctive terms that lie above the threshold of $mean + sd$ for each complete list. Table B.4.1 and B.4.2 show the results for profile and clustering evaluation. Unfortunately, the t-test failed due to too few frequent terms in the *World* profiles. Although, there are good discriminators found in the set, these are not frequent for the *World* set and since we only choose frequent representative and distinctive terms for profile evaluation, the comparison is not possible. This is undoubtedly also the reason why Dickens is rated closer to almost all test documents, the Dickensian ones and the *World set* ones.

Generally the *World set* agrees more on infrequent, discriminatory items with respect to Dickens than on frequent average ones, which might also be influenced by the input term sample. Clustering is slightly erratic with some higher ranked iterations, but on average it is rather low with about 0.44 for the *adjusted Rand Index*.

Table 4.2.15: Profile consistency over 15 iterations and 122 intersecting representative and distinctive terms for Dickens

Dickens 'positive and negative markers
+until, looking, quiet, air,window, corner,head,
round, being, state, presented,hard, remarkable,
off, expression, again, moment, anything, position,
night, shake, lighted, behind, holding,
house, sound, anybody, glance,back
tight, eye, leaning
-given, use, till, return,
able, determined, advice, than, give,
temper, entirely, nor, must,
pleased, presence, only, things,
visit, received, without, cannot,
anxious, ashamed, therefore, however, judgment,
probably, affair, feel,
promise, understanding, accept, hardly,
reason, longer, felt,neither, feeling,
did, advantage, stay, too,
make, person, though, seeing,
immediately, wishes, obliged, order,
can, disagreeable, yet,
offer, fortune, nothing,
proposal, distress, account, possible,
produced, wished, appear, expect, greatest,
own, talked, almost, thus, desire,
necessity, need, confess, taste, discovered,
shall, talk, either, justice, also,
condition, attended, husbands, thing,
pain, pay,least, greatly, fit, possession

Iteration	Profile length	Terms after intersection
1	228	228
2	224	146
3	230	137
4	226	125
5	224	125
6	239	123
7	235	123
8	239	123
9	234	123
10	238	123
11	221	122
12	230	122
13	230	122
14	232	122
15	233	122
mean	231	132
sd	6	27
SE	1	7

When, considering profile consistency in table 4.2.15 and table 4.2.16, a curious phenomenon can be observed. Although Dickens' consistent terms do not include many body parts, they appear in plenty over the *World set*, e.g. *legs, faces, chin, face, heads*. However, consulting the list of frequent terms for both profiles, these are not frequent for the *World set* but still chosen for discrimination to the outside set, namely Dickens' set. The reason why they are not also listed for Dickens is that, even though single ones are highly rated on individual profiles, they do not appear among all of them, which is obviously a necessity to be among intersecting terms for an author.

ICA on Dickens vs. World

For this single ICA experiment, we chose a lower number of components, since previous trials showed an improved performance on the larger reference set. Thus, we set the number of to-be-extracted components to 50 and discard terms in the profile at a level of 1.0. Tables B.5.1 and B.5.2 show the results for testing on unseen Dickens and Collins documents. Except for one document, *D699*, all of Dickens' test documents are rated closer to the Dickens profile. Using a paired t-test on the overall model performance yields

Table 4.2.16: Profile consistency over 15 iterations and 115 intersecting representative and distinctive terms for the World

Worlds' positive and negative markers

+

-until, glancing, corner, head, smoking, legs, heavily, stopping, hat, dust, shaking, various, bar, staring, smoke, rubbing, tight, boys, lighted, faces, chin, state, glass, heres, returned, folded, chimney, remark, ashes, air, shining, pavement, heads, blue, staircase, mysterious, iron, red, boy, gloomy, shook, outside, gentleman, lying, ceiling, visitor, looking, window, crowd, inquired, streets, extent, floor, behind, asleep, whats, dark, devoted, reference, gradually, coat, spot, street, solitary, brick, roof, wall, bright, windows, gentlemen, arm, shadows, yard, door, cheerful, referred, knocked, visitors, breath, stairs, dull, softly, lights, hurried, wouldnt, pursued, sky, takes, round, through, repeated, beside, stopped, light, couldnt, expression,stare, thats, sits, clerks, shadow, breast, chair, hanging, night, hand, clock, asks, nod, leaves, office, chambers, awful, face, fire

Iteration	Profile length	Terms after intersection
1	165	165
2	163	128
3	186	122
4	169	116
5	195	116
6	174	115
7	185	115
8	185	115
9	181	115
10	177	115
11	172	115
12	178	115
13	174	115
14	181	115
15	176	115
mean	177	120
sd	9	13
SE	2	3

significant results for unseen Dickens documents with $p < 0.0001$ and a confidence interval of 0.0036 to *Inf*.

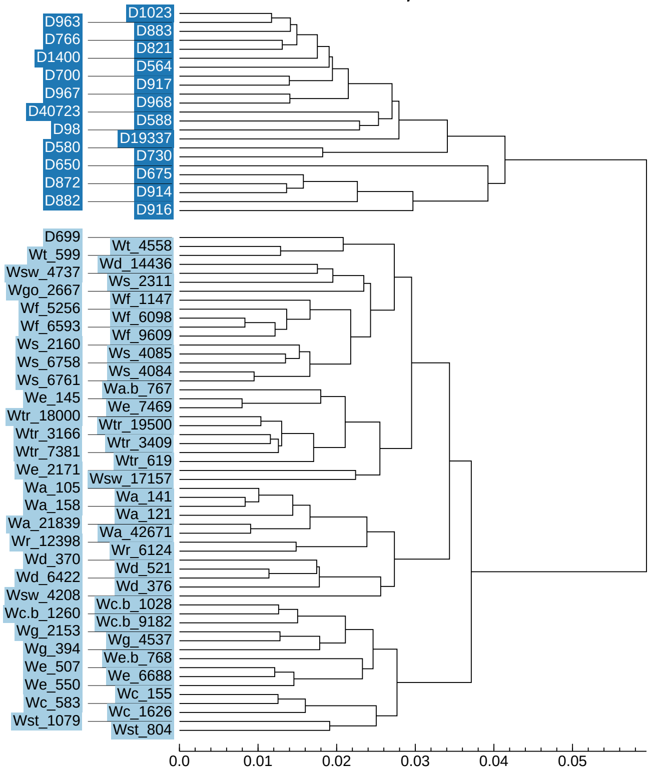
Considering the reference set documents as test instances is less favourable, as 17 out of 55 are rated closer to Dickens' profile, but the overall model evaluation is still significant with $p < 0.0001$ and a 0.00093 to *Inf* confidence interval, which is shifting more towards a possibly zero difference in mean between the two samples.

Clustering is fairly high on all the Dickens test documents and on most iterations of the *World set*, while it is also occasionally very low for the *World set* and this could maybe be explained by the fact that leaving out certain documents upsets a certain balance and different, less discriminatory features are selected. Figure 4.2.4 then shows the clustering result based on the 14th iteration with only one document of Dickens misclassified.

Tables 4.2.17 and 4.2.18 show the consistency level and the consistent features for both author sets over the 15 iterations. Dickens' consistency is fair with a mean profile length of 245 terms and 109 intersecting terms. The *World set* on the other hand has a mean length of 240 terms per profile and only 74 of them are constant over all iterations.

Regarding the terms in Dickens' profile, we observe a number of body parts, e.g. *legs, faces, head, hands, chin, arm* and *hair*. In addition, there seem to be a large number of scene-setting terms, such as *smoking, glancing, looking* and *shaking*, that could be used in collocations describing the background situation, which were reported elsewhere as seemingly characteristic of Dickens' style (Mahlberg 2007). The *World set* has only a few positive terms, but a large number of infrequent terms, which correspond to some extent to Dickens' frequent terms. Thus, it seems that there is generally more agreement on what should be infrequent on average than frequent, which somewhat indicates that Dickens had an unusual style for his time. Generally, since we are comparing to Dickens' set, there

Figure 4.2.4: ICA on clustering on characteristic terms on *DickensWorldSet* with "complete link" method based on iteration 14.



is likely to be a strong correlation of Dickensian documents that dominate in concepts, because there is bound to be more overlap between his documents. Since the reference set is made up of individual documents, these are maybe unlikely to agree strongly on a lot of terms, but rather agree that they are not close to common Dickens concepts.

ICA with Representative and Distinctive Components on Dickens vs. World

For the last experiment using ICA with feature selection on components, we again lower the number of components to 55 from 79 possible ones, so concepts are less spread out and components for a document set are chosen at an α level of 0.5. Again, we choose terms for profiles by taking the mean over the absolute values over the original profile and add the standard deviation.

Table B.6.1 and table B.6.2 show the results for testing on Dickens and *World* documents. The results are not as good as with the single ICA model, although the *World* test documents perform slightly better with only 14 out of 55 being misclassified. Both model evaluations using t-test are significant with p less than 0.0001 and a positive confidence interval of 0.0027 to *Inf* for Dickens' unseen documents and p less than 0.000 and a interval of 0.0011 to *Inf* for the *World* documents. However, clustering is highly irregular, even on Dickens' iterations, which were rather consistent with the ICA model. Figure 4.2.5 shows the clustering on the basis of one of the better profiles of iteration four, showing two misclassifications, namely *D699* and *D916*.

Considering consistency of terms over different iterations, using combined ICA and *Representativeness & Distinctiveness* seems also to perform slightly worse than the isolated ICA model. As shown in table 4.2.19, Dickens' mean profile length is 242 terms, but there

Table 4.2.17: ICA on *DickensWorldSet*. Feature Consistency over 15 iterations, showing 109 negative and positive features shared by all of Dickens profiles.

Dickens 'positive and negative markers
+stopping, smoking, glancing, window, legs, until, folded, head, glass, hat, various, bar, state, heavily, smoke, inquired, faces, asleep, corner, boys, air, boy, breath, night, rubbing, referred, red, table, behind, dust, remark, whispered, knocked, hot, round, looking, forth, ceiling, outside, floor, heres, visitors, reference, stopped, gloomy, hands, paper, key, again, lighted, breaking, chin, existence, office, wall, dark, arm, gradually, establishment,expression, staring, wet, softly, staircase, chair, through, shook, shaking, stars, lying, beside, hair, looked, hard, wouldnt, brick, bright, iron, bird, ashes, devoted, light, another, returned, bottle
–however, desire, fortune, only, obliged, ashamed, necessary, did, visit, entirely, return, receive, though, regard, own, make, use, than, judgment, advice, longer, seeing, give, given

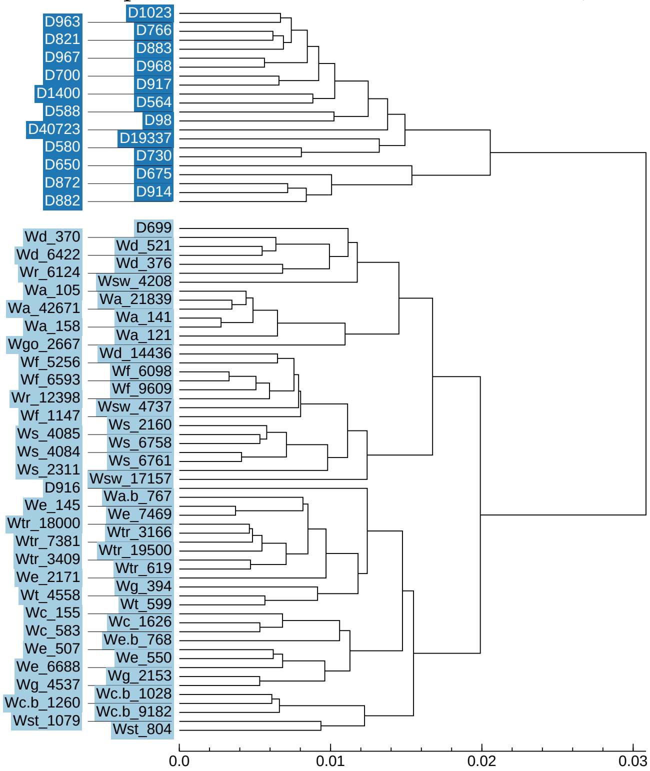
Iteration	Profile length	Terms after intersection
1	246	246
2	225	150
3	250	128
4	240	111
5	252	109
6	250	109
7	247	109
8	244	109
9	250	109
10	244	109
11	244	109
12	247	109
13	245	109
14	247	109
15	250	109
mean	245	122
sd	6	36

Table 4.2.18: ICA on *DickensWorldSet*. Feature Consistency over 15 iterations, showing 74 negative and positive features shared by all of World profiles.

Worlds 'positive and negative markers
+given, give, return
–opposite, inquired, boys, shining, yard, thats, outside, through, drinking, shadow, blue, whats, lying, bottle, ceiling, softly, stare, wall, lights, gloomy, gentleman, floor, breath, chimney, glass, boy, asleep, red, behind, eyed, repeated, remark, spot, staircase, shook, forth, hair, couldnt, shake, rubbing, state, various, faces, upright, lighted, looking, folded, mysterious,touching, smoke, chin, round, wouldnt, window, glancing, hat, gradually, heres, arm, shaking, legs, air, bar, stopping, smoking, dust, staring, corner, head, until, heavily

Iteration	Profile length	Terms after intersection
1	241	241
2	248	227
3	243	225
4	240	219
5	240	210
6	246	159
7	224	124
8	244	111
9	233	98
10	239	95
11	249	90
12	230	87
13	246	81
14	245	78
15	239	74
mean	240	141
sd	7	65

Figure 4.2.5: Clustering on combined ICA & RD characteristic terms on *DickensWorldSet* with "complete link" method based on iteration 4.



is only agreement on 84 of them. Table 4.2.20 shows that the *World* set has a mean length of 235 and with only 26 common terms. Regarding the shared and consistent terms, there is reasonable overlap with the previous two experiments considering body parts and scenic elements.

ICA models seem to do less well on mixed sets involving a variety of authors and seem to have difficulties in finding similarities in terms of joined characteristic deviations. Moreover, the set is ordered and documents of the same author are often extracted at the same time for testing. In order to investigate, whether this might be a factor, one could repeat the experiment with *leave-one-out* cross-validation. This second experiment using a larger reference set has clearly shown that the composition of the reference set is vital for detecting the *desirable* discriminatory elements of an author.

For Dickens, we obtain a large number of body parts, even in intersection of his profiles, as well as scenic elements that he might use for ongoing characterisation. For the *World set*, agreement is rather on average infrequent terms, i.e. absence of particular terms, such as body parts, than what is common for that time. However, since we we left out the 70 most frequent terms, this might influence the result as well in terms of size and which final terms are chosen for the *World set*.

In conclusion to this comparison, we tentatively note that isolated ICA performed best on the *World set*, although since we have not exhausted all parameter combinations, the combined model may also perform more consistently given another setting. Unfortunately, *representative & distinctive* terms could not be evaluated correctly here, but this last comparison finally confirmed the unsuitability of the current evaluation for *Representativeness & Distinctiveness* and identified the need to evaluate profile distances in a different way, which does not necessarily require frequent markers. There is considerable overlap in

Table 4.2.19: Combined ICA & RD on *DickensWorldSet*. Profile consistency over 15 iterations, showing 84 negative and positive features shared by all of Dickens profiles.

Dickens 'positive and negative markers
+legs, looking, until, stopping, head, dust, arm, smoking, smoke, asleep, outside, glass, slowly, folded, eyed, hat, round, heavily, glancing, shaking, behind, bar, dark, window, staring, roof, reference,through, heres, door, couldnt, visitor, hands, knocked, lying, breath, floor, wall, shoulder, staircase,faces, iron, whats, wouldnt, face, tight, hand, eyes, confused, leaning, stopped, awful, holding, light, turned, top, thats, corner, stare, brick, shake, turning, heavy, hair, lighted, air, shook, chair, show, putting, beside, table, ceiling
–least, than, return, only, visit, pleased, able, entirely, till, though, given

Iteration	Profile length	Terms after intersection
1	247	247
2	227	136
3	241	108
4	236	96
5	243	89
6	238	86
7	236	86
8	222	86
9	241	84
10	242	84
11	242	84
12	242	84
13	242	84
14	242	84
15	242	84
mean	239	101
sd	7	43

Table 4.2.20: Combined ICA & RD on *DickensWorldSet*. Profile consistency over 15 iterations, showing 26 negative and positive features shared by all World profiles.

Worlds 'positive and negative markers
enough, beginning, expected, really, possible, wanted, might, before, living, relief, surprise, difficulty,talk
–takes, gentleman, bar, asleep, midst, hearts, bottle, becomes, drink, heres, knows, fellow, ears

Iteration	Profile length	Terms after intersection
1	242	242
2	246	189
3	241	168
4	248	162
5	248	145
6	231	94
7	263	68
8	232	42
9	236	37
10	235	26
11	235	26
12	235	26
13	235	26
14	235	26
15	235	26
mean	240	87
sd	8	74

characteristic terms identified by our different models, as we can see a marked appearance of scene-setting elements and a large number of body parts.

4.3 Discussion and Interpretation of Characteristic Term Results

The experiments conducted in this work are still only tentative, as we have not exhausted all possibilities with respect to all combinations of parameters. What can be said in general is that the choice of characteristic terms of an author by a particular model are highly dependent on the comparison set. When opposing Dickens and Collins, we obtain mainly markers that seem well able to separate those two authors. This does not mean that the terms are necessarily very characteristic for each author individually, but they become characteristic when both Dickens and Collins are compared.

All methods showed difficulty in finding common frequent terms of the *World set*. What is also interesting is the extent to which the *Representativeness & Distinctiveness* model and the ICA-based models seem to agree on certain characteristic markers, given the fact that one method is supervised and the other unsupervised.

4.3.1 *Comparing to Tabata's Random Forests*

One objective of this study was to compare to the previous work of Tabata 2012 using *Random Forests* classification also comparing Dickens to Collins and the larger reference set of the 18th/19th century.

Random forests classification *Random Forests* were first introduced in Breiman 2001 and are based on ensemble learning from a large number of decision trees. Like common decision trees, *Random Forests* can be used for both classification and regression, but with the additional advantages of ensemble learning through combining different individual models.

Building a forest of decision trees is based on different attributes in the nodes, where attributes at each node are chosen with respect to information gain that support classification. Different trees have access to a different random subset of the feature set. Given a training corpus, a different subset of this training data is selected with replacement to train each tree, while the remainder is used to estimate error and variable importance. The fact that variable importance is provided alongside the result makes it suitable for style analysis, where not only the decision is of importance, but also the motivation that led to this decision.

Comparing characteristic terms of Dickens vs. Collins In order to recall, which terms where identified by Tabata 2012, these are again displayed in table 4.3.1 and table 4.3.4. We generally compare to the consistent terms of the intersections identified by our models. For *Representativeness & Distinctiveness*, we use only the frequent terms for each author. The terms for ICA and ICA combined with representative and distinctive components are taken from the respective intersecting features over all iterations. All tables show the intersections on terms of our models and the ones described in Tabata 2012, given the same input matrices. As we can observe from table 4.3.2 and table 4.3.5, there is fair agreement for the first model, but even more agreement when comparing to single ICA terms, as shown in table 4.3.3 and table 4.3.6. For this comparison. we leave out terms for ICA and representative and distinctive components, since there were only few common terms, which seemed thus less interesting.

Comparing characteristic terms of Dickens vs. World Tabata's terms are shown in table 4.3.7 and our results for the three models in table 4.3.8, table 4.3.9 and table 4.3.10 respectively. There is considerably less overlap for all models, with the first model still sharing most terms.

Table 4.3.1: Dickens' markers, when compared to Collins according to Tabata's work using Random Forests.

Dickens' markers

very, many, upon, being, much, and, so, with, a, such, indeed, air, off, but, would, down, great, there, up, or, were, head, they, into, better, quite, brought, said, returned, rather, good, who, came, having, never, always, ever, replied, boy, where this, sir, well, gone, looking, dear, himself, through, should, too, together, these, like, an, how, though, then, long, going, its

Table 4.3.2: Intersection of Dickens' markers according to Representativeness & Distinctiveness and Tabata's Dickens' markers on the Collins' comparison.

Dickens' markers

upon, being, but, so, though, much, such, and, with, very, off, up, down, a, then, many

Table 4.3.3: Intersection of Dickens' markers returned by ICA and Tabata's Dickens' markers on the Collins' comparison.

Dickens' markers

upon, its, down, great, much, being, such, though, like, then, many, where, never, returned, head, always, off, well, indeed

Table 4.3.4: Collins' markers, when compared to Dickens according to Tabata's work using Random Forests.

Collins' markers

first, words, only, end, left, moment, room, last, letter, to, enough, back, answer, leave, still, place, since, heard, answered, time, looked, person, mind, on, woman, at, told, she, own, under, just, ask, once, speak, found, passed, her, which, had, me, felt, from, asked, after, can, side, present, turned, life, next, word, new, went, say, over, while, far, london, don't, your, tell, now, before

Table 4.3.5: Intersection of Collins' markers yielded by Representativeness & Distinctiveness and Tabata's Collins' markers.

Collins' markers

first, only, left, words, end, to, enough, heard, letter, moment, answer, leave, on, looked, since, under, passed, place, felt, had

Table 4.3.6: Intersection of Collins' markers according to ICA and Tabata's Collins' markers.

Collins' markers

first, letter, only, asked, woman, room, looked, words, own, back, answered, left, still, moment, tell, enough, can, mind, life, heard, speak, answer, leave

Table 4.3.7: Tabata's Dickens markers, when compared to the reference corpus.

Positive Dickens' markers

eyes, hands, again, are, these, under, right, yes, up, sir, child, looked, together, here, back, it, at, am, long, quite, day, better, mean, why, turned, where, do, face, new, there, dear, people, they, door, cried, in, you, very, way, man

Negative Dickens' markers

lady, poor, less, of, things, leave, love, not, from, should, can, last, saw, now, next, my, having, began, our, letter, had, I, money, tell, such, to, nothing, person, be, would, those, far, miss, life, called, found, wish, how, must, more, herself, well, did, but, much, make, other, whose, as, own, take, go, no, gave, shall, some, against, wife, since, first, them, word

In conclusion, there seems to be considerably more overlap between our terms and Tabata's results on the first comparison for Dickens and Collins and the ICA model seems

Table 4.3.8: Intersection of Dickens markers according to Representativeness & Distinctiveness and Tabata's set, when compared to the reference corpus.

Positive Dickens' markers
again, back, must,did,make,own,shall
Negative Dickens' markers
things,can, nothing,person

Table 4.3.9: Intersection of Dickens markers according to ICA and Tabata's set, when compared on the reference corpus.

Positive Dickens' markers
hands,again, looked
Negative Dickens' markers
did, make, own

Table 4.3.10: Intersection of Dickens markers according to combined ICA & RD and Tabata's set, when compared to the reference corpus.

Positive Dickens' markers
door,hands, face, eyes,turned
Negative Dickens' markers

to agree even more with Tabata's terms than the representative and distinctive terms selected. For the *World* set, there is considerably less overlap, which might be attributed to the different samples of input terms or also the possibility that a more inconsistent set, such as the reference set here affects our models a lot differently than two fairly coherent author sets. What is notable is that our methods, especially the ICA model return more body parts and terms than Tabata's analysis, which could be part of frequent collocations, such as *staring*, *looking*, *glancing*, *smoking* that could form part of Dickensian background ongoing characterisation that was already identified previously.

4.3.2 *Towards a More Suitable Evaluation for Representative and Distinctive Terms*

As has been shown during this work, the current evaluation scheme is not suitable for characteristic terms chosen by the *Representativeness & Distinctiveness* measure. Even when only the more frequent items of the profiles are chosen, is it unlikely that the values correspond directly to relative frequencies used in evaluation. Since the method is supervised, we cannot evaluate on the weights directly as in the ICA evaluation. Moreover, representative and distinctive values are calculated over a number of documents and given a single test document, we could not achieve the same result. Thus, we propose evaluation of representative and distinctive terms on the basis of their respective representative values.

Given a representative and distinctive profile, containing a number of individual terms for an author, we select only the representative values for those terms. Also, we obtain another rival profile for comparison, also containing a number of other individual terms and select their representative values. In addition, we take a test corpus containing a sufficient number of documents of the author under investigation. The assumption is, that representative and distinctive terms for an author should also be closer in representativeness to the test corpus than the rival author. Another basic assumption is that the test corpus is large enough to detect representative terms. Thus, we calculate a representative profile for the test corpus on the basis of profile terms for author A and then author B. We then calculate histogram differences between the test corpus and both author profiles and compare closeness.

As a representative value does not reveal whether a term is frequent or less frequent for an author a disadvantage of this method is the need for a test corpus rather than a single test document, but values need to be calculated on the basis of comparison between different documents of an author. Naturally, this approach should be subjected to analysis and close scrutiny before admittance as a reliable evaluation scheme. However, if valid this would provide comparison based on comparable values, which would provide a more reliable method of evaluating representative and distinctive terms.

Contribution and Open Ends

The present work was yet tentative and exploratory and aimed at investigating *Representativeness & Distinctiveness* and *Independent Component Analysis* for characteristic term extraction in authorship attribution.

In the process, we made attempts at developing evaluation methods for non-traditional stylometry that combined provide some measure of the degree of reliability and validity of the chosen characteristic terms. The measure of profile consistency could be further extended to exactly measure the degree of consistency taking into account profile length and number of profiles intersected.

In addition, one might consider different types of input features, such as part-of-speech tags to our current models. Also, it might be worthwhile to further investigate the influence of the exact composition of the authorship sets on the selected characteristic terms. While different subsets of an author's work seem to yield similar sets of markers, the opposing set seems to considerably influence the terms that are chosen for discrimination.

Dickens is said to be an unusual writer compared to his contemporaries, but for ascertaining general applicability of the proposed methods to authorship attribution it may be worthwhile conducting similar studies with other maybe less unusual authors to determine to what extent an individual style can still be detected.

Overall, our results are generally encouraging insofar as to suggest that there is in fact something consistent and detectable with respect to style in Dickens and that the presented methods should be further explored to improve results according to the criteria developed in this study.

5 CONCLUSION AND FUTURE WORK

This thesis was an investigation into Dickens' style using two statistical measures to extract some salient features of the author. Apart from actually extracting style markers, we were also concerned about important characteristics of the results, such as discrimination and separation ability as well as consistency of discriminators given different subsets of an author's set. We also found indications that for most methods the composition of the reference set is vital for selection of representative characteristic terms.

If our findings with respect to Dickens can be generalized to authors in general, results strongly indicate that there is something of a measurable style in the writings of an author. These findings may even overlap with studies using different approaches, which additionally support their validity and general applicability. Thus, different *prisoners* using different methods arrived at similar conclusions.

The present study could not give justice to all aspects of the problem, but hopefully convincingly showed that the presented methods could be beneficial for stylometry. In order to draw more definite conclusions from the results, the presented statistical methods and evaluation schemes require consolidation.

Nevertheless, in conducting this study, we should at least have succeeded in letting some light into the *cave* of style analysis, so *shapes* will be better visible.

> " It was further assumed that, owing to the well-known persistence of unconscious habit, *personal peculiarities* in the construction of sentences, in the use of long or short words, in the number of words in a sentence, etc., will in the long run manifest themselves with such regularity that their graphic representation may become a means of identification, at least by exclusion."
>
> *- Mendenhall 1901*

BIBLIOGRAPHY

[1] Harald Baayen et al. *JADT 2002: 6 es Journées internationales d'Analyse statistique des Données Textuelles An experiment in authorship attribution*. 2002.

[2] Kyungim Baek et al. "PCA vs. ICA: a comparison on the FERET data set". In: *in Proc. of the 4th International Conference on Computer Vision, ICCV'02*. 2002, pp. 824–827.

[3] Kirk Baker. "Singular Value Decomposition Tutorial". 2005.

[4] R. A. Bosch and J. A. Smith. "Separating hyperplanes and the authorship of the disputed federalist papers." In: *American Mathematical Monthly* 105.7 (1998), pp. 601–608.

[5] Leo Breiman. "Random Forests". In: *Machine Learning*. 2001, pp. 5–32.

[6] J. F. Burrows. "Not unless you ask nicely: The interpretative nexus between analysis and information." In: *Literary and Linguistic Computing* 7 (1992), pp. 91–109.

[7] John Burrows. "All the Way Through: Testing for Authorship in Different Frequency Strata." In: *LLC* 22.1 (2007), pp. 27–47.

[8] John Burrows. "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship". In: *Literary and Linguistic Computing* 17.3 (Sept. 2002), pp. 267–287.

[9] Alexander Michael Simon Clark. "Forensic Stylometric Authorship Analysis under the Daubert Standard". In: *Journal of Law & Literature eJournal* (2011).

[10] Paul Clough. "Old and new challenges in automatic plagiarism detection". In: *National Plagiarism Advisory Service, 2003*. 2003, pp. 391–407.

[11] Pierre Comon. "Independent component analysis, a new concept?" In: *Signal Process.* 36.3 (Apr. 1994), pp. 287–314.

[12] Hugh Craig and John Drew. "Did Dickens write "Temperate Temperance"?: (An Attempt to Identify Authorship of an Anonymous Article in All the Year Round)". In: *Victorian Periodicals Review* 44 (3 2011), pp. 267–290.

[13] Glenn Fung, Olvi Mangasarian, and John Jay. "The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization". In: *in 'Proc. 2003 Conf. on Diversity in Computing', ACM*. ACM Press, 2003, pp. 42–46.

[14] Ramyaa Congzhou He and Khaled Rasheed. "Using Machine Learning Techniques for Stylometry". In: *IC-AI*. Ed. by Hamid R. Arabnia and Youngsong Mun. CSREA Press, 2004, pp. 897–903.

[15] Jeanny Herault and Christian Jutten. "Space or time adaptive signal processing by neural network models". In: *AIP conference proceedings* 151 (1986), p. 206.

[16] Timo Honkela and Aapo Hyvärinen. "Linguistic Feature Extraction using Independent Component Analysis". In: *Proceedings of IJCNN'04*. Budabest, Hungary, 2004, pp. 279–284.

[17] David L. Hoover. "The Rarer They Are, the More There Are, the Less They Matter". In: *Proceedings of the Digital Humanities 2012*. DH2012. 2012.

[18] Lawrence Hubert and Phipps Arabie. "Comparing partitions". In: *Journal of Classification* 2.1 (Dec. 1985), pp. 193–218.

[19] Aapo Hyvärinen and Erkki Oja. "Independent component analysis: algorithms and applications". In: *Neural Networks* 13 (2000), pp. 411–430.

[20] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2009.

[21] S.K. Khamis. "Inference and Disputed Authorship: The Federalist by F. Mosteller; D. L. Wallace". In: *Review of the International Statistical Institute* 34.2 (1966), pp. 277–279.

[22] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. "Computational methods in authorship attribution". In: *J. Am. Soc. Inf. Sci. Technol.* 60.1 (Jan. 2009), pp. 9–26.

[23] Michaela Mahlberg. "Clusters, key clusters and local textual functions in Dickens." In: *Corpora* 2.1 (2007), pp. 1–31.

[24] M. B. Malyutov. *Authorship attribution of texts: a review*. 2005.

[25] T. C. Mendenhall. "The characteristic curves of composition". In: *Science* 214S (1887), pp. 237–246.

[26] T.C. Mendenhall. "A Mechanical Solution of a Literary Problem." In: *The Popular Science Monthly* 60 (1901), pp. 97–105.

[27] Frederick Mosteller and David L. Wallace. *Inference and Disputed Authorship: The Federalist*. New Ed. The David Hume Series of Philosophy and Cognitive Science Reissues. Center for the Study of Language and Information, Dec. 2008.

[28] Plato and B. Jowett. *The Republic*. The Modern library of the world's best books. Modern Library, 2011.

[29] Jelena Prokić, Çağri Çöltekin, and John Nerbonne. "Detecting shibboleths". In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. EACL 2012. Avignon, France: Association for Computational Linguistics, 2012, pp. 72–80.

[30] Jonathon Shlens. *A Tutorial on Principal Component Analysis*. Tech. rep. Systems Neurobiology Laboratory, Salk Insitute for Biological Studies, Dec. 2005.

[31] Efstathios Stamatatos. "A survey of modern authorship attribution methods". In: *J.Am.Soc.Inf.Sci.Technol.* 60.3 (Mar. 2009), pp. 538–556.

[32] Tomoji Tabata. "Approaching Dickens' Style through Random Forests". In: *Proceedings of the Digital Humanities 2012*. DH2012. 2012.

[33] Jaakko J. Väyrynen, Lasse Lindqvist, and Timo Honkela. "Sparse Distributed Representations for Words with Thresholded Independent Component Analysis". In: *IJCNN*. 2007, pp. 1031–1036.

[34] Brian Vickers. "Shakespeare and Authorship Studies in the Twenty-First Century". In: *Shakespeare Quarterly* 62.1 (2011), pp. 106–142.

[35] G. S. Watson. "Inference and Disputed Authorship: The Federalist. by Frederick Mosteller; David L. Wallace". English. In: *The Annals of Mathematical Statistics* 37.1 (1966), pp. 308–312.

[36] G. K. Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, 1932.

A

AUTHORSHIP DATASETS

A.1 DICKENS VS. COLLINS DATA SET (2)

Table A.1.1: Dickens' augmented set for second comparison as part of *DickensCollinsSet2*.

No.	Author	Texts	Abbr.	Tabata label
1	Dickens	Bleak House	D1023	D52_BH
2	Dickens	Great Expectations	D1400	D60b_GE
3	Dickens	Little Dorrit	D963	D55_LD
4	Dickens	David Copperfield	D766	D49_DC
5	Dickens	A Christmas Carol	D19337	D43b_CB
6	Dickens	Life And Adventures Of Martin Chuzzlewit	D968	D43_MC
7	Dickens	The Mystery of Edwin Drood	D564	D70_ED
8	Dickens	A Tale of Two Cities	D98	D59_TTC
9	Dickens	Master Humphrey's Clock	D588	D40a_MHC
10	Dickens	The Battle of Life: A Love Story	D40723	D43b_CB
11	Dickens	The Life And Adventures Of Nicholas Nickleby	D967	D38_NN
12	Dickens	Barnaby Rudge	D917	D41_BR
13	Dickens	Sketches of Young Couples	D916	D37a_OEP
14	Dickens	The Uncommercial Traveller	D914	D60a_UT
15	Dickens	Our Mutual Friend	D883	D64_OMF
16	Dickens	Pictures From Italy	D650	D46a_PFI
17	Dickens	Sketches by Boz	D882	D33_SB
18	Dickens	A Child's History of England	D699	D51_CHE
19	Dickens	Reprinted Pieces	D872	D56_RP
20	Dickens	Dombey and Son	D821	D46b_DS
21	Dickens	Oliver Twist	D730	D37b_OT
22	Dickens	The Old Curiosity Shop	D700	D40b_OCS
23	Dickens	American Notes	D675	D42_AN
24	Dickens	The Pickwick Papers	D580	D36_PP
25	Dickens	The Letters of Charles Dickens: Vol. 1	D25852	-
26	Dickens	The Letters of Charles Dickens: Vol. 2	D25853	-
27	Dickens	The Letters of Charles Dickens: Vol. 3	D25854	-
28	Dickens	Mrs. Lirriper's Lodgings	D1416	-
29	Dickens	Captain Boldheart & the Latin-Grammar Master	D23765	-
30	Dickens	The Seven Poor Travellers	D1392	-
31	Dickens	Doctor Marigold	D1415	-
32	Dickens	The Holly-Tree	D1394	-
33	Dickens (et al.)	A Budget of Christmas Tales	Dal28198	-
34	Dickens	The Perils of Certain English Prisoners	D1406	-
35	Dickens	A Message from the Sea	D1407	-
36	Dickens	Somebody's Luggage	D1414	-
37	Dickens	Mugby Junction	D1419	-
38	Dickens	Mrs. Lirriper's Legacy	D1421	-
39	Dickens	The Wreck of the Golden Mary	D1465	-
40	Dickens	The Cricket on the Hearth	D20795	-
41	Dickens	Mugby Junction	D27924	-
42	Dickens	The Magic Fishbone	D23344	-
43	Dickens	Charles Dickens' Children Stories	D37121	-
44	Dickens (et al.)	A House to Let	Dal2324	-
45	Dickens(/Collins)	No Thoroughfare	DC1423	-

Table A.1.2: Collins' augmented set for second comparison as part of *DickensCollinsSet2*.

No.	Author	Texts	Abbr.	Tabata label
1	Collins	After Dark	C1626	C56_AD
2	Collins	Antonina	C3606	C50_Ant
3	Collins	Armadale	C1895	C66_Armadale
4	Collins	Man and Wife	C1586	C70_MW
5	Collins	Little Novels	C1630	C87_LN
6	Collins	Jezebel's Daughter	C3633	C80_JD
7	Collins	I Say No	C1629	C84_ISN
8	Collins	Hide and Seek	C7893	C54_HS
9	Collins	Basil	C4605	C52_Basil
10	Collins	A Rogue's Life	C1588	C57_ARL
11	Collins	The Woman in White	C583	C60_WIW
12	Collins	The Two Destinies	C1624	C76_TD
13	Collins	The Queen of Hearts	C1917	C59_QOH
14	Collins	The New Magdalen	C1623	C73_TNM
15	Collins	The Moonstone	C155	C68_MS
16	Collins	The Legacy of Cain	C1975	C89_LOC
17	Collins	The Law and the Lady	C1622	C75_LL
18	Collins	The Haunted Hotel: A Mystery of Modern Venice	C170	C78_HH
19	Collins	The Fallen Leaves	C7894	C79_FL
20	Collins	The Evil Genius	C1627	C86_EG
21	Collins	No Name	C1438	C62_NN
22	Collins	Poor Miss Finch	C3632	C72_PMF
23	Collins	Rambles Beyond Railways	C28367	C51_RBR
24	Collins	The Black Robe	C1587	C81_BR
25	Collins	Miss or Mrs.?	C1621	-
26	Collins	My Lady's Money	C1628	-
27	Collins	The Dead Alive	C7891	-
28	Collins	The Frozen Deep	C1625	-
29	Collins	The Guilty River	C3634	-

A.2 DICKENS VS. WORLD DATA SET

Table A.2.1: 18th century reference corpus to oppose Dickens as part of the *DickensWorldSet*.

No.	Author	Texts	Abbr.	Date	Word-token
1	Defoe	Captain Singleton	Wd_6422	1720	110,916
2	Defoe	Journal of Prague year	Wd_376	1722	83,494
3	Defoe	Military Memoirs of Capt. George Carleton	Wd_14436	1728	80,617
4	Defoe	Moll Flanders	Wd_370	1724	138,094
5	Defoe	Robinson Crusoe	Wd_521	1719	232,453
6	Fielding	A journey from this world to the next	Wf_1147	1749	45,024
7	Fielding	Amelia	Wf_6098	1751	212,339
8	Fielding	Jonathan Wild	Wf_5256	1743	70,086
9	Fielding	Joseph Andrews I&II	Wf_9609	1742	126,342
10	Fielding	Tom Jones	Wf_6593	1749	347,219
11	Goldsmith	The Vicar of Wakefield	Wgo_2667	1766	63,076
12	Richardson	Clarrissa I - IX	Wr_12398	1748	939,448
13	Richardson	Pamela	Wr_6124	1740	439,562
14	Smollett	Peregrine Pickle	Ws_4084	1752	330,557
15	Smollett	Travels through France and Italy	Ws_2311	1766	121,032
16	Smollett	The Adventures of Ferdinand Count Fathom	Ws_6761	1753	157,032
17	Smollett	Humphrey Clinker	Ws_2160	1771	150,281
18	Smollett	The Adventures of Sir Launcelot Greaves	Ws_6758	1760	89,010
19	Smollett	The Adventures of Roderick Random	Ws_4085	1748	191,539
20	Sterne	A Sentimental Journey	Wst_804	1768	41,028
21	Sterne	The Life and Opinions of Tristram Shandy	Wst_1079	1759-67	184,428
22	Swift	A Tale of a Tub	Wsw_4737	1704	44,225
23	Swift	Gulliver's Travels	Wsw_17157	1726	103,806
24	Swift	The Journal to Stella	Wsw_4208	1710-3	191,740
					sum: 4,493,348

Table A.2.2: 19th century reference corpus to oppose Dickens set as part of the *DickensWorld-Set*.

No.	Author	Texts	Abbr.	Date	Word-token
1	Bronte, A.	Agnes Grey	Wa.b_767	1847	68,352
2	Austen	Emma	Wa_158	1815	160,899
3	Austen	Mansfield Park	Wa_141	1814	159,921
4	Austen	Pride and Prejudice	Wa_42671	1813	121,874
5	Austen	Northanger Abbey	Wa_121	1803	77,810
6	Austen	Sense and Sensibility	Wa_21839	1811	119,793
7	Austen	Persuasion	Wa_105	1816 (1818)	83,380
8	Bronte, C.	The Professor	Wc.b_1028	1857	88,281
9	Bronte, C.	Villette	Wc.b_9182	1853	193,819
10	Bronte, C.	Jane Eyre	Wc.b_1260	1847	188,092
11	Bronte, E.	Wuthering Heights	We.b_768	1847	117,344
12	Eliot	Daniel Deronda	We_7469	1876	311,400
13	Eliot	Silas Marner	We_550	1861	71,449
14	Eliot	Middlemarch	We_145	1871-2	317,975
15	Eliot	The Mill on the Floss	We_6688	1860	207,505
16	Eliot	Brother Jacob	We_2171	1864	16,693
17	Eliot	Adam Bede	We_507	1859	215,253
18	Gaskell	Cranford	Wg_394	1851-3	71,037
19	Gaskell	Sylvia's Lovers	Wg_4537	1863	191,176
20	Gaskell	Mary Barton	Wg_2153	1848	161,098
21	Thackeray	Vanity Fair	Wt_599	1848	303,530
22	Thackeray	Barry Lyndon	Wt_4558	1844	125,986
23	Trollope	Doctor Thorne	Wtr_3166	1857	220,867
24	Trollope	Barchester Towers	Wtr_3409	1857	197,691
25	Trollope	The Warden	Wtr_619	1855	72,068
26	Trollope	Phineas Finn	Wtr_18000	1869	263,393
27	Trollope	Can You Forgive Her	Wtr_19500	1865	316,349
28	Trollope	The Eustace Diamonds	Wtr_7381	1873	269,981
29	Collins	After Dark	Wc_1626	1882	136,356
30	Collins	The Moonstone	Wc_155	1868	196,506
31	Collins	The Woman in White	Wc_583	1859	246,917
					sum: 5,292,795

B EVALUATION RESULTS

B.1 REPRESENTATIVE & DISTINCTIVE TERMS OF DICKENS VS. COLLINS (2)

Table B.1.1: Representativeness & Distinctiveness on *DickensCollinsSet2*. Results of evaluating distances for profiles P_D and P_C to test closeness to Dickens' documents also showing t-test results for hypothesis assuming greater mean for *Collins* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis of shared representative and distinctive terms of both profiles.

				Author Profile Comparison			**Clustering**
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
1	D1023	0.0322	0.0105	-0.0217	1.00	-0.0341 ... Inf	0.89
	D1392	0.0347	0.0089	-0.0257	1.00	-0.0391 ... Inf	
	D1394	0.0344	0.0087	-0.0256	1.00	-0.0351 ... Inf	
	D1400	0.0309	0.0103	-0.0206	1.00	-0.0314 ... Inf	
	D1406	0.0384	0.0102	-0.0282	1.00	-0.0369 ... Inf	
2	D1407	0.0253	0.0088	-0.0165	1.00	-0.0240 ... Inf	0.84
	D1414	0.0282	0.0111	-0.0172	1.00	-0.0233 ... Inf	
	D1415	0.0263	0.0103	-0.0160	1.00	-0.0231 ... Inf	
	D1416	0.0287	0.0110	-0.0178	1.00	-0.0271 ... Inf	
	D1419	0.0258	0.0105	-0.0154	1.00	-0.0241 ... Inf	
3	D1421	0.0386	0.0081	-0.0305	1.00	-0.0397 ... Inf	0.89
	D1465	0.0310	0.0086	-0.0223	1.00	-0.0312 ... Inf	
	D19337	0.0292	0.0103	-0.0189	1.00	-0.0292 ... Inf	
	D20795	0.0257	0.0100	-0.0156	0.99	-0.0251 ... Inf	
	D23344	0.0419	0.0064	-0.0356	1.00	-0.0563 ... Inf	
4	D23765	0.0328	0.0069	-0.0259	1.00	-0.0349 ... Inf	0.95
	D25852	0.0372	0.0108	-0.0263	0.99	-0.0441 ... Inf	
	D25853	0.0378	0.0116	-0.0263	0.98	-0.0473 ... Inf	
	D25854	0.0349	0.0117	-0.0232	0.99	-0.0401 ... Inf	
	D27924	0.0276	0.0102	-0.0175	1.00	-0.0258 ... Inf	
5	D37121	0.0445	0.0090	-0.0355	0.99	-0.0572 ... Inf	0.84
	D40723	0.0393	0.0093	-0.0299	1.00	-0.0416 ... Inf	
	D564	0.0278	0.0108	-0.0170	0.99	-0.0289 ... Inf	
	D580	0.0411	0.0105	-0.0305	0.98	-0.0554 ... Inf	
	D588	0.0400	0.0096	-0.0304	1.00	-0.0443 ... Inf	
6	D650	0.0318	0.0132	-0.0186	1.00	-0.0290 ... Inf	0.89
	D675	0.0290	0.0111	-0.0179	1.00	-0.0281 ... Inf	
	D699	0.0314	0.0123	-0.0191	1.00	-0.0299 ... Inf	
	D700	0.0276	0.0109	-0.0167	1.00	-0.0265 ... Inf	
	D730	0.0289	0.0112	-0.0177	1.00	-0.0275 ... Inf	
7	D766	0.0322	0.0115	-0.0207	1.00	-0.0324 ... Inf	0.84
	D821	0.0352	0.0110	-0.0241	1.00	-0.0354 ... Inf	
	D872	0.0379	0.0117	-0.0262	1.00	-0.0344 ... Inf	
	D882	0.0372	0.0110	-0.0262	0.99	-0.0436 ... Inf	
	D883	0.0298	0.0121	-0.0177	0.99	-0.0287 ... Inf	
8	D914	0.0284	0.0109	-0.0175	0.99	-0.0297 ... Inf	0.84
	D916	0.0499	0.0103	-0.0396	1.00	-0.0598 ... Inf	
	D917	0.0338	0.0107	-0.0231	0.99	-0.0368 ... Inf	
	D963	0.0288	0.0111	-0.0177	0.99	-0.0299 ... Inf	
	D967	0.0447	0.0101	-0.0346	0.99	-0.0557 ... Inf	
9	D968	0.0329	0.0103	-0.0225	1.00	-0.0349 ... Inf	0.89
	D98	0.0268	0.0106	-0.0162	1.00	-0.0208 ... Inf	
	Dal2324	0.0360	0.0091	-0.0270	1.00	-0.0404 ... Inf	
	Dal28198	0.0321	0.0096	-0.0225	1.00	-0.0302 ... Inf	
	DC1423	0.0289	0.0083	-0.0206	1.00	-0.0301 ... Inf	
	mean	0.0333	0.0102	-0.0230			
	sd	0.0057	0.0013	0.0062			
	SE	0.0009	0.0002	0.0009			

Table B.1.2: Representativeness & Distinctiveness on *DickensCollinsSet2*. Results of evaluating distances for profiles P_D and P_C to test closeness to Collins' documents also showing t-test results for hypothesis assuming greater mean for *Dickens* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis of shared representative and distinctive terms of both profiles.

Author Profile Comparison							Clustering
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
10	C1438	0.0415	0.0081	-0.0334	0.00	0.0198 ... Inf	0.84
	C155	0.0410	0.0090	-0.0320	0.00	0.0218 ... Inf	
	C1586	0.0453	0.0088	-0.0365	0.00	0.0239 ... Inf	
	C1587	0.0391	0.0086	-0.0305	0.00	0.0148 ... Inf	
	C1588	0.0384	0.0096	-0.0287	0.00	0.0167 ... Inf	
11	C1621	0.0581	0.0086	-0.0495	0.00	0.0320 ... Inf	0.95
	C1622	0.0542	0.0077	-0.0464	0.00	0.0244 ... Inf	
	C1623	0.0477	0.0080	-0.0396	0.00	0.0194 ... Inf	
	C1624	0.0596	0.0079	-0.0517	0.00	0.0347 ... Inf	
	C1625	0.0612	0.0095	-0.0517	0.01	0.0187 ... Inf	
12	C1626	0.0368	0.0096	-0.0272	0.00	0.0130 ... Inf	0.84
	C1627	0.0450	0.0090	-0.0360	0.00	0.0206 ... Inf	
	C1628	0.0479	0.0085	-0.0395	0.00	0.0199 ... Inf	
	C1629	0.0455	0.0092	-0.0363	0.00	0.0202 ... Inf	
	C1630	0.0387	0.0086	-0.0301	0.00	0.0171 ... Inf	
13	C170	0.0520	0.0069	-0.0451	0.00	0.0278 ... Inf	0.89
	C1895	0.0488	0.0071	-0.0417	0.00	0.0230 ... Inf	
	C1917	0.0418	0.0074	-0.0344	0.00	0.0156 ... Inf	
	C1975	0.0469	0.0072	-0.0397	0.00	0.0233 ... Inf	
	C28367	0.0426	0.0090	-0.0336	0.00	0.0225 ... Inf	
14	C3606	0.0411	0.0094	-0.0317	0.00	0.0210 ... Inf	0.95
	C3632	0.0480	0.0072	-0.0408	0.00	0.0269 ... Inf	
	C3633	0.0490	0.0070	-0.0420	0.00	0.0255 ... Inf	
	C3634	0.0520	0.0065	-0.0455	0.00	0.0291 ... Inf	
	C4605	0.0331	0.0071	-0.0260	0.01	0.0082 ... Inf	
	mean	0.0462	0.0082	-0.0380			
	sd	0.0072	0.0010	0.0075			
	SE	0.0014	0.0002	0.0015			

Figure B.1.1: Clustering on representative and distinctive terms on *DickensCollinsSet2* with "complete link" method based on the 4th iteration profile terms of both authors.

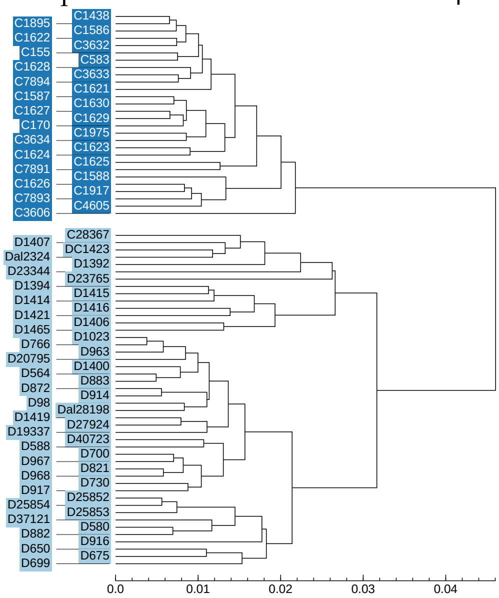


Table B.2.1: ICA on *DickensCollinsSet1*. Results of evaluating distances for profiles P_D and P_C to test closeness to Dickens' documents also showing t-test results for hypothesis assuming greater mean for *Collins* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis of shared characteristic terms of both profiles.

				Author Profile Comparison			Clustering
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
1	D33_SB	0.0107	0.0157	0.0050	0.00	0.0027 ... Inf	0.83
	D36_PP	0.0107	0.0160	0.0054	0.02	0.0011 ... Inf	
	D37a_OEP	0.0099	0.0155	0.0056	0.00	0.0023 ... Inf	
	D37b_OT	0.0086	0.0150	0.0064	0.00	0.0038 ... Inf	
	D38_NN	0.0076	0.0158	0.0083	0.00	0.0058 ... Inf	
2	D40a_MHC	0.0086	0.0166	0.0079	0.00	0.0051 ... Inf	0.83
	D40b_OCS	0.0066	0.0168	0.0102	0.00	0.0074 ... Inf	
	D41_BR	0.0065	0.0170	0.0105	0.00	0.0081 ... Inf	
	D42_AN	0.0083	0.0160	0.0076	0.00	0.0052 ... Inf	
	D43_MC	0.0064	0.0175	0.0111	0.00	0.0078 ... Inf	
3	D46a_PFI	0.0089	0.0161	0.0072	0.00	0.0046 ... Inf	0.83
	D46b_DS	0.0061	0.0166	0.0105	0.00	0.0082 ... Inf	
	D49_DC	0.0077	0.0154	0.0077	0.00	0.0056 ... Inf	
	D51_CHE	0.0107	0.0158	0.0051	0.02	0.0010 ... Inf	
	D52_BH	0.0076	0.0165	0.0089	0.00	0.0064 ... Inf	
4	D54_HT	0.0090	0.0161	0.0070	0.00	0.0038 ... Inf	0.83
	D55_LD	0.0072	0.0164	0.0092	0.00	0.0070 ... Inf	
	D56_RP	0.0085	0.0170	0.0085	0.00	0.0064 ... Inf	
	D59_TTC	0.0095	0.0152	0.0057	0.00	0.0034 ... Inf	
	D60a_UT	0.0084	0.0168	0.0084	0.00	0.0064 ... Inf	
5	D60b_GE	0.0110	0.0153	0.0044	0.00	0.0018 ... Inf	0.83
	D64_OMF	0.0108	0.0151	0.0043	0.00	0.0016 ... Inf	
	D70_ED	0.0100	0.0161	0.0061	0.00	0.0035 ... Inf	
	mean	0.0087	0.0161	0.0074			
	sd	0.0015	0.0007	0.0020			
	SE	0.0003	0.0001	0.0004			

Table B.2.2: ICA on *DickensCollinsSet1*. Results of evaluating distances for profiles P_D and P_C to test closeness to Collins' documents also showing t-test results for hypothesis assuming greater mean for *Dickens* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis of shared characteristic terms of both profiles.

			Author Profile Comparison				Clustering
Iteration	Test Doc.	Dist.D	Dist.C	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
	C50_Ant	0.0129	0.0137	0.0008	0.69	-0.0034 . . . Inf	
	C51_RBR	0.0124	0.0156	0.0033	0.96	-0.0063. . . Inf	
6	C52_Basil	0.0156	0.0105	-0.0051	0.00	0.0025. . . Inf	0.83
	C54_HS	0.0148	0.0128	-0.0020	0.18	-0.0016 . . . Inf	
	C56_AD	0.0157	0.0109	-0.0048	0.00	0.0019. . . Inf	
	C57_ARL	0.0167	0.0118	-0.0048	0.00	0.0021 . . . Inf	
	C59_QOH	0.0170	0.0097	-0.0073	0.00	0.0047. . . Inf	
7	C60_WIW	0.0180	0.0064	-0.0116	0.00	0.0088. . . Inf	0.83
	C62_NN	0.0178	0.0066	-0.0112	0.00	0.0087 . . . Inf	
	C66_Armadale	0.0180	0.0066	-0.0114	0.00	0.0090 . . . Inf	
	C68_MS	0.0174	0.0078	-0.0097	0.00	0.0070 . . . Inf	
	C70_MW	0.0171	0.0075	-0.0096	0.00	0.0066 . . . Inf	
8	C72_PMF	0.0184	0.0067	-0.0117	0.00	0.0083 . . . Inf	0.83
	C73_TNM	0.0181	0.0061	-0.0119	0.00	0.0086 . . . Inf	
	C75_LL	0.0179	0.0053	-0.0125	0.00	0.0105 . . . Inf	
	C76_TD	0.0181	0.0053	-0.0128	0.00	0.0107 . . . Inf	
	C78_HH	0.0178	0.0070	-0.0108	0.00	0.0082 . . . Inf	
9	C79_FL	0.0166	0.0069	-0.0097	0.00	0.0059 . . . Inf	0.83
	C80_JD	0.0174	0.0057	-0.0117	0.00	0.0095 . . . Inf	
	C81_BR	0.0174	0.0073	-0.0101	0.00	0.0062 . . . Inf	
	C84_ISN	0.0175	0.0091	-0.0084	0.00	0.0037 . . . Inf	
	C86_EG	0.0171	0.0059	-0.0112	0.00	0.0089 . . . Inf	
	mean	0.0168	0.0084	-0.0084			
	sd	0.0016	0.0029	0.0045			
	SE	0.0003	0.0006	0.0009			

Table B.2.3: ICA on *DickensCollinsSet2*. Results of evaluating distances for profiles P_D and P_C to test closeness to Dickens' documents, also showing t-test results for hypothesis assuming greater mean for *Collins* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis of shared characteristic terms of both profiles.

				Author Profile Comparison			**Clustering**
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
1	D1023	0.0032	0.0081	0.0049	0.00	0.0046 ... Inf	0.95
	D1392	0.0046	0.0064	0.0017	0.00	0.0011 ... Inf	
	D1394	0.0045	0.0069	0.0024	0.00	0.0018 ... Inf	
	D1400	0.0036	0.0080	0.0044	0.00	0.0040 ... Inf	
	D1406	0.0040	0.0073	0.0033	0.00	0.0028 ... Inf	
2	D1407	0.0044	0.0062	0.0018	0.00	0.0013 ... Inf	0.84
	D1414	0.0040	0.0068	0.0028	0.00	0.0023 ... Inf	
	D1415	0.0046	0.0066	0.0021	0.00	0.0014 ... Inf	
	D1416	0.0038	0.0073	0.0036	0.00	0.0030 ... Inf	
	D1419	0.0037	0.0070	0.0033	0.00	0.0028 ... Inf	
3	D1421	0.0047	0.0067	0.0020	0.00	0.0014 ... Inf	0.74
	D1465	0.0045	0.0064	0.0019	0.00	0.0013 ... Inf	
	D19337	0.0030	0.0075	0.0045	0.00	0.0040 ... Inf	
	D20795	0.0029	0.0077	0.0048	0.00	0.0044 ... Inf	
	D23344	0.0063	0.0046	-0.0017	1.00	-0.0023 ... Inf	
4	D23765	0.0057	0.0041	-0.0016	1.00	-0.0021 ... Inf	0.89
	D25852	0.0042	0.0077	0.0035	0.00	0.0029 ... Inf	
	D25853	0.0044	0.0077	0.0032	0.00	0.0027 ... Inf	
	D25854	0.0046	0.0075	0.0029	0.00	0.0023 ... Inf	
	D27924	0.0032	0.0070	0.0038	0.00	0.0033 ... Inf	
5	D37121	0.0036	0.0070	0.0034	0.00	0.0029 ... Inf	0.84
	D40723	0.0037	0.0070	0.0033	0.00	0.0028 ... Inf	
	D564	0.0037	0.0077	0.0039	0.00	0.0035 ... Inf	
	D580	0.0036	0.0074	0.0038	0.00	0.0033 ... Inf	
	D588	0.0038	0.0070	0.0032	0.00	0.0027 ... Inf	
6	D650	0.0035	0.0075	0.0040	0.00	0.0035 ... Inf	0.95
	D675	0.0035	0.0076	0.0041	0.00	0.0037 ... Inf	
	D699	0.0044	0.0072	0.0028	0.00	0.0023 ... Inf	
	D700	0.0035	0.0079	0.0044	0.00	0.0040 ... Inf	
	D730	0.0039	0.0075	0.0036	0.00	0.0031 ... Inf	
7	D766	0.0036	0.0079	0.0043	0.00	0.0039 ... Inf	0.84
	D821	0.0034	0.0079	0.0045	0.00	0.0041 ... Inf	
	D872	0.0029	0.0078	0.0049	0.00	0.0045 ... Inf	
	D882	0.0038	0.0074	0.0036	0.00	0.0031 ... Inf	
	D883	0.0042	0.0077	0.0035	0.00	0.0031 ... Inf	
8	D914	0.0026	0.0078	0.0052	0.00	0.0048 ... Inf	0.84
	D916	0.0043	0.0070	0.0028	0.00	0.0022 ... Inf	
	D917	0.0032	0.0078	0.0046	0.00	0.0042 ... Inf	
	D963	0.0038	0.0078	0.0041	0.00	0.0037 ... Inf	
	D967	0.0036	0.0077	0.0041	0.00	0.0037 ... Inf	
9	D968	0.0031	0.0079	0.0048	0.00	0.0045 ... Inf	0.84
	D98	0.0034	0.0076	0.0042	0.00	0.0037 ... Inf	
	Dal2324	0.0048	0.0063	0.0016	0.00	0.0010 ... Inf	
	Dal28198	0.0029	0.0079	0.0050	0.00	0.0046 ... Inf	
	DC1423	0.0062	0.0052	-0.0009	1.00	-0.0015 ... Inf	
	mean	0.0039	0.0072	0.0033			
	sd	0.0008	0.0009	0.0016			
	SE	0.0001	0.0001	0.0002			

Table B.2.4: ICA on *DickensCollinsSet2*. Results of evaluating distances for profiles P_D and P_C to test closeness to Collins' documents, also showing t-test results for hypothesis assuming greater mean for *Dickens* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis of shared characteristic terms of both profiles.

Author Profile Comparison							**Clustering**
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
10	C1438	0.0079	0.0024	-0.0055	0.00	0.0051 ... Inf	0.84
	C155	0.0078	0.0029	-0.0049	0.00	0.0045 ... Inf	
	C1586	0.0079	0.0026	-0.0053	0.00	0.0050 ... Inf	
	C1587	0.0080	0.0024	-0.0055	0.00	0.0052 ... Inf	
	C1588	0.0064	0.0047	-0.0017	0.00	0.0011 ... Inf	
11	C1621	0.0073	0.0038	-0.0035	0.00	0.0030 ... Inf	0.84
	C1622	0.0081	0.0024	-0.0057	0.00	0.0053 ... Inf	
	C1623	0.0080	0.0027	-0.0053	0.00	0.0049 ... Inf	
	C1624	0.0078	0.0028	-0.0050	0.00	0.0045 ... Inf	
	C1625	0.0068	0.0044	-0.0024	0.00	0.0018 ... Inf	
12	C1626	0.0067	0.0035	-0.0031	0.00	0.0027 ... Inf	0.84
	C1627	0.0078	0.0027	-0.0051	0.00	0.0047 ... Inf	
	C1628	0.0076	0.0032	-0.0044	0.00	0.0040 ... Inf	
	C1629	0.0078	0.0028	-0.0050	0.00	0.0046 ... Inf	
	C1630	0.0079	0.0019	-0.0060	0.00	0.0057 ... Inf	
13	C170	0.0078	0.0023	-0.0055	0.00	0.0051 ... Inf	0.84
	C1895	0.0080	0.0019	-0.0061	0.00	0.0058 ... Inf	
	C1917	0.0073	0.0030	-0.0044	0.00	0.0039 ... Inf	
	C1975	0.0079	0.0028	-0.0051	0.00	0.0047 ... Inf	
	C28367	0.0046	0.0061	0.0015	1.00	-0.0020 ... Inf	
14	C3606	0.0054	0.0057	0.0003	0.84	-0.0009 ... Inf	0.84
	C3632	0.0080	0.0023	-0.0057	0.00	0.0053 ... Inf	
	C3633	0.0078	0.0025	-0.0053	0.00	0.0049 ... Inf	
	C3634	0.0076	0.0032	-0.0044	0.00	0.0040 ... Inf	
	C4605	0.0069	0.0043	-0.0026	0.00	0.0021 ... Inf	
	mean	0.0074	0.0032	-0.0042			
	sd	0.0009	0.0011	0.0019			
	SE	0.0002	0.0002	0.0004			

Figure B.2.1: Clustering with ICA extracted characteristic terms on *DickensCollinsSet*2 with “complete link” method based on iteration one.

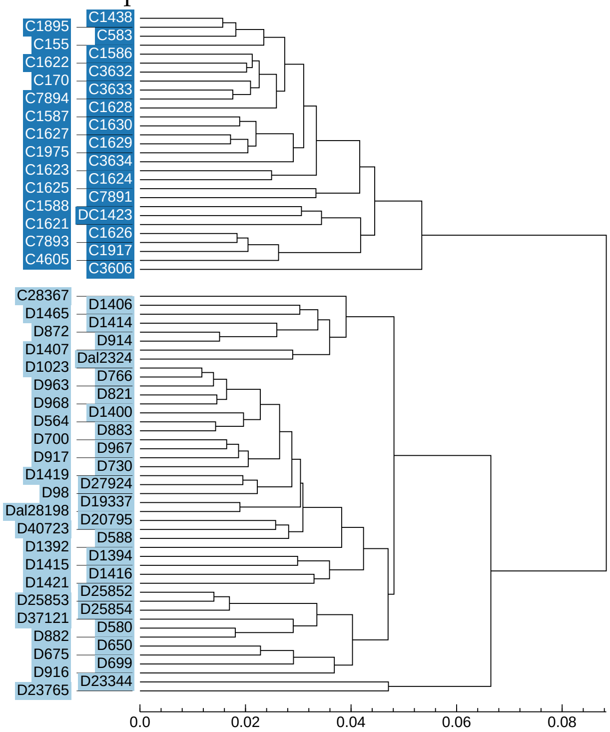


Table B.3.1: Combined ICA&RD on *DickensCollinsSet1*. Results of evaluating distances for profiles P_D and P_C to test closeness to Dickens' documents also showing t-test results for hypothesis assuming greater mean for *Collins* profile to test document. Clustering and corresponding *adjusted Rand* is on the on the basis of shared characteristic terms of both profiles.

Author Profile Comparison							**Clustering**
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
1.	D33_SB	0.0126	0.0248	0.0122	0.00	0.0087 ... Inf	0.83
	D36_PP	0.0128	0.0246	0.0118	0.00	0.0053 ... Inf	
	D37a_OEP	0.0127	0.0244	0.0117	0.00	0.0064 ... Inf	
	D37b_OT	0.0107	0.0246	0.0139	0.00	0.0088 ... Inf	
	D38_NN	0.0105	0.0257	0.0152	0.00	0.0094 ... Inf	
2	D40a_MHC	0.0115	0.0224	0.0109	0.00	0.0069 ... Inf	0.83
	D40b_OCS	0.0106	0.0228	0.0122	0.00	0.0081 ... Inf	
	D41_BR	0.0116	0.0227	0.0112	0.00	0.0070 ... Inf	
	D42_AN	0.0103	0.0228	0.0125	0.00	0.0089 ... Inf	
	D43_MC	0.0119	0.0237	0.0118	0.00	0.0064 ... Inf	
3	D46a_PFI	0.0157	0.0265	0.0108	0.00	0.0055 ... Inf	0.83
	D46b_DS	0.0146	0.0266	0.0120	0.00	0.0046 ... Inf	
	D49_DC	0.0144	0.0240	0.0096	0.00	0.0046 ... Inf	
	D51_CHE	0.0165	0.0259	0.0094	0.01	0.0024 ... Inf	
	D52_BH	0.0132	0.0252	0.0120	0.00	0.0071 ... Inf	
4	D54_HT	0.0102	0.0211	0.0110	0.00	0.0065 ... Inf	0.91
	D55_LD	0.0082	0.0211	0.0130	0.00	0.0099 ... Inf	
	D56_RP	0.0126	0.0211	0.0085	0.00	0.0049 ... Inf	
	D59_TTC	0.0144	0.0192	0.0048	0.02	0.0011 ... Inf	
	D60a_UT	0.0127	0.0208	0.0081	0.00	0.0049 ... Inf	
5	D60b_GE	0.0134	0.0195	0.0061	0.01	0.0016 ... Inf	0.83
	D64_OMF	0.0146	0.0194	0.0047	0.04	0.0003 ... Inf	
	D70_ED	0.0121	0.0203	0.0082	0.00	0.0049 ... Inf	
	mean	0.0125	0.0230	0.0105			
	sd	0.0020	0.0023	0.0027			
	SE	0.0004	0.0005	0.0006			

Table B.3.2: Combined ICA&RD on *DickensCollinsSet1*. Results of evaluating distances for profiles P_D and P_C to test closeness to Collins' documents also showing t-test results for hypothesis assuming greater mean for *Dickens* profile to test document. Clustering and corresponding *adjusted Rand* is on the on the basis of shared characteristic terms of both profiles.

				Author Profile Comparison			**Clustering**
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
	C50_Ant	0.0152	0.0177	0.0025	0.88	-0.0060 ... Inf	
	C51_RBR	0.0120	0.0201	0.0081	1.00	-0.0119 ... Inf	
6	C52_Basil	0.0175	0.0123	-0.0052	0.00	0.0022 ... Inf	0.91
	C54_HS	0.0158	0.0161	0.0002	0.54	-0.0046 ... Inf	
	C56_AD	0.0186	0.0128	-0.0058	0.00	.0025 ... Inf	
	C57_ARL	0.0167	0.0151	-0.0016	0.19	-0.0014 ... Inf	
	C59_QOH	0.0185	0.0122	-0.0063	0.00	0.0035 ... Inf	
7	C60_WIW	0.0233	0.0111	-0.0122	0.00	0.0082 ... Inf	0.83
	C62_NN	0.0242	0.0103	-0.0139	0.00	0.0101 ... Inf	
	C66_Armadale	0.0238	0.0120	-0.0118	0.00	0.0074 ... Inf	
	C68_MS	0.0227	0.0103	-0.0124	0.00	0.0091 ... Inf	
	C70_MW	0.0242	0.0106	-0.0136	0.00	0.0092 ... Inf	
8	C72_PMF	0.0229	0.0099	-0.0130	0.00	0.0084 ... Inf	0.91
	C73_TNM	0.0240	0.0092	-0.0149	0.00	0.0097 ... Inf	
	C75_LL	0.0236	0.0083	-0.0153	0.00	0.0124 ... Inf	
	C76_TD	0.0235	0.0080	-0.0156	0.00	0.0129 ... Inf	
	C78_HH	0.0235	0.0119	-0.0116	0.00	0.0069 ... Inf	
9	C79_FL	0.0203	0.0115	-0.0087	0.01	0.0028 ... Inf	0.83
	C80_JD	0.0215	0.0078	-0.0137	0.00	0.0110 ... Inf	
	C81_BR	0.0221	0.0106	-0.0115	0.00	0.0060 ... Inf	
	C84_ISN	0.0215	0.0127	-0.0088	0.01	0.0028 ... Inf	
	C86_EG	0.0223	0.0090	-0.0133	0.00	0.0108 ... Inf	
	mean	0.0208	0.0118	-0.0090			
	sd	0.0035	0.0031	0.0064			
	SE	0.0007	0.0007	0.0014			

Table B.3.3: Combined ICA&RD on *DickensCollinsSet2*. Results of evaluating distances for profiles P_D and P_C to test closeness to Dickens' documents also showing t-test results for hypothesis assuming greater mean for *Collins* profile to test document. Clustering and corresponding *adjusted Rand* is on the on the basis of shared characterisic terms of both profiles.

Author Profile Comparison							**Clustering**
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
1	D1023	0.0023	0.0078	0.0056	0.00	0.0052 ... Inf	0.84
	D1392	0.0049	0.0066	0.0016	0.00	0.0010 ... Inf	
	D1394	0.0044	0.0067	0.0023	0.00	0.0017 ... Inf	
	D1400	0.0027	0.0075	0.0048	0.00	0.0044 ... Inf	
	D1406	0.0040	0.0068	0.0028	0.00	0.0023 ... Inf	
2	D1407	0.0050	0.0062	0.0012	0.00	0.0006 ... Inf	0.70
	D1414	0.0038	0.0064	0.0027	0.00	0.0022 ... Inf	
	D1415	0.0041	0.0064	0.0023	0.00	0.0017 ... Inf	
	D1416	0.0036	0.0069	0.0034	0.00	0.0028 ... Inf	
	D1419	0.0037	0.0069	0.0032	0.00	0.0027 ... Inf	
3	D1421	0.0047	0.0070	0.0023	0.00	0.0016 ... Inf	0.84
	D1465	0.0048	0.0063	0.0015	0.00	0.0009 ... Inf	
	D19337	0.0033	0.0074	0.0041	0.00	0.0036 ... Inf	
	D20795	0.0028	0.0080	0.0052	0.00	0.0047 ... Inf	
	D23344	0.0067	0.0054	-0.0013	1.00	-0.0019 ... Inf	
4	D23765	0.0058	0.0042	-0.0016	1.00	-0.0021 ... Inf	0.84
	D25852	0.0039	0.0074	0.0035	0.00	0.0030 ... Inf	
	D25853	0.0042	0.0072	0.0030	0.00	0.0025 ... Inf	
	D25854	0.0042	0.0071	0.0029	0.00	0.0023 ... Inf	
	D27924	0.0037	0.0068	0.0031	0.00	0.0026 ... Inf	
5	D37121	0.0037	0.0067	0.0030	0.00	0.0025 ... Inf	0.84
	D40723	0.0033	0.0072	0.0039	0.00	0.0034 ... Inf	
	D564	0.0028	0.0076	0.0049	0.00	0.0044 ... Inf	
	D580	0.0032	0.0072	0.0039	0.00	0.0035 ... Inf	
	D588	0.0034	0.0072	0.0038	0.00	0.0033 ... Inf	
6	D650	0.0036	0.0072	0.0036	0.00	0.0031 ... Inf	0.84
	D675	0.0035	0.0072	0.0038	0.00	0.0033 ... Inf	
	D699	0.0041	0.0068	0.0026	0.00	0.0021 ... Inf	
	D700	0.0027	0.0076	0.0049	0.00	0.0046 ... Inf	
	D730	0.0034	0.0070	0.0036	0.00	0.0032 ... Inf	
7	D766	0.0027	0.0077	0.0049	0.00	0.0046 ... Inf	0.84
	D821	0.0027	0.0078	0.0051	0.00	0.0048 ... Inf	
	D872	0.0024	0.0076	0.0052	0.00	0.0049 ... Inf	
	D882	0.0032	0.0073	0.0041	0.00	0.0037 ... Inf	
	D883	0.0032	0.0075	0.0043	0.00	0.0039 ... Inf	
8	D914	0.0022	0.0076	0.0054	0.00	0.0051 ... Inf	0.84
	D916	0.0038	0.0069	0.0031	0.00	0.0026 ... Inf	
	D917	0.0026	0.0077	0.0051	0.00	0.0048 ... Inf	
	D963	0.0027	0.0077	0.0050	0.00	0.0046 ... Inf	
	D967	0.0027	0.0076	0.0049	0.00	0.0045 ... Inf	
9	D968	0.0024	0.0078	0.0054	0.00	0.0050 ... Inf	0.84
	D98	0.0032	0.0072	0.0040	0.00	0.0035 ... Inf	
	Dal2324	0.0046	0.0061	0.0015	0.00	0.0010 ... Inf	
	Dal28198	0.0028	0.0078	0.0050	0.00	0.0047 ... Inf	
	DC1423	0.0061	0.0050	-0.0011	1.00	-0.0016 ... Inf	
	mean	0.0036	0.0070	0.0034			
	sd	0.0010	0.0008	0.0017			
	SE	0.0002	0.0001	0.0003			

Table B.3.4: Combined ICA&RD on *DickensCollinsSet2*. Results of evaluating distances for profiles P_D and P_C to test closeness to Collins' documents also showing t-test results for hypothesis assuming greater mean for *Dickens* profile to test document. Clustering and corresponding *adjusted Rand* is on the on the basis of shared characteristic terms of both profiles.

				Author Profile Comparison			**Clustering**
Iteration	Test Doc.	Dist.D.	Dist.C.	(Dist.C-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
10	C1438	0.0084	0.0025	-0.0059	0.00	0.0055 . . . Inf	0.84
	C155	0.0081	0.0026	-0.0055	0.00	0.0050 . . . Inf	
	C1586	0.0083	0.0025	-0.0058	0.00	0.0054 . . . Inf	
	C1587	0.0084	0.0023	-0.0061	0.00	0.0057 . . . Inf	
	C1588	0.0067	0.0050	-0.0017	0.00	0.0011 . . . Inf	
11	C1621	0.0075	0.0035	-0.0040	0.00	0.0034 . . . Inf	0.89
	C1622	0.0082	0.0021	-0.0061	0.00	0.0057 . . . Inf	
	C1623	0.0081	0.0027	-0.0054	0.00	0.0050 . . . Inf	
	C1624	0.0081	0.0028	-0.0053	0.00	0.0048 . . . Inf	
	C1625	0.0072	0.0039	-0.0033	0.00	0.0027 . . . Inf	
12	C1626	0.0073	0.0041	-0.0032	0.00	0.0027 . . . Inf	0.70
	C1627	0.0080	0.0023	-0.0056	0.00	0.0053 . . . Inf	
	C1628	0.0077	0.0029	-0.0048	0.00	0.0044 . . . Inf	
	C1629	0.0080	0.0021	-0.0058	0.00	0.0055 . . . Inf	
	C1630	0.0081	0.0012	-0.0069	0.00	0.0067 . . . Inf	
13	C170	0.0077	0.0023	-0.0055	0.00	0.0051 . . . Inf	0.74
	C1895	0.0079	0.0019	-0.0060	0.00	0.0057 . . . Inf	
	C1917	0.0073	0.0033	-0.0040	0.00	0.0035 . . . Inf	
	C1975	0.0077	0.0021	-0.0055	0.00	0.0051 . . . Inf	
	C28367	0.0048	0.0066	0.0018	1.00	-0.0023 . . . Inf	
14	C3606	0.0050	0.0057	0.0008	0.99	-0.0013 . . . Inf	0.74
	C3632	0.0079	0.0022	-0.0057	0.00	0.0053 . . . Inf	
	C3633	0.0080	0.0024	-0.0056	0.00	0.0052 . . . Inf	
	C3634	0.0075	0.0032	-0.0043	0.00	0.0039 . . . Inf	
	C4605	0.0068	0.0045	-0.0023	0.00	0.0018 . . . Inf	
	mean	0.0075	0.0031	-0.0045			
	sd	0.0009	0.0013	0.0021			
	SE	0.0002	0.0003	0.0004			

Figure B.3.1: Clustering on combined ICA & RD characteristic terms on *DickensCollinsSet*2 with "complete link" method based on iteration 11.

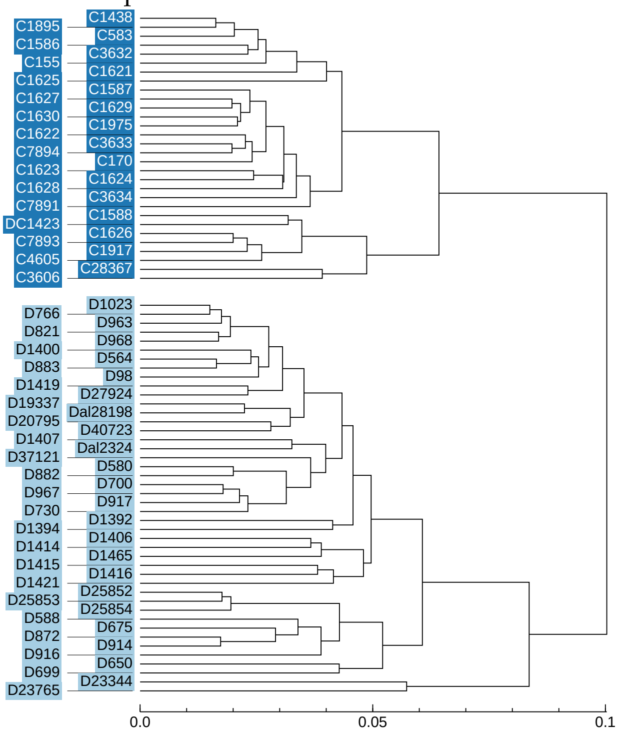


Table B.4.1: Representativeness & Distinctiveness on *DickensWorldSetSet*. Results of evaluating distances for profiles P_D and P_W to test closeness to Dickens' documents with failed t-test due to too few frequent terms in *World* profile. Clustering and corresponding *adjusted Rand* is on the basis of shared representative and distinctive terms of both profiles

Author Profile Comparison								Clustering
Iteration	Test Doc.	Dist.D.	Dist.W.	(Dist.W-Dist.D)	p-value	conf.interval	(lower/upper bound)	adjust.Rand
1	D1023	0.0106	0.1836	0.1730	NA	NA	NA	0.41
	D1400	0.0108	0.1513	0.1405	NA	NA	NA	
	D19337	0.0091	0.1282	0.1191	NA	NA	NA	
	D40723	0.0105	0.2821	0.2716	NA	NA	NA	
	D564	0.0099	0.2141	0.2042	NA	NA	NA	
2	D580	0.0071	0.1058	0.0987	NA	NA	NA	0.90
	D588	0.0076	0.0296	0.0220	NA	NA	NA	
	D650	0.0069	0.0058	-0.0011	NA	NA	NA	
	D675	0.0070	0.0776	0.0706	NA	NA	NA	
	D699	0.0099	0.0657	0.0558	NA	NA	NA	
3	D700	0.0108	0.1033	0.0925	NA	NA	NA	0.44
	D730	0.0110	0.1578	0.1468	NA	NA	NA	
	D766	0.0115	0.1441	0.1325	NA	NA	NA	
	D821	0.0113	0.1276	0.1163	NA	NA	NA	
	D872	0.0111	0.1962	0.1851	NA	NA	NA	
4	D882	0.0100	0.1500	0.1400	NA	NA	NA	0.90
	D883	0.0103	0.2120	0.2017	NA	NA	NA	
	D914	0.0095	0.1034	0.0938	NA	NA	NA	
	D916	0.0095	0.0808	0.0713	NA	NA	NA	
	D917	0.0106	0.1678	0.1573	NA	NA	NA	
5	D963	0.0104	0.1350	0.1247	NA	NA	NA	0.90
	D967	0.0104	0.1720	0.1616	NA	NA	NA	
	D968	0.0104	0.1519	0.1415	NA	NA	NA	
	D98	0.0099	0.1301	0.1201	NA	NA	NA	
	mean	0.0098	0.1365	0.1266	NA	NA	NA	
	sd	0.0014	0.0610	0.0601	NA	NA	NA	
	SE	0.0003	0.0124	0.0123	NA	NA	NA	

Table B.4.2: Representativeness & Distinctiveness on *DickensWorldSetSet*. Results of evaluating distances for profiles P_D and P_W to test closeness to World documents with failed t-test due to too few frequent terms in *World* profile. Clustering and corresponding *adjusted Rand* is on the basis of shared representative and distinctive terms of both profiles

Author Profile Comparison								**Clustering**
Iteration	Test Doc.	Dist.D.	Dist.W.	(Dist.W-Dist.D)	p-value	conf.interval	(lower/upper bound)	adjust.Rand
	Wa_105	0.0143	0.1387	0.1244	NA	NA	NA	
6	Wa_121	0.0119	0.0000	-0.0119	NA	NA	NA	0.41
	Wa_141	0.0132	0.0000	-0.0132	NA	NA	NA	
	Wa_158	0.0138	0.0000	-0.0138	NA	NA	NA	
	Wa_21839	0.0132	0.0000	-0.0132	NA	NA	NA	
	Wa_42671	0.0133	0.0000	-0.0133	NA	NA	NA	
7	Wa.b_767	0.0110	0.2868	0.2758	NA	NA	NA	0.41
	Wc_155	0.0111	0.0792	0.0680	NA	NA	NA	
	Wc_1626	0.0100	0.1890	0.1789	NA	NA	NA	
	Wc_583	0.0105	0.2530	0.2425	NA	NA	NA	
	Wc.b_1028	0.0099	0.2270	0.2171	NA	NA	NA	
8	Wc.b_1260	0.0105	0.1110	0.1005	NA	NA	NA	0.32
	Wc.b_9182	0.0107	0.0964	0.0857	NA	NA	NA	
	Wd_14436	0.0128	0.0687	0.0560	NA	NA	NA	
	Wd_370	0.0146	0.1795	0.1650	NA	NA	NA	
	Wd_376	0.0134	0.1707	0.1574	NA	NA	NA	
9	Wd_521	0.0128	0.1085	0.0956	NA	NA	NA	0.32
	Wd_6422	0.0137	0.1959	0.1821	NA	NA	NA	
	We_145	0.0106	0.1466	0.1360	NA	NA	NA	
	We_2171	0.0093	0.0744	0.0651	NA	NA	NA	
	We_507	0.0105	0.1490	0.1384	NA	NA	NA	
10	We_550	0.0098	0.1864	0.1766	NA	NA	NA	0.29
	We_6688	0.0101	0.2212	0.2111	NA	NA	NA	
	We_7469	0.0101	0.1579	0.1478	NA	NA	NA	
	We.b_768	0.0098	0.1619	0.1521	NA	NA	NA	
	Wf_1147	0.0114	0.0731	0.0616	NA	NA	NA	
11	Wf_5256	0.0136	0.1920	0.1784	NA	NA	NA	0.41
	Wf_6098	0.0150	0.1868	0.1718	NA	NA	NA	
	Wf_6593	0.0144	0.1032	0.0888	NA	NA	NA	
	Wf_9609	0.0141	0.0576	0.0436	NA	NA	NA	
	Wg_2153	0.0123	0.1994	0.1871	NA	NA	NA	
12	Wg_394	0.0104	0.0981	0.0877	NA	NA	NA	0.41
	Wg_4537	0.0104	0.0984	0.0880	NA	NA	NA	
	Wgo_2667	0.0123	0.2115	0.1992	NA	NA	NA	
	Wr_12398	0.0133	0.1080	0.0947	NA	NA	NA	
	Wr_6124	0.0144	0.1716	0.1572	NA	NA	NA	
13	Ws_2160	0.0125	0.2222	0.2097	NA	NA	NA	0.41
	Ws_2311	0.0121	0.2131	0.2009	NA	NA	NA	
	Ws_4084	0.0135	0.2082	0.1947	NA	NA	NA	
	Ws_4085	0.0128	0.3127	0.2999	NA	NA	NA	
	Ws_6758	0.0115	0.2235	0.2121	NA	NA	NA	
14	Ws_6761	0.0118	0.1102	0.0984	NA	NA	NA	0.41
	Wst_1079	0.0124	0.1315	0.1192	NA	NA	NA	
	Wst_804	0.0110	0.0613	0.0504	NA	NA	NA	
	Wsw_17157	0.0108	0.0319	0.0212	NA	NA	NA	
	Wsw_4208	0.0154	0.1514	0.1360	NA	NA	NA	
15	Wsw_4737	0.0123	0.0593	0.0470	NA	NA	NA	0.38
	Wt_4558	0.0116	0.1317	0.1200	NA	NA	NA	
	Wt_599	0.0114	0.1337	0.1223	NA	NA	NA	
	Wtr_18000	0.0133	0.0793	0.0660	NA	NA	NA	
	Wtr_19500	0.0122	0.0942	0.0820	NA	NA	NA	
	mean	0.0121	0.1346	0.1225	NA	NA	NA	
	sd	0.0016	0.0761	0.0763	NA	NA	NA	
	SE	0.0002	0.0107	0.0107	NA	NA	NA	

Table B.5.1: ICA on *DickensWorldSet*. Results of evaluating distances for profiles P_D and P_W to test closeness to Dickens' documents also showing t-test results for hypothesis assuming greater mean for *World* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis of shared characterisic terms of both profiles.

			Author Profile Comparison				**Clustering**
Iteration	Test Doc.	Dist.D.	Dist.W.	(Dist.W-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
1	D1023	0.0031	0.0079	0.0048	0.00	0.0043 ... Inf	0.95
	D1400	0.0029	0.0081	0.0052	0.00	0.0048 ... Inf	
	D19337	0.0035	0.0077	0.0041	0.00	0.0036 ... Inf	
	D40723	0.0038	0.0075	0.0037	0.00	0.0032 ... Inf	
	D564	0.0034	0.0079	0.0045	0.00	0.0040 ... Inf	
2	D580	0.0031	0.0077	0.0046	0.00	0.0041 ... Inf	0.95
	D588	0.0039	0.0072	0.0033	0.00	0.0027 ... Inf	
	D650	0.0044	0.0075	0.0030	0.00	0.0025 ... Inf	
	D675	0.0046	0.0071	0.0025	0.00	0.0019 ... Inf	
	D699	0.0073	0.0048	-0.0025	1.00	-0.0031 ... Inf	
3	D700	0.0026	0.0081	0.0055	0.00	0.0051 ... Inf	0.95
	D730	0.0030	0.0079	0.0049	0.00	0.0045 ... Inf	
	D766	0.0031	0.0078	0.0047	0.00	0.0043 ... Inf	
	D821	0.0029	0.0081	0.0052	0.00	0.0048 ... Inf	
	D872	0.0034	0.0074	0.0040	0.00	0.0035 ... Inf	
4	D882	0.0037	0.0077	0.0040	0.00	0.0034 ... Inf	0.90
	D883	0.0024	0.0082	0.0058	0.00	0.0054 ... Inf	
	D914	0.0030	0.0078	0.0049	0.00	0.0044 ... Inf	
	D916	0.0048	0.0071	0.0024	0.00	0.0017 ... Inf	
	D917	0.0018	0.0081	0.0063	0.00	0.0059 ... Inf	
5	D963	0.0029	0.0082	0.0053	0.00	0.0049 ... Inf	0.95
	D967	0.0028	0.0080	0.0052	0.00	0.0047 ... Inf	
	D968	0.0029	0.0081	0.0052	0.00	0.0048 ... Inf	
	D98	0.0034	0.0080	0.0046	0.00	0.0041 ... Inf	
	mean	0.0034	0.0077	0.0042			
	sd	0.0011	0.0007	0.0017			
	SE	0.0002	0.0001	0.0004			

Table B.5.2: ICA on *DickensWorldSet*. Results of evaluating distances for profiles P_D and P_W to test closeness to World documents also showing t-test results for hypothesis assuming greater mean for *Dickens* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis of shared characteristic terms of both profiles.

			Author Profile Comparison				**Clustering**
Iteration	Test Doc.	Dist.D.	Dist.W.	(Dist.W-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
	Wa_105	0.0074	0.0034	-0.0039	0.00	0.0035 ... Inf	
6	Wa_121	0.0072	0.0057	-0.0014	0.00	0.0008 ... Inf	0.95
	Wa_141	0.0077	0.0044	-0.0033	0.00	0.0028 ... Inf	
	Wa_158	0.0076	0.0045	-0.0031	0.00	0.0026 ... Inf	
	Wa_21839	0.0077	0.0044	-0.0033	0.00	0.0028 ... Inf	
	Wa_42671	0.0077	0.0043	-0.0033	0.00	0.0028 ... Inf	
7	Wa.b_767	0.0068	0.0044	-0.0024	0.00	0.0017 ... Inf	0.95
	Wc_155	0.0054	0.0067	0.0013	1.00	-0.0020 ... Inf	
	Wc_1626	0.0048	0.0074	0.0026	1.00	-0.0033 ... Inf	
	Wc_583	0.0055	0.0068	0.0012	1.00	-0.0019 ... Inf	
	Wc.b_1028	0.0051	0.0063	0.0013	1.00	-0.0019 ... Inf	
8	Wc.b_1260	0.0051	0.0059	0.0009	0.99	-0.0015 ... Inf	0.03
	Wc.b_9182	0.0051	0.0059	0.0008	0.98	-0.0014 ... Inf	
	Wd_14436	0.0070	0.0036	-0.0035	0.00	0.0030 ... Inf	
	Wd_370	0.0073	0.0033	-0.0041	0.00	0.0036 ... Inf	
	Wd_376	0.0068	0.0040	-0.0028	0.00	0.0022 ... Inf	
9	Wd_521	0.0067	0.0050	-0.0017	0.00	0.0011 ... Inf	0.95
	Wd_6422	0.0070	0.0043	-0.0028	0.00	0.0022 ... Inf	
	We_145	0.0069	0.0049	-0.0020	0.00	0.0015 ... Inf	
	We_2171	0.0048	0.0067	0.0019	1.00	-0.0025 ... Inf	
	We_507	0.0054	0.0070	0.0016	1.00	-0.0023 ... Inf	
10	We_550	0.0053	0.0071	0.0018	1.00	-0.0024 ... Inf	0.03
	We_6688	0.0059	0.0067	0.0008	0.99	-0.0014 ... Inf	
	We_7469	0.0065	0.0056	-0.0009	0.01	0.0003 ... Inf	
	We.b_768	0.0052	0.0065	0.0013	1.00	-0.0019 ... Inf	
	Wf_1147	0.0073	0.0040	-0.0033	0.00	0.0027 ... Inf	
11	Wf_5256	0.0075	0.0048	-0.0028	0.00	0.0022 ... Inf	0.03
	Wf_6098	0.0079	0.0041	-0.0038	0.00	0.0033 ... Inf	
	Wf_6593	0.0080	0.0041	-0.0039	0.00	0.0035 ... Inf	
	Wf_9609	0.0077	0.0046	-0.0031	0.00	0.0026 ... Inf	
	Wg_2153	0.0054	0.0057	0.0003	0.79	-0.0009 ... Inf	
12	Wg_394	0.0056	0.0057	0.0001	0.56	-0.0007 ... Inf	0.95
	Wg_4537	0.0049	0.0071	0.0022	1.00	-0.0028 ... Inf	
	Wgo_2667	0.0071	0.0043	-0.0028	0.00	0.0023 ... Inf	
	Wr_12398	0.0079	0.0031	-0.0048	0.00	0.0044 ... Inf	
	Wr_6124	0.0076	0.0039	-0.0037	0.00	0.0031 ... Inf	
13	Ws_2160	0.0072	0.0056	-0.0016	0.00	0.0010 ... Inf	0.95
	Ws_2311	0.0064	0.0056	-0.0008	0.01	0.0002 ... Inf	
	Ws_4084	0.0074	0.0056	-0.0018	0.00	0.0012 ... Inf	
	Ws_4085	0.0074	0.0055	-0.0018	0.00	0.0013 ... Inf	
	Ws_6758	0.0067	0.0064	-0.0004	0.17	-0.0003 ... Inf	
14	Ws_6761	0.0073	0.0043	-0.0030	0.00	0.0025 ... Inf	0.95
	Wst_1079	0.0061	0.0054	-0.0007	0.04	0.0001 ... Inf	
	Wst_804	0.0054	0.0064	0.0010	0.99	-0.0017 ... Inf	
	Wsw_17157	0.0062	0.0052	-0.0010	0.00	0.0004 ... Inf	
	Wsw_4208	0.0069	0.0038	-0.0031	0.00	0.0025 ... Inf	
15	Wsw_4737	0.0062	0.0045	-0.0017	0.00	0.0011 ... Inf	0.95
	Wt_4558	0.0066	0.0048	-0.0018	0.00	0.0012 ... Inf	
	Wt_599	0.0059	0.0058	-0.0001	0.38	-0.0005 ... Inf	
	Wtr_18000	0.0072	0.0041	-0.0031	0.00	0.0026 ... Inf	
	Wtr_19500	0.0072	0.0043	-0.0029	0.00	0.0023 ... Inf	
	mean	0.0066	0.0052	-0.0014			
	sd	0.0010	0.0011	0.0020			
	SE	0.0001	0.0002	0.0003			

Table B.6.1: Combined ICA& RD on *DickensWorldSet*. Results of evaluating distances for profiles P_D and P_C to test closeness to Dickens' documents also showing t-test results for hypothesis assuming greater mean for *World* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis shared characteristic terms of both profiles.

Author Profile Comparison							**Clustering**
Iteration	Test Doc.	Dist.D.	Dist.W.	(Dist.W-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
1	D1023	0.0028	0.0068	0.0040	0.00	0.0034 ... Inf	0.03
	D1400	0.0028	0.0060	0.0032	0.00	0.0027 ... Inf	
	D19337	0.0034	0.0065	0.0032	0.00	0.0026 ... Inf	
	D40723	0.0034	0.0065	0.0031	0.00	0.0026 ... Inf	
	D564	0.0033	0.0070	0.0036	0.00	0.0030 ... Inf	
2	D580	0.0030	0.0074	0.0044	0.00	0.0039 ... Inf	-0.01
	D588	0.0040	0.0070	0.0030	0.00	0.0024 ... Inf	
	D650	0.0047	0.0071	0.0024	0.00	0.0018 ... Inf	
	D675	0.0050	0.0069	0.0019	0.00	0.0013 ... Inf	
	D699	0.0069	0.0050	-0.0019	1.00	-0.0025 ... Inf	
3	D700	0.0027	0.0072	0.0045	0.00	0.0040 ... Inf	0.85
	D730	0.0030	0.0072	0.0042	0.00	0.0036 ... Inf	
	D766	0.0027	0.0063	0.0037	0.00	0.0031 ... Inf	
	D821	0.0026	0.0072	0.0047	0.00	0.0041 ... Inf	
	D872	0.0037	0.0068	0.0031	0.00	0.0026 ... Inf	
4	D882	0.0042	0.0066	0.0025	0.00	0.0019 ... Inf	0.90
	D883	0.0024	0.0071	0.0047	0.00	0.0042 ... Inf	
	D914	0.0035	0.0070	0.0036	0.00	0.0030 ... Inf	
	D916	0.0049	0.0064	0.0015	0.00	0.0009 ... Inf	
	D917	0.0018	0.0073	0.0054	0.00	0.0050 ... Inf	
5	D963	0.0028	0.0059	0.0031	0.00	0.0025 ... Inf	0.03
	D967	0.0032	0.0063	0.0032	0.00	0.0027 ... Inf	
	D968	0.0032	0.0066	0.0034	0.00	0.0029 ... Inf	
	D98	0.0035	0.0064	0.0030	0.00	0.0024 ... Inf	
	mean	0.0035	0.0067	0.0032			
	sd	0.0011	0.0005	0.0014			
	SE	0.0002	0.0001	0.0003			

Table B.6.2: Combined ICA& RD on *DickensWorldSet*. Results of evaluating distances for profiles P_D and P_C to test closeness to World documents also showing t-test results for hypothesis assuming greater mean for *Dickens* profile to test document. Clustering and corresponding *adjusted Rand* is on the basis shared characteristic terms of both profiles.

				Author Profile Comparison			Clustering
Iteration	Test Doc.	Dist.D.	Dist.W.	(Dist.W-Dist.D)	p-value	conf.interval (lower/upper bound)	adjust.Rand
	Wa_105	0.0079	0.0029	-0.0050	0.00	0.0045 ... Inf	
6	Wa_121	0.0076	0.0050	-0.0026	0.00	0.0020 ... Inf	0.85
	Wa_141	0.0083	0.0045	-0.0038	0.00	0.0034 ... Inf	
	Wa_158	0.0083	0.0051	-0.0031	0.00	0.0026 ... Inf	
	Wa_21839	0.0081	0.0049	-0.0032	0.00	0.0027 ... Inf	
	Wa_42671	0.0081	0.0049	-0.0031	0.00	0.0026 ... Inf	
7	Wa.b_767	0.0072	0.0037	-0.0035	0.00	0.0029 ... Inf	0.95
	Wc_155	0.0057	0.0049	-0.0007	0.02	0.0001 ... Inf	
	Wc_1626	0.0051	0.0054	0.0003	0.83	-0.0008 ... Inf	
	Wc_583	0.0056	0.0047	-0.0010	0.00	0.0004 ... Inf	
	Wc.b_1028	0.0056	0.0047	-0.0009	0.00	0.0004 ... Inf	
8	Wc.b_1260	0.0053	0.0051	-0.0002	0.29	-0.0004 ... Inf	-0.01
	Wc.b_9182	0.0053	0.0053	0.0001	0.56	-0.0006 ... Inf	
	Wd_14436	0.0080	0.0065	-0.0016	0.00	0.0010 ... Inf	
	Wd_370	0.0085	0.0060	-0.0025	0.00	0.0019 ... Inf	
	Wd_376	0.0081	0.0066	-0.0015	0.00	0.0009 ... Inf	
9	Wd_521	0.0065	0.0053	-0.0012	0.00	0.0006 ... Inf	0.85
	Wd_6422	0.0070	0.0051	-0.0019	0.00	0.0013 ... Inf	
	We_145	0.0065	0.0036	-0.0029	0.00	0.0024 ... Inf	
	We_2171	0.0055	0.0059	0.0004	0.85	-0.0011 ... Inf	
	We_507	0.0047	0.0050	0.0003	0.83	-0.0008 ... Inf	
10	We_550	0.0048	0.0065	0.0016	1.00	-0.0023 ... Inf	0.35
	We_6688	0.0050	0.0061	0.0011	1.00	-0.0017 ... Inf	
	We_7469	0.0060	0.0048	-0.0012	0.00	0.0007 ... Inf	
	We.b_768	0.0045	0.0057	0.0012	1.00	-0.0018 ... Inf	
	Wf_1147	0.0075	0.0042	-0.0033	0.00	0.0027 ... Inf	
11	Wf_5256	0.0078	0.0044	-0.0034	0.00	0.0028 ... Inf	0.35
	Wf_6098	0.0081	0.0040	-0.0040	0.00	0.0035 ... Inf	
	Wf_6593	0.0081	0.0041	-0.0040	0.00	0.0035 ... Inf	
	Wf_9609	0.0078	0.0048	-0.0029	0.00	0.0024 ... Inf	
	Wg_2153	0.0049	0.0056	0.0007	0.98	-0.0013 ... Inf	
12	Wg_394	0.0056	0.0048	-0.0007	0.02	0.0002 ... Inf	0.35
	Wg_4537	0.0045	0.0061	0.0017	1.00	-0.0023 ... Inf	
	Wgo_2667	0.0075	0.0048	-0.0026	0.00	0.0020 ... Inf	
	Wr_12398	0.0080	0.0043	-0.0036	0.00	0.0031 ... Inf	
	Wr_6124	0.0074	0.0049	-0.0024	0.00	0.0018 ... Inf	
13	Ws_2160	0.0073	0.0059	-0.0014	0.00	0.0008 ... Inf	0.35
	Ws_2311	0.0067	0.0058	-0.0009	0.01	0.0002 ... Inf	
	Ws_4084	0.0075	0.0046	-0.0029	0.00	0.0023 ... Inf	
	Ws_4085	0.0073	0.0052	-0.0021	0.00	0.0015 ... Inf	
	Ws_6758	0.0066	0.0061	-0.0006	0.08	-0.0001 ... Inf	
14	Ws_6761	0.0074	0.0049	-0.0024	0.00	0.0018 ... Inf	0.35
	Wst_1079	0.0064	0.0061	-0.0004	0.18	-0.0003 ... Inf	
	Wst_804	0.0055	0.0064	0.0009	0.98	-0.0015 ... Inf	
	Wsw_17157	0.0066	0.0050	-0.0016	0.00	0.0010 ... Inf	
	Wsw_4208	0.0070	0.0045	-0.0025	0.00	0.0019 ... Inf	
15	Wsw_4737	0.0071	0.0055	-0.0016	0.00	0.0009 ... Inf	0.35
	Wt_4558	0.0067	0.0056	-0.0011	0.00	0.0004 ... Inf	
	Wt_599	0.0055	0.0061	0.0006	0.96	-0.0012 ... Inf	
	Wtr_18000	0.0070	0.0045	-0.0025	0.00	0.0019 ... Inf	
	Wtr_19500	0.0067	0.0045	-0.0022	0.00	0.0016 ... Inf	
	mean	0.0067	0.0051	-0.0016			
	sd	0.0012	0.0008	0.0016			
	SE	0.0002	0.0001	0.0002			

Table B.7.1: Ranking of Dickens' terms over the first five iterations using Representativeness & Distinctiveness on the *DickensCollinsSet*1.

Rank	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
1	first	upon	first	only	letter
2	discovered	first	only	first	left
3	produced	left	letter	letter	only
4	only	return	discovered	future	first
5	left	future	future	left	future
6	resolution	only	tried	discovered	wait
7	upon	letter	return	upon	words
8	future	discovered	second	return	news
9	letter	news	end	later	discovered
10	being	end	left	lines	upon
11	words	happened	to	words	serious
12	attempt	words	words	wait	advice
13	return	advice	met	position	later
14	end	produced	produced	resolution	return
15	but	written	upon	produced	written
16	serious	lines	advice	end	happened
17	followed	wait	wait	news	end
18	wait	resolution	resolution	second	lines
19	events	enough	written	advice	resolution
20	suddenly	serious	serious	serious	answer
21	later	much	position	happened	chance
22	news	later	news	moment	questions
23	lines	position	promised	written	produced
24	advice	waited	happened	experience	write
25	so	absence	with	chance	leave
26	absence	chance	down	waited	warning
27	chance	already	later	absence	second
28	written	moment	lines	entirely	waited
29	position	longer	moment	events	enough
30	happened	with	change	motives	absence

Table B.7.2: Ranking of Dickens' terms over the first five iterations using separate ICA on the *DickensCollinsSet*1.

Rank	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
1	upon	upon	upon	upon	upon
2	its	much	down	old	great
3	down	down	much	great	old
4	great	dear	great	down	down
5	much	great	such	oliver	its
6	being	come	its	then	such
7	come	being	many	much	much
8	such	then	being	being	many
9	though	says	come	such	where
10	like	such	though	dear	oliver
11	then	like	old	replied	every
12	many	where	oliver	come	some
13	old	well	then	though	these
14	where	always	joe	should	being
15	sir	oliver	replied	some	nicholas
16	says	old	never	many	then
17	good	know	where	boy	replied
18	never	sir	these	sir	night
19	returned	head	here	its	though
20	night	never	night	where	come
21	dear	here	boffin	joe	says
22	know	dorrit	off	head	like
23	head	clennam	young	boffin	never
24	these	going	well	quite	micawber
25	always	its	always	says	always
26	some	though	some	it.s	mother
27	any	replied	every	micawber	peggotty
28	off	off	boy	nicholas	people
29	here	dombey	gentleman	returned	long
30	fire	nicholas	X.em	know	dombey

Table B.7.3: Ranking of Dickens' terms over the first five iterations using ICA combined with representative and distinctive components on the *DickensCollinsSet*1.

Rank	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
1	upon	upon	upon	sir	great
2	much	much	young	upon	upon
3	sir	says	gentleman	dear	much
4	says	being	boffin	much	pickwick
5	great	great	much	boffin	says
6	should	young	miss	says	many
7	dear	boffin	sir	being	these
8	where	dear	bella	should	such
9	old	should	should	it.s	its
10	being	where	being	know	being
11	down	its	wegg	bella	where
12	its	down	great	young	some
13	then	like	nicholas	don.t	young
14	though	come	then	gentleman	about
15	came	always	pickwick	wegg	never
16	boffin	never	franklin	well	our
17	many	bella	sergeant	miss	sir
18	richard	then	says	lady	down
19	come	miss	off	come	people
20	george	got	dear	quite	weller
21	king	going	john	then	every
22	pickwick	about	down	say	sam
23	these	many	most	never	off
24	know	off	having	old	any
25	never	wegg	gentlemen	that.s	old
26	long	joe	than	down	then
27	head	gentleman	it.s	richard	gentleman
28	john	our	came	about	dorrit
29	joe	know	many	came	two
30	bella	well	its	george	always