Using Style Markers for Detecting Plagiarism in Natural Language Documents

HS-IDA-MD-03-004

Marco Kimler

Submitted by Marco Kimler to the University of Skövde as a dissertation towards the degree of M.Sc. by examination and dissertation in the Department of Computer Science.

August 2003

I certify that all material in this dissertation which is not my own work has been identified and that no material is included for which a degree has already been conferred to me.

Marco Kimler

Abstract

Most of the existing plagiarism detection systems compare a text to a database of other texts. These external approaches, however, are vulnerable because texts not contained in the database cannot be detected as source texts. This paper examines an internal plagiarism detection method that uses style markers from authorship attribution studies in order to find stylistic changes in a text. These changes might pinpoint plagiarized passages. Additionally, a new style marker called "specific words" is introduced. A pre-study tests if the style markers can "fingerprint" an author's style and if they are constant with sample size. It is shown that vocabulary richness measures do not fulfil these prerequisites. The other style markers - simple ratio measures, readability scores, frequency lists, and entropy measures - have these characteristics and are, together with the new specific words measure, used in a main study with an unsupervised approach for detecting stylistic changes in plagiarized texts at sentence and paragraph levels. It is shown that at these small levels the style markers generally cannot detect plagiarized sections because of intra-authorial stylistic variations (i.e. noise), and that at bigger levels the results are strongly affected by the sliding window approach. The specific words measure, however, can pinpoint single sentences written by another author.

Acknowledgements

Above all, I would like to thank my supervisor Kim Laurio for being a never drying-up spring of ideas and inspiration, and for patiently reading and commenting my alpha-quality drafts. In addition, I am grateful to my fellow students and my fellow exchange students, for fruitful discussions on life, this thesis and everything. Thanks to Christoph Bunzmann for reading my drafts and giving invaluable suggestions of many kinds. Special thanks go to Johanna, who morally supported me and understood that I preferred evenings with this work to evenings with her.

Contents

Chapter 1

Introduction

Plagiarism - the “wrongful act of taking the product of another person’s mind, and presenting it as one’s own” (Lindey 1952, p. 3, cited in Gibaldi 1999, p. 30) - is a growing problem for the scientific society in general and universities in particular (Clough 2000b; Culwin and Lancaster 2000). By copying existing material from the web or written literature, students subvert the use of assignments, essays and dissertations. Universities ignoring or even tolerating the problem endanger their reputation for honesty and fairness. Therefore, lecturers are forced to check whether papers handed in by students contain cases of plagiarism. However, a manual analysis is not feasible because of the number of students and the lack of time: a computational method is much needed.

A lot of computational methods for the detection of plagiarism exist. Most of them use an external approach: the suspicious text is compared to other texts in a database; if one passage in the target text already exists in the database, plagiarism is assumed (see chapter 2 for a more thorough discussion of these approaches). However, the results largely depend on the size and the quality of the database which is used; texts which are not stored cannot be identified as source texts. Furthermore, external analysis is not very elegant as it uses a computationally intensive “brute-force” approach: the whole database is scanned, though most of the documents are not related at all. Finally, most of the existing plagiarism detection services are commercial and are - due to the prices of up to one US dollar per checked text - too expensive for many academic institutions (Culwin and Lancaster 2000).

Little work has been done on internal approaches, which try to find suspicious passages *in* a text. This is somewhat surprising because this approach is intuitively used by lecturers and teachers: while reading a text handed in by a student, they can easily identify changes in the author's style (e.g. more complex sentence structures, more elegant phrases and words). These changes often indicate cases of plagiarism (Clough 2000b). In order to identify these potentially plagiarized sections, no other document is needed. Hence this approach can be called "intra-document analysis".

It is tempting to mimic this human strategy for detecting plagiarism and to analyse the style of a text in order to identify differences. Other fields, for example authorship attribution studies, have developed approaches to measure the style of a text and to compare different texts. These stylistic measures shall be examined in this study for their applicability at a sentence or paragraph level.

1.1 Aim

The aim of this work is to apply style markers from the field of authorship attribution to a single document in order to detect plagiarism.

Style markers have been frequently used at a text level to attribute whole texts to one author. This study investigates whether these style markers are also applicable at a sentence or paragraph level to detect stylistic changes that might pinpoint plagiarism.

1.2 Objectives

The following objectives are steps towards fulfilling the aim stated above:

1. *Review of existing literature and style marker identification:*

   An overview of stylometric approaches shall be presented. From the results of the works reviewed a first set of potentially applicable style markers is identified and presented.

2. *Software tools for performing a stylistic analysis:*

   The process of analysing a text should be automated. Therefore it is necessary to have software which is capable of parsing the input text, extracting the style markers, and presenting the results in an easily interpretable form. Since tools for intra-document stylistic analysis do not exist, it is necessary to create a new set of tools that fulfils these prerequisites.

3. *Pre-study: Testing the general applicability of style markers:*

   The style markers chosen above have been widely used at a text level. However, it is unclear whether these approaches also work at a sentence or paragraph level, which would be the prerequisite for plagiarism detection. A pre-study has to check the general applicability of the style markers in this field.

4. *Main study: Applying style markers to detect plagiarism:*

   The style markers which seem generally applicable according to the results of the pre-study have to be applied to plagiarized texts. For the main study, artificially plagiarized documents are to be created by inserting sentences of one text into another text of different authorship.

   If the measures really work at a sentence or paragraph level (i.e., if sections written by different authors indeed lead to different values), these differences can be detected and visualized. In the main study the aim posed in section 1.1 will be investigated by using an unsupervised approach.

To sum up, this study will perform two first steps to answer the question whether it is possible to use an internal approach using style markers from authorship attribution to detect plagiarism in documents. Therefore, the general applicability of style markers to small sentence and paragraph level will be analysed. Then, the style markers which seem generally applicable are applied to plagiarized texts in order to detect plagiarism with an unsupervised approach.

1.3 Dissertation Outline

The rest of this thesis is structured as follows: Chapter 2 will provide some general background on plagiarism and stylometry, while chapter 3 provides references to related work done in these fields. This chapter will also present a set of existing style markers which will be used in this work. A new style marker called specific words is introduced in chapter 4. Chapter 5 will outline the general methodology used, and bring up some common issues concerning preprocessing and style marker extraction. In chapter 6, solutions to these issues are suggested. Furthermore, some implementational details will be given. Chapter 7 outlines a pre-study that checks which style markers are generally applicable for intra-document analysis. The main study experiments, in which the style markers are applied to artificially plagiarized documents, are described in chapter 8. Chapter 9 discusses the presented results from both pre-study and main study. The most important conclusions drawn from the results are presented in chapter 10, together with a summary of the contributions of this work and ideas for future work.

Chapter 2

Background

2.1 On plagiarism

In the introduction, plagiarism has been defined as the “wrongful act of taking the product of another person’s mind, and presenting it as one’s own” (Lindey 1952, p. 3, cited in Gibaldi 1999, p. 30). Since this definition by Lindey is very broad and vague, the rest of this section tries to comment on the definition and to go into more detail.

Lindey’s notion of ‘the product of another person’s mind’ is very broad. As Evans (2000) notes, this includes all works from photographs and musical compositions to text documents. Concerning texts, plagiarism ranges from inserting sections of different authorship into one’s own text (verbatim copying, paraphrasing) and illegal teamwork (collusion) to complete ghost-writing (Martin 1994; Clough 2000b). Furthermore, plagiarism comprises the action of taking foreign thoughts, ideas, and lines of argument without reference (Martin 1994). This work focuses on the detection of inserted text sections.

When hearing the term plagiarism, many people think of lazy students, who do not want to spend hours on writing their own paper, but copy whole sections from textbooks or from other students. According to Evans (2000), however, this flagrant plagiarism is less common than unintentional cases of plagiarism arising from ignorance. Many students forget to give references or reference incorrectly because they never learned how to cite properly (Carroll and Appleton 2001). Students from cultures where it is normal to memorize and literally reproduce knowledge may not be used to the Anglo-Saxon scientific culture of presenting own thoughts and referencing old ones (Lesko 1996). Many writers do not know that it is also necessary to reference their own work (so called auto-plagiarism or self-plagiarism) (Evans 2000). Another case of unintentional plagiarism is cryptomnesia, where authors present ideas which they think are original, but in fact are based on memories they have forgotten (Carroll 2001).

All examples given so far are cases of plagiarism, even the unintentional ones. The only case of copied knowledge which may be used without reference is “common knowledge” (Carroll and Appleton 2001): nobody has to reference a source when stating ‘Second World War ended in 1945’. The problem is that common knowledge is rarely defined and varies from field to field (Carroll and Appleton 2001, p. 14).

The unintentional cases of plagiarism can be counteracted relatively easily. As Evans points out, “understanding plagiarism is a key to its discovery and prevention” (Evans 2000, Definition section). Universities have to provide clear definitions of what is correct and forbidden, and have to teach their students how to cite works (Carroll and Appleton 2001, p. 15); authors should be concise in referencing, and use style manuals (Evans 2000).

The intentional cases of plagiarism can also be fought. Offering interesting assignments will motivate students and quicken their interests. “Designing out opportunities for plagiarism” (Carroll and Appleton 2001, p. 9) - for example creating individualized tasks - will make it harder to cheat. Assessments can check whether students have really dealt with the subject (Evans 2000).

In cases where prevention does not work, plagiarism detection techniques can help to check if a text is original. Generally, two plagiarism detection approaches can be identified:

- External plagiarism detection methods compare a target text to other texts in a repository, and search for documents which might be the source from which the author plagiarized.
- Internal plagiarism detection methods try to find suspicious passages (e.g. stylistic inconsistencies) *in* a text. No other texts are needed for comparison.

Both external and internal approaches have their individual advantages and shortcomings, which will be discussed in section 2.2. Some problems, however, are innate to both approaches. Firstly, they cannot determine the motivation behind plagiarism: whether a passage was copied flagrantly or because of ignorance cannot be ascertained. Secondly, they can hardly distinguish common knowledge from 'personal' knowledge. Therefore common knowledge would be incorrectly regarded as plagiarized. Lastly, it is hard to show if copied material is correctly cited (i.e. no plagiarism), incorrectly cited (plagiarism in some cases, for example if not completely referenced), or not referenced at all (plagiarism).

Despite these shortcomings, plagiarism detection systems are valuable tools for checking students' papers for fraud and assist lecturers in highlighting suspicious passages themselves. However, the final decision of what is plagiarism, and how it influences the judgement of the papers, is still in the hand of the lecturer.

2.2 The temptation of an internal approach to plagiarism detection

Most of the existing plagiarism detection tools use external approaches. On the one hand this is understandable, since external techniques have various advantages: They are well studied in the literature, and they can directly point to sources from which an author has plagiarized. Internal approaches can never prove plagiarism, they can just pinpoint sections which are different in some sense.

On the other hand, external techniques also have shortcomings: The comparison to millions of mostly completely unrelated texts is computationally intensive and - as being brute-force-like - not very elegant. The database has to be as big as possible, and it has to be kept up-to-date, since texts which are not in the repository cannot be identified as source texts.

Furthermore external approaches are not very "natural". Lecturers and teachers also perform an indirect plagiarism check: while reading a student's paper, they can identify regions in the text which are "somewhat different" because the language level changes rapidly or because different words or sentence structures are used (Clough 2000b, p. 5). Clough (2000b) and Evans (2000) give further hints to detect plagiarism internally.

Since internal techniques can circumvent these shortcomings, their use is very tempting. Hence, if a working internal approach was found, it would be very powerful. Natural Language Processing (NLP) and stylometry have done much work on the quantitative analysis of style, and have developed style markers which can distinguish different authors. However, these style markers have not been applied at a sentence or paragraph level to detect plagiarism. This is what will be done in this work.

2.3 Stylometry

Stylometry is “an attempt to capture the essence of the style of a particular author by reference to a variety of quantitative criteria” (McEnery and Oakes 2000, p. 548). These quantifiable characteristics are called discriminators or style markers.

Stylometry assumes that every author or genre has a set of quantifiable characteristics (i.e. style markers) that are claimed constant for all works of this author but can discriminate different authors (Holmes 1994, p. 87). Because of the similarity to human fingerprints, the metaphor of “fingerprinting” will be used throughout this thesis.

Some authors claim for their style markers that it is not possible to manipulate them consciously (Holmes 1998, p. 111). This assumption, however, is not very realistic. Especially simple style markers, like sentence length or readability scores, can be easily affected by changing punctuation. But also more complex measures, like most frequent words or vocabulary richness measures, are not immune to conscious manipulation. Authors can artificially limit their available vocabulary by avoiding words of foreign origin or by substituting seldom special words (e.g. *convertible*, *coupé*, *limousine*, *roadster*, *saloon/sedan*, *SUV*) by using more general ones (*car*). Tirvengadum (1996) shows that the French author Romain Gary consciously changed his writing style for a novel under the pseudonym Émile Ajar. That stylistic change - which is impossible according to Holmes - was detected by comparing frequency distributions (see section 3.2.1).

It seems paradoxical that on the one hand stylometrists claim that style markers are constant for one author, but on the other hand take advantage of the fact that they change slightly over time, which allows works to be dated (Holmes 1994, p. 99). McEnery and Oakes (2000) note that genre also has an effect on stylistic fingerprints, and that genre changes can suppress differences in authorship.

These contradictions lead to severe criticism of stylometry. Rudman (1998) states “that for every paper announcing an authorship attribution method that ‘works’ [...], there is a counter paper that points out real or imagined crucial shortcomings" (Rudman 1998, p. 352). However, despite all criticism many papers have shown that it is possible to distinguish the styles of different authors (see chapter 3 for examples). Furthermore, the above-mentioned problems affect the use of style markers in plagiarism detection only marginally. Since student assignments are usually about one topic, genre changes do not occur. The stylistic change over time does not carry weight when texts are written in a period of a few weeks or months. Many students plagiarize to save time, and hence it is unlikely that these students invest time to change the style so that the stylistic fingerprint changes.

Consequently, it seems reasonable to use style markers for internal plagiarism detection. A survey of style markers and related work where they have been used is given in the following chapter.

Chapter 3

Related Work

3.1 Plagiarism Detection

As indicated in the introduction, most plagiarism detection tools use external methods. The external approaches can be grouped into methods analysing constrained or unconstrained text and in copy detection methods.

Most of the early external approaches focus on *constrained text*, especially software source code (Clough 2003). Since this thesis concentrates on the analysis of natural language texts, the reader is referred to related literature in this field. Verco and Wise (1996) and Clough (2000b) provide good overviews of existing techniques for constrained texts.

Since the 1990s, the focus of external techniques shifted towards *unconstrained text*, i.e. natural language text. The most important tools for plagiarism detection in free texts are the web-based services Plagiarism.org[1] and Copycatch.[2] These tools compare an electronically submitted document to an internal database of documents and provide a report with pointers to possible sources from which a similar passage might have been copied. See Culwin and Lancaster (2000) and Bull et al. (2001)

[1] Available at http://www.plagiarism.org

[2] Available at http://www.copycatch.freeserve.co.uk

for reviews of these services.

Another focus of external techniques are copy detection techniques, which try to prevent (by encrypting and watermarking) and detect plagiarism of texts. The most important copy detection systems are CHECK (Si, Leong and Lau 1997), SCAM (Shivakumar and Garcia-Molina 1995), and YAP (Verco and Wise 1996). Clough (2003) provides a good introduction to the field of copy detection.

Internal plagiarism techniques, which try to find suspicious passages *in* the text, are used rarely. Hersee (2001) is one of the few papers using an internal plagiarism detection method. Hersee applies the cusum technique (Farringdon 1996), but fails in detecting plagiarism. This may be due to the cusum technique, which has been severely criticized in the literature (de Haan 1998; McEnery and Oakes 2000, see also section 3.3). Therefore the cusum technique will not be used in this work.

3.2 Stylometry

Stylometry is an umbrella term standing for a broad range of applications where style markers are used to find changes in one or more texts. The most important field is authorship attribution, where style markers are applied to assign a text of unknown authorship to one particular author. Holmes (1992) and McEnery and Oakes (2000) provide introductions to the field of authorship attribution. Other fields using stylometry are for example genre detection (Clough 2000a; Stamatatos, Fakotakis and Kokkinakis 2000), collaborative writing (Glover 1996; Hoad and Zobel 2002), or literary forensics (Chaski 1997).

Section 3.2.1 will introduce five types of style markers, which have been used frequently and successfully in stylometry. An overview of recent work which uses these style markers follows in section 3.2.2.

3.2.1 Style markers

For this thesis, five types of style markers have been identified: simple ratio measures, readability scores, vocabulary richness measures, frequency lists, and relative entropy. All these style markers work at a lexical level.[3] Style markers at a syntactic level[4] are not considered, since they need a syntactically annotated text (see section 3.2.2 for details). Other approaches which were not chosen are shortly presented in section 3.2.2.

Simple ratio measures

Simple ratio measures are defined as style markers which are the proportion of two easily extractable text variables. As the following discussion will show, simple ratio measures have been frequently criticized, but because they are easy to compute, they will be included in the analysis in this work. Moreover, they may serve for setting a baseline for more complex style markers.

Sentence length, or *words per sentence*, is defined as the ratio of words and sentences in a text (equation 3.1 on equation page 18). Yule (1939) used sentence length to analyse different works from the Middle Ages (Kempis' *De Imitatione Christi*) to the 19th century (Coleridge, Lamb, Macaulay) and concludes that "sentence-length *is* a characteristic of an author's style" (Yule 1939, p. 370, original italics). Due to some shortcomings, especially concerning conscious control and change of punctuation due to editing, sentence length has been rarely used over the last years.

The *syllables per word* measure (equation 3.2 on page 18) was used by Fucks (1952) for his analysis of eight German and English authors. Although he focused on distinguishing German from English texts, he also points to "the peculiarities of

[3] The lexical level can be defined as the 'word level', i.e. words constitute the types. A typical type on a lexical level could be *car*.

[4] The syntactic level can be seen as the 'word-type level'. The level changes from words (e.g. *car*) to grammatical groups (e.g. *substantive, singular, subject...*)

style structure of a certain author" (Fucks 1952, p. 128), which could be used for distinguishing authors. Therefore, it seems promising to use the *syllables per word* measure also in the analysis of potentially plagiarized texts.

Readability scores

"Readability describes the ease with which a document can be read" (Stephens 2000, p. 1). According to Johnson, readability is affected by three factors: interest and motivation of the reader, the legibility of the document (font type and size, line length and spacing, etc.), and the complexity of the sentences (Johnson 1998, p. 1). While the first two factors are rather subjective and hard to quantify, various formulae exist for measuring the sentence complexity. Most of these formulae evaluate the number of words and syllables in a text, only a few take the grammatical structure into account.

Flesch Reading Ease score (Clough 2000a), in the following denoted by*Flesch*, uses average sentence length and number of syllables per word to calculate a percentage representing the ease of readability (equation 3.3 on page 18), i.e. higher values denote texts which are easier to read.

Flesch-Kincaid Formula (Johnson 1998) (short: *Kincaid*), uses the same information to calculate a score standing for the grade level of a reader who can to understand the text (equation 3.4 on page 18).

Gunning FOG Readability Test (short: *FOG*) (Johnson 1998) only considers words with three or more syllables (so called "complex words") to estimate the age of the reader who is able to understand the text (equation 3.5 on page 18).

Though the parameters in these formulae (like 206.835 in equation 3.3 on page 18) pretend exactness, readability scores are empirical equations and can only *estimate* how easy a text is to read. However, they are easy to compute, seem to work (see, for example, Clough 2000a), and do not put restrictions on the text size which is analysed. Therefore these three measures will be used in this thesis.

Other readability measures exist, but they are either hard to compute or inflexible. *Powers-Sumner-Kearl formula*, *McLaughlin 'SMOG' Formula* and *FORCAST formula* (Johnson 1998) need a certain amount of words or sentences to compute a readability score. Therefore they are not usable for an analysis at a sentence level, since sentence length varies and a constant sample length can not be assured. *Fry Readability Graph* uses a coordinate system which is used to determine the reading age. This procedure cannot easily be converted into an algorithm. Other approaches measuring the complexity of the sentence structure (Yngve 1960, cited in Glover 1996) require a comprehensive parsing of the text. Because of the mentioned shortcomings those readability measures were not chosen for this work.

Vocabulary richness measures

Vocabulary richness measures quantify the diversity of an author's vocabulary by evaluating information about the frequencies of occurring words. The most familiar vocabulary richness measure is the type/token[5] ratio $\frac{V_1}{V}$, which is, however, not constant over sample size and hence not usable for authorship attribution studies (McEnery and Oakes 2000, p. 551).

Most vocabulary richness measures focus on a part of the frequency spectrum, for example once occurring words V_1 (called hapax legomena) or twice occurring words V_2 (hapax dislegomena). Generally, a type occurring N times in a text is denoted by V_N. V denotes the number of tokens in a text.

Honoré (1979) proposes a formula based on the ratio of once occurring words (hapax legomena) and the length of the text, *Honoré's R* (equation 3.6 on page 18). Honoré claims that it "directly tests the propensity of an author to choose between

[5] Speaking with object-oriented vocabulary, types are the *classes* of a text (the word *car*), tokens their *instantiations* (each occurrence of the word *car*). Alternatively, type often denotes the number of different words in a text, while token refers to the number of occurrences for each type (Hockey 2000, p. 89).

the alternatives of using a word used previously and deploying a new word" (Honoré 1979, p. 175).

While the type/token ratio $\frac{V_1}{V}$ is very unstable with respect to variation in sample size (McEnery and Oakes 2000, p. 551), Sichel (1975) states that the ratio of hapax dislegomena to the vocabulary size is constant for different sample sizes (Tweedie and Baayen 1998). *Sichel's S* is given in equation 3.7 on page 18.

For his analysis of the Latin book *De Imitatione Christi*, whose authorship is disputed between Thomas à Kempis and Jean Charlier de Gerson, Yule (1944) developed a measure which does not only evaluate hapax legomena or dislegomena words, but the whole spectrum of types. *Yule's Characteristic K* is now often used in stylometry in multivariate analyses. It appears mostly in a revised form,[6] as in equation 3.8 on page 18.

Brunet (1978) developed a parametric formula called *Brunet's W* (equation 3.9 on page 18) for his thorough analysis of the French writer Jean Giraudoux. Holmes and Forsyth (1995) propose that W is constant for $0.165 \leq a \leq 0.172$. In this work, a is set to 0.172, the original value used by Brunet (Brunet 1978, p. 49).

Frequency lists

Measures based on frequency lists extract a specified list of types from two samples and compare the frequencies in these lists to each other. In contrast to the style markers presented so far, frequency lists do not result in one single value which can be calculated with a simple formula. A more complex process of list extraction and comparison is necessary, which is presented in chapter 6.

[6] The original form, $K = 10,000\frac{S_2 - S_1}{S_1^2}$ (Yule 1944, p. 47, equation 3.6) is equivalent, but not used in recent literature. See Yule (1944, pp. 12-13) for a definition of S_N.

In his study of Jane Austen's novels, Burrows (1987) compares the frequencies of the thirty most frequent words from different novels to each other. Mosteller and Wallace (1964) manually chose a list of words to attribute the disputed *Federalist Papers* to either Alexander Hamilton or John Madison. Frequency lists are not limited to a word level; McCombe (2002) for example successfully uses letter unigrams, letter bigrams and letter trigrams.

The extracted frequency lists can for example be compared with a correlation matrix (Burrows 1987), principal components analysis (Holmes and Forsyth 1995), hierarchical cluster analysis (Hoover 2001), or a χ^2 test (Kilgarriff 1996).

This thesis will use frequency lists at a word level and at a letter unigram, bigram and trigram level since at all these levels promising results have been reported (see also section 3.2.2).

Relative entropy

In information theory, *entropy* is a well known measure of amount of information contained in a message (Pierce 1980, p. 80). In other words, entropy quantifies the diversity and redundancy of a text, i.e. it is a kind of vocabulary richness measure. Entropy has been used for stylistic analysis by Bruno (1974).

A variation of entropy is *relative entropy*, which is also known as *Kullback-Leibler divergence*. Relative entropy does not measure the diversity of a message itself, but quantifies how different one sample p is compared to the overall population q (Manning and Schütze 1999). For intra-document analysis, p is the focused section and q is the complete text. If i is the index for every token in a text, relative entropy H_{rel} is defined as $H_{rel} = -\sum p_i \log_2 \frac{p_i}{q_i}$ (see also equation 3.10 on page 18).

Simple ratio measures

$$\text{WPS} = \frac{N_{\text{Words}}}{N_{\text{Sentences}}} \tag{3.1}$$

$$\text{SPW} = \frac{N_{\text{Syllables}}}{N_{\text{Words}}} \tag{3.2}$$

Readability scores

$$\text{Flesch} = 206.835 - 1.015 \cdot \text{WPS} - 84.6 \cdot \frac{N_{\text{Syllables}}}{N_{\text{Words}}} \tag{3.3}$$

$$\text{Kincaid} = 0.39 \cdot \text{WPS} + 11.8 \cdot \frac{N_{\text{Syllables}}}{N_{\text{Words}}} - 15.59 \tag{3.4}$$

$$\text{FOG} = 0.4 \cdot \left(\text{WPS} + \frac{N_{\text{Complex Words}}}{N_{\text{Words}}}\right) \tag{3.5}$$

Vocabulary richness measures

$$\text{Honoré's } R = \frac{100 \cdot \log N}{1 - \frac{V_1}{V}} \tag{3.6}$$

$$\text{Sichel's } S = \frac{V_2}{V} \tag{3.7}$$

$$\text{Yule's } K = 10^4 \cdot \frac{\left(\sum_{i=1}^{N} i^2 V_i\right) - N}{N^2} \tag{3.8}$$

$$\text{Brunet's } W = N^{V^{-a}} \tag{3.9}$$

Relative Entropy

$$H_{rel} = -\sum p_i \log_2 \frac{p_i}{q_i} \tag{3.10}$$

Figure 3.1: Formulae of style markers used in this study. See the text for more information on the parameters of the formulae.

3.2.2 Recent work in stylometry

While the previous section focused on the definition and the background of a set of style markers, this section will present recent work where they have been used. Many of the newer papers do not introduce new style markers, but evaluate old ones together in multivariate analyses.

The discriminators most often used in the literature are lists of most frequent words, which are compared to each other. With Mosteller and Wallace (1964) and Burrows (1987) two of the most influencing works in authorship attribution use this method. The idea of using frequent words as a style marker has since then been used in a multitude of works, e.g. Holmes and Forsyth (1995), Stamatatos, Fakotakis and Kokkinakis (2000), and Hoover (2001, 2002).

Other approaches using frequency distributions do not work at a word, but at a letter level. McCombe (2002) compares methods using words with ones using letter unigrams, letter bigrams, and letter trigrams. She concludes that letter unigrams discriminate the authors best, while letter bigrams and letter trigrams have little discriminatory power. Kjell (1994), however, presents promising results by using letter bigrams which are classified by a neural network. Khmelev and Tweedie (2001) and Kukushkina, Polikarpov and Khmelev (2000) successfully use Markov chains of letters (i.e. letter bigrams) to discriminate authors.

Besides frequency distributions, vocabulary richness measures are often used in stylometry. Many of the newer papers use various vocabulary richness measures like Yule's K, Honoré's R, Sichel's S, and Brunet's W (see previous section) together and evaluate them with multivariate analyses such as hierarchical clustering and PCA. Examples are Holmes (1992), Holmes and Forsyth (1995), and Baayen (1996).

Readability scores are used rather seldom in authorship attribution. Clough (2000a), however, applies readability scores to analyse the style of British newspapers and shows that hierarchical clustering can distinguish tabloids (like *Sun*) from broadsheets (like *The Times*).

3.3 Style markers not chosen for this work

The methods described in section 3.2.1 and section 3.2.2 have shown promising results and seem to be usable for discriminating authors in an intra-document analysis. However, there are also approaches which seem not to be applicable in this study since they have certain shortcomings or are too complex for this work. These style markers will be presented in the following.

The cusum technique (Morton 1978; Farringdon 1996), which was used by Hersee (2001) for internal plagiarism detection, has severe shortcomings and has been harshly criticized in the literature. Referring to the high degree of inherent subjectivity - beginning from the arbitrary choice of style markers, so called *habits* and the unstandardized visualization of data to the analysis of the results - McEnery and Oakes (2000, p. 555) discount the cusum technique as “impressionistic and prone to distorted interpretation”. de Haan expresses “doubts as to its validity” (de Haan 1998, p. 69) because of the same reasons. Because of these flaws the cusum technique will not be used in this study.

Baayen (1996) and Stamatatos, Fakotakis and Kokkinakis (2001) show that the analysis at a syntactic level leads to better attribution results. This, however, requires that the texts are syntactically annotated, which is often a non-trivial task, as automatic parsers are still imperfect (Holmes 1998, p. 116). Therefore, the analysis in this work is limited to a lexical level.

Methods based on neural networks (Matthews and Merriam 1993; Merriam and Matthews 1994; Kjell 1994) have shown promising results for solving common authorship attribution problems, but cannot be used for plagiarism detection. The reason is that, before the analysis, it is not known which sentences are plagiarized. Therefore it is impossible to provide the network with training data to set up the network weights. The same problems are encountered by approaches using genetic algorithms (as Holmes and Forsyth 1995).

Chapter 4

A new measure: Specific words

The style markers presented in section 3.2.1 are mostly used at a text level (usually 1,000 words and more), and none of these measures was developed for internal plagiarism detection. In the following, a new metric, called "specific words" will be introduced, which has been developed for this study and can detect specialties at a very small level of a few sentences.

4.1 Idea and hypothesis

Stylometry assumes that the style of an author can be fingerprinted, i.e. the style of one author is constant for all works of this author, but different to the styles of other authors (see section 2.3). For vocabulary similar assumptions can be formulated:

- Different authors use a different vocabulary when writing a text. This may be due to the amount of vocabulary being at an author's disposal (which can be quantified by vocabulary richness measures), or due to different preferences when an author faces a choice which synonym to use. Mosteller and Wallace (1964), for example, utilized the fact that Madison and Hamilton preferred different synonyms when writing *The Federalist Papers* to attribute the disputed articles.

- On the other hand, the vocabulary of one author does not significantly change during a text. Of course, an author will use new words, for example, when changing to a new topic. However, the personal basic vocabulary, including preferences which synonyms or expressions to use, does not change.

Based on these assumptions, the following hypothesis can be formulated:

Hypothesis. *Because the vocabulary of one author does not significantly change during a text, the number of specific words is relatively constant for the text. A short section by another author, however, will have many specific words because the author who wrote that inserted section uses another vocabulary.*

The term *specific word*, which is used in the hypothesis, is defined as follows:

Definition 1. *A* specific word *is a word which only occurs in a focused section of a text but not in the rest of the text. The word is specific to this focused section.*

The hypothesis is supported if an analysis of different sections in a text shows that sentences of different authorship actually result in a higher number of specific words. The hypothesis must be rejected if this is not the case.

4.2 Algorithm

In the previous section a hypothesis how specific words could be used to detect a change in authorship has been formulated. Now this hypothesis must be tested with an algorithm.

The idea behind the algorithm is to split a text into several parts, to extract the absolute or relative number of specific words for each of these parts and to compare the values to each other.

Figure 4.1 shows the pseudocode for the extraction algorithm of specific word

```
Number of specific words $N_{SW} = 0$
Create word frequency list $WFL_T$ for whole text
Split text into sentences
for each sentence S
    Create word frequency list $WFL_S$ for actual sentence
    for each word $i$ in $WFL_S$
        Get frequency $f_S(i)$ of word $i$ in $WFL_S$
        Get frequency $f_T(i)$ of word $i$ in $WFL_T$
        if $f_S(i)$ equals $f_T(i)$, i.e. word is specific to sentence
            $N_{SW} = N_{SW} + f_S(i)$
    Plot new point with $y = N_{SW}$ into coordinate system
    Reset $N_{SW}$
```

Figure 4.1: Pseudocode for the extraction and visualization of the absolute specific words measure at a sentence level.

values at a sentence level. The key idea behind the algorithm is to count all words which only occur in this sentence but not in the rest of the text. The number of words for which this is true is defined as the absolute number of specific words of this sentence. To account for different sentence lengths, the absolute number of specific words is divided by the length of the sentence. This relative specific words measure is used throughout the rest of this thesis.

The values of the specific words measure can now be analysed for example in a coordinate system. Examples for such diagrams can be found in figures 8.4 and 8.5. If changes of authorship do not lead to outliers in the graph, the hypothesis that the specific words measure can detect changes of authorship must be rejected.

Chapter 5

Method

The aim of this thesis is to apply style markers to a single document and to investigate if they can detect stylistic changes at a sentence or a paragraph level (see section 1.1). From that prerequisite, several phases in the analysis of a text can be deduced (figure 5.1): Style markers (which might have to be preprocessed) are extracted from the text; after that the results from applying the style markers are evaluated, for example by visualising them graphically or by applying statistical tests.

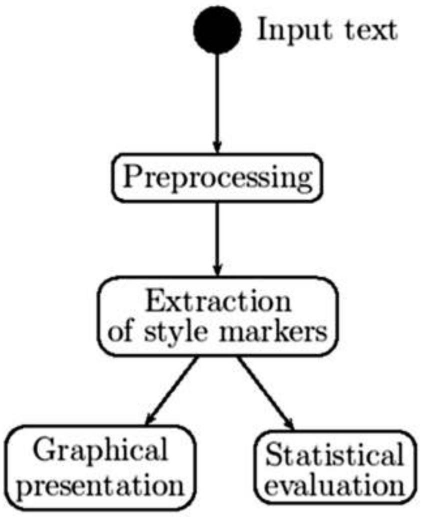


Figure 5.1: Different phases in the analysis of a text.

The rest of this chapter discusses these phases in more detail. Concerning preprocessing and extraction, common problems will be highlighted to which solutions will be presented in the implementation chapter. In addition, the sliding window approach as an analysis method will be explained. At last, several techniques for interpreting the data are presented: visual inspection using coordinate systems, hierarchical cluster analysis, principal components analysis (PCA), self-organizing maps (SOMs), and miscellaneous statistical tests.

5.1 Preprocessing

Before a text can be analysed with style markers, it has to be preprocessed, i.e. brought into a format from which style markers can be easily extracted.

In the case of simple ASCII files, in the preprocessing phase the text is just loaded and split into sentences and words. The syllables are counted and type-token lists are created. See the subsequent sections for more detailed information about these preprocessing steps.

In case of more complex file formats like HTML, RTF, TEX, and various word processor formats, the preprocessing phase would also include steps like determining the file type, separating meta information from real text, and annotating the text. This allows an advanced processing like the exclusion of text which limits the author's stylistic freedom, e.g. headlines or references. In this thesis, the advanced preprocessing is not performed as the analysis is limited to ASCII files.

Even the easy splitting procedures are far from trivial. In the following some of the problems with sentence and word splitting, letter and syllable counting are discussed. Solutions for these issues will be presented in chapter 6.

To be able to split a text into sentences, one must know what a sentence is. In Western languages, as English, punctuation marks serve as sentence delimiters. Full stops, exclamation marks and question marks are widely recognized as sentence delimiters (Hockey 2000, p. 110), although not all of their occurrences mark the end of a sentence.[1] The question is how other punctuation marks like colons and semicolons are handled. In some, but not all cases they fulfil the claim by Corns (1990) that a sentence should be grammatically complete. The approach used for splitting sentences will be presented in section 6.1.1.

The problem of splitting sentences into words (tokenization) is likewise complicated. Hockey's definition of a word as a "group of letters separated from other words by a delimitation character" (Hockey 2000, p. 53) shifts the problem to delimitation characters. Certainly, whitespace characters (e.g. blanks, tabulators, carriage returns) and punctuation marks delimit words. But how are words containing apostrophes and hyphens handled? Apostrophes can demarcate a genitive 's as in *my father's car* (which does not separate words) and contracted representations of a two word phrase as in *it's*. Hyphenated words may be counted as one word or several words. See section 6.1.2 for the definition of word used in this thesis.

Another issue is whether words in the text should be postprocessed. Are words in upper case and lower case handled as one type or as several types? Should a semantic categorization occur, i.e. should *in spite of* be treated as one token as its synonym *despite*? Is a part-of-speech tagging to be performed?

As indicated above, these issues will be solved in the implementation chapter, where the concepts will be defined, and regular expressions for sentence and word segmentation will be given.

[1] Full stops are also used to demarcate abbreviations, like Mr., Dr., etc. Exclamation marks sometimes mark an exclamation, as Alas! In these cases the punctuation marks are primarily no sentence delimiters.

5.2 Extraction of style markers

In the preprocessing phase the text was split into its constituents - sentences and words. In the next phase, the style markers are extracted from the text.

Like the preprocessing phase, the extraction of style markers is not unproblematic, and some terms have to be defined. As above, the rest of the section will bring up some of the issues, while the definitions and solutions to the problems are presented in the implementation chapter.

Some style markers evaluate syllable information. But what is a syllable? If a word contains only letters, the number of syllables can be estimated relatively easily by using heuristics (see section 6.1.3 for details). The matter gets complicated if words containing letters and digits occur. On the one hand, the five-character number *28144* is much easier to grasp than the 44 character and 12 syllable phrase *twenty-eight thousand, one hundred and forty-four*, but probably harder to get than other five letter words. How many syllables does *28144* have? See section 6.1.3 for the heuristic used for syllable counting.

But not only definition issues have to be tackled. How are the style markers stored internally? Should they be extracted each time one of the analysis options is changed, or should they be extracted once and saved in a temporary data structure? How would this temporary data structure look like? See the implementation chapter for suggested solutions to these issues.

5.3 Sliding window analysis

In the previous section the style markers were extracted from the preprocessed text. In the next step, the sentences are grouped together to examine the text at different levels, and to analyse how different levels affect the detectability of stylistic changes. A grouping is necessary because the analysis at a sentence level may not work because of noise, or because style markers are not applicable at levels of a few

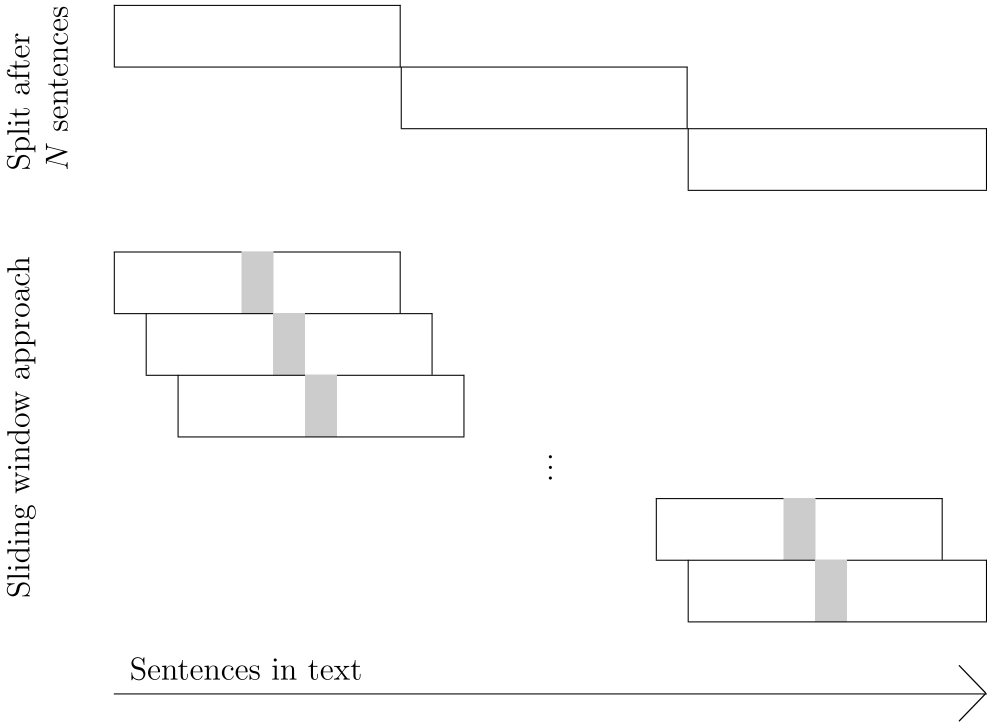


Figure 5.2: Text split into groups of N sentences with a 'normal' split after N sentences (top) and sliding window approach (bottom). The grey section demarcates the central sentence of the sliding window.

words.

A straightforward approach would be to split the text after every Nth sentence, and to analyse the resulting portions which are each N sentences long (see figure 5.2, top). The problem with this approach is that it may detect a stylistic change in one of the portions, but it cannot pinpoint *which* of the sentences in the portion are stylistically different. The approach is not fine-grained enough to detect changes at a very small level.

An approach which groups together sentences while preserving fine granularity is the sliding window approach. The sliding window approach focuses on a central sentence, but also considers sentences surrounding that central element. The advantage is that - apart from the first and last sentences in a group - every sentence is central in one text, and hence the groups largely overlap (figure 5.2). In that way, more data points can be evaluated than in the normal splitting method. Hence, the sliding window approach is finer grained and can detect changes which affect only some of the elements in the grouping. The advantages of grouping sentences, however, are preserved.

Another advantage of the sliding window approach is the flexibility. Besides the width of the sliding window, also the weights of the sentences in the window can be varied. It is, for example, possible that the weight factor decreases from the mean, giving more weight to the central sentence. See section 6.3.2 for examples of sliding window weight functions.

The sliding window approach also has shortcomings. For long sliding windows it is not possible to select the first and last sentences of a text as central sentences since the sliding window would be out of document bounds. Therefore an analysis of the beginning and the end of a document is not possible. Furthermore, by grouping sentences the sliding window approach blurs actual changes, and it becomes harder to detect small variations.

5.4 Visualization and analysis

In the previous steps, the style markers were extracted from the text and are now stored in a data structure. The next step is to process that raw data so that it can be easily interpreted. One alternative is to present the data graphically so that the users can “see” the data and draw conclusions from the visualizations. If there is enough data, statistical tests can be used to decide whether samples are significantly different or not.

In both cases, the evaluation method is unsupervised, i.e. it is not known to the algorithm which sections are plagiarized and how plagiarized sections look like.

5.4.1 Graphical presentation

A straightforward solution to visualize the results is to use one two dimensional coordinate system for each style marker. The x axis represents the sentences, the y axis the values of the style markers. These coordinate systems can be displayed by a program and can be analysed by the user. This representation is easy to implement and - since each variable is displayed separately - allows a statement *which* of the style markers detected a change in style. On the other hand, several graphs have to be analysed in parallel, and the user might fail to notice complex correlations between the variables. Furthermore, the analysis is subjective.

Multivariate analyses evaluate two or more variables at once and create a representation in an objective and reproducible way. The algorithms may find differences which humans might not have found by looking at the graphs. In the following, three multivariate approaches will be presented: hierarchical cluster analysis, principal components analysis (PCA) and self-organizing maps (SOMs).

Hierarchical cluster analysis produces a tree-like structure, a so-called *dendrogram*, to visualize similarity of elements. In the dendrogram, similar elements are grouped together in a small sub-cluster, which is itself included in another cluster of less similar elements (Oakes 1998). The dendrogram[4] in figure 5.3 shows that humans are closely related to chimpanzees, but relatively distantly related to mice.

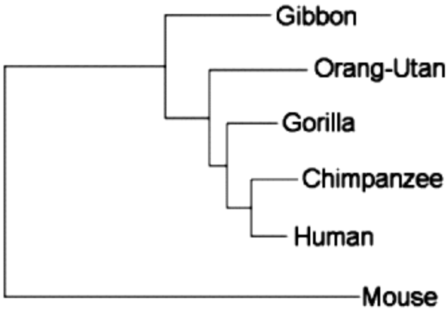


Figure 5.3: Dendrogram[4] showing the relationship between humans, various apes and mice. Adapted from the Phylodendron website[5].

[4] In the case of phylogenetic relationships (like here), the dendrogram is called ‘phylogenetic tree’. However, the tree is the result of performing cluster analysis with morphological or genome data, and therefore is a dendrogram.

[5] http://iubio.bio.indiana.edu/treeapp/treeprint-sample1.html

Principal components analysis (PCA) breaks down a highly dimensional input space (e.g. 50 variables) to a low dimensional space (e.g. 2 or 3 dimensions) and visualizes the resulting space in a coordinate system. Therefore, the original variables are transformed into a new set of uncorrelated variables, which are sorted in decreasing importance so that the first few principal components retain as much of the variation contained in the original variables as possible (Binongo and Smith 1999, p. 445). As a result, it is possible to represent many variables in a 2D coordinate system without losing much of the information contained in the data.[6] See Binongo and Smith (1999) for details on the PCA algorithm and its application to stylometry.

Self-organizing maps (SOMs) use neural networks to cluster input data. Figure 5.4 shows the general principle of SOMs: by adjusting the weights of the network (i.e. training), the nodes of the 3x2 grid migrate to fit the data points (Tamayo et al. 1999). See Kohonen (2000) and Lubovac (2000) for a description of the SOM algorithm. SOMs have the advantage that the number of resulting clusters and their arrangement (called *grid*) can be explicitly defined. Self-organizing maps are seldom used in stylometry but have been successfully applied in bioinformatics for interpreting gene expression patterns (Tamayo et al. 1999) and classifying cancer (Golub et al. 1999).

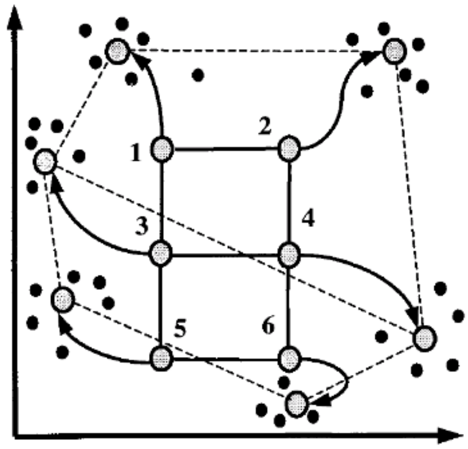


Figure 5.4: Principle of SOMs. The six nodes (numbered circles) migrate along the trajectories to fit the data points (black dots). From Tamayo et al. (1999, p. 2908, Figure 1).

[6] If two variables are correlated, the information contained in them is redundant. PCA removes that redundancy. In stylometry the first two principal components represent about 90% of the variation of the variables (Holmes and Forsyth 1995).

5.4.2 Statistical analysis

All graphical evaluations are at least to some degree subjective and must be interpreted. Statistical tests lead to completely objective results like "There is a significant difference", provided that the prerequisites are met. The prerequisites differ from test to test, but all tests need relatively much data to be able to draw conclusions. Therefore, at a sentence level the tests may not be applicable because the data base is too small to provide meaningful results. But wherever the tests fulfil the prerequisites, they are preferable to graphical evaluations.

The standard test for comparing means of two samples is the *t* test. Its advantage is that it is more powerful than non-parametric tests (Woods, Fletcher and Hughes 1986). The disadvantage is that the data must be normally distributed, otherwise "the reliability of the *t* test statistic may be compromised" (Sheskin 2000, p. 247). Therefore it must be checked if the data is really normally distributed.

Oakes (1998) suggests the χ^2 test to check whether data is normally distributed. Therefore the data is grouped into equally sized intervals. The resulting distribution (i.e. the observed data) is compared to normally distributed data (the expected data) with the χ^2 test, and a decision is made whether the null hypothesis that the data is normally distributed is rejected or not. The analysis whether the values of the style markers are normally distributed will be performed in a pre-study.

If the *t* test cannot be applied because the χ^2 test showed that the data is not normally distributed, non-parametric tests can be used to investigate if changes in a text are significant. Oakes (1998) suggests to use the Mann-Whitney *U* test and the median test. The Mann-Whitney *U* test ranks all data points and decides according to the ranks of each sample if the difference of two samples is significant. The median test evaluates how many data points of each sample are above and below the overall median, and derives therefrom if the data is significantly different. See Oakes (1998) for a detailed description of the algorithms.

Kilgarriff (1996) used the χ^2 test to measure the similarity of two corpora by comparing frequency lists. The overall χ^2 value is defined as the sum of the χ^2 values representing the deviation of one word in the frequency list (observed value) from the joint value of this word (expected value). This overall χ^2 value is the basis for the decision if two samples are significantly different. See Oakes (1998, p. 28) for a detailed description of this variation of the χ^2 test. In this study, this test will be used for measuring the similarity of two texts or text parts.

5.5 Summary

The analysis of texts with style markers can be classified into three phases: preprocessing, extraction of style markers, and graphical or statistical evaluation.

The preprocessing phase brings a text into a format from which the style markers can be easily extracted. This includes splitting the text into sentences and words, and handling of special file types. Since the issue of how to split a text is controversial, this chapter brought up some problems of text splitting to which solutions will be presented in chapter 6.

In the extraction phase the style markers are extracted from the preprocessed text. To facilitate the analysis at different levels, the style markers are grouped by the sliding window approach.

The data can be analysed by visualizing the data or by performing statistical test. Four visualization methods are proposed: a straightforward solution using coordinate systems, hierarchical cluster analysis, principal components analysis (PCA) and self-organizing maps (SOMs).

The t test has been identified as a powerful test for comparing sample means. Since the t test assumes normally distributed data, it will be checked in a pre-study if the values of the style markers are normally distributed. In case the data is not normally distributed, the Mann-Whitney U test and the median test will be used for analysis. The χ^2 test has been found suitable for comparing frequency lists.

Chapter 6

Implementation

In the method chapter some common problems concerning preprocessing and style marker extraction were brought up. This chapter will suggest solutions to these issues and provide definitions. Furthermore, some implementational details will be given.

6.1 Text Preprocessing

6.1.1 Sentence splitting

The first issue brought up in the method chapter concerned sentence splitting. For this thesis, the following definition of a sentence is used:

Definition 2. *A* sentence *is a group of words which is delimited from other sentences by one of the following sentence delimiters: full stop, exclamation mark, question mark, colon, and semicolon. Exceptions, for example punctuation marks denoting abbreviations, exclamations etc., must be handled.*

Colon and semicolon are added to the common list of sentence delimiters (.!?) since preliminary tests showed that most of these occurrences represent grammatically complete units (compare section 5.1).

Paul Clough developed a rule-based sentence splitter (Version 1.0. Paul Clough, Sheffield, Great Britain)[1] in Perl, which correctly disambiguates around 98% of the sentences in the British National Corpus and the Brown corpus (Clough 2001, pp. 17 and 19). Furthermore, the progam reliably detects abbreviations which do not end a sentence by performing a dictionary look-up. Since this algorithm outperforms other implementations (e.g. the Perl CPAN modules `Text::Sentence` and `Lingua::EN::Sentence`), it was chosen for this thesis. However, the regular expression was slightly changed to also handle sentences which are delimited by colons and semicolons, resulting in the following pattern (see Clough 2001 for a description of the expression):

```
$text =~ /([\’\"‘]*[({[]?[a-zA-Z0-9]+.*?)([\.!?:;])
          (?:(?=([(([{\"\’‘)}\]<]*[ ]+)[(([{\"\’‘)}\] ]*
          ([A-Za-z0-9][a-z]*))|(?=([()\"\’‘)}\<\] ]+)\s))/gs
```

The resulting single sentences are stored in a Perl array.

6.1.2 Word splitting

Definition 3. *A* word *is a group of letters which is delimited from other words by a whitespace character. Words containing apostrophes (e.g.* it’s*) and hyphens (e.g.* self-made*) are counted as one word.*

The decision to count words containing apostrophes and hyphens as one word is very common, and is for example used by other programs like the UNIX-tool *wc* and the grammar-checker of *Microsoft Word*. The following regular expression splits a text into words:

```
$text =~ /\b([\w\d][-’\w\d]*)\b/ig
```

The resulting split words are stored in a Perl array, which is itself part of the sentence array described above (array of arrays).

[1] Available at http://www.dcs.shef.ac.uk/∼cloughie/programs

6.1.3 Syllable counting

For counting syllables the Perl CPAN module `Lingua::EN::Syllable` (Version 0.251. Greg Fast, Aurora, IL)[2] is used. It uses a heuristic which generally counts each vowel group as one syllable and handles exceptions from this basic rule with an exception list. For simplicity, digits in words are handled like consonants, i.e. *28144* is counted as a one-syllable word, *Win4Lin* as a two-syllable word. The author reports off-by-one errors for about 10-15% of the words in a not further specified word list.

6.1.4 Type-token lists

For frequency lists and relative entropy, type-token lists must be extracted from the text. Therefore, the text has to be split into tokens (e.g. words, letter unigrams, bigrams, trigrams), and a list with every type and the number of its occurrences (tokens) has to be created.

For this work, types are not case-sensitive, i.e. *car* and *Car* are the same type. The classification into types is lexical rather than semantic, that is, *in spite of* and *despite* are two different types though they have the same meaning. On the other hand, the word *pole* is treated as one type, regardless if one speaks of a ski pole, a magnetic pole, or a pole position. Part-of-speech tagging is not performed.

Types are extracted at a sentence level, i.e. for each sentence one list of all types and the number of their occurrences (tokens) is extracted and stored in a Perl hash. These hashes are stored in an array (array of hashes). This extraction of type-token lists occurs only once at a sentence level. If, in the later analysis, one of the analysis options is changed (e.g., if the number of sentences to be grouped changes), the corresponding entries of the array are merged together to compute style marker values, and the type-token lists do not have to be extracted again.

[2] Available at: http://search.cpan.org/dist/Lingua-EN-Syllable/Syllable.pm

Four different types of lists are extracted: words, letters, letter bigrams (pairs of letters), and letter trigrams (triples of letters). For the three last lists, types containing characters other than letters (e.g. a␣a) are excluded since preliminary tests showed that the results are better when using types consisting only of letters.

For a text with N sentences, the type-token list extraction results in four arrays, which have each N entries. Each of these entries is a hash which contains the types and tokens of the sentence it represents.

If sentences are grouped together for analysis, the hashes can be merged, resulting in a new hash which contains a type-token list for that group.

6.2 Extraction of style markers

In the preprocessing phase, the text has been split into sentences and words, and type-token lists have been created. In the next phase, the values for the style markers are extracted. While the preprocessing steps are performed only once at a sentence level, the extraction happens each time the analysis options change.

Before the style markers are extracted, the sentences are grouped together. That means, if sentences 1 to 10 shall be analysed in a group, the first 10 entries of the sentence array from 6.1.1 are merged together. Moreover, the first 10 entries of each type-token array from 6.1.4 are merged.

The extraction of simple ratio measures and readability scores is straightforward. The information these measures need - number of sentences, number of words, and syllable information - can be easily extracted from the corresponding arrays. Likewise obvious is the calculation of vocabulary richness measures: the number of types occurring once, twice, or generally N times can be extracted from the hashes representing the type-token lists (see section 6.1.4). These variables just have to be inserted into the corresponding formulae, and the values of the style markers can be calculated.

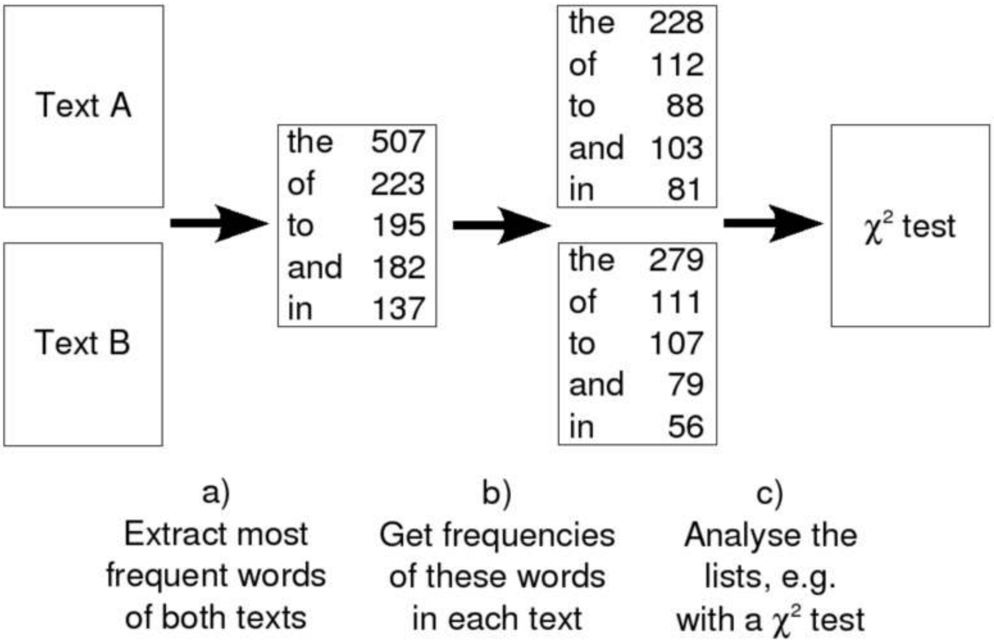

Text A
Text B
the 507
of 223
to 195
and 182
in 137
the 228
of 112
to 88
and 103
in 81
the 279
of 111
to 107
and 79
in 56
χ^2 test
a)
Extract most frequent words of both texts
b)
Get frequencies of these words in each text
c)
Analyse the lists, e.g. with a χ^2 test

Figure 6.1: The analysis of 2 texts by comparing the 5 most frequent words.

The extraction of frequency lists is more complex. This measure compares one section of a text to the rest of the text, or one text to another text. Therefore it is necessary to consider both the sentences for which the measure should be computed and the remaining sentences. From both of these parts (rather than just one)[3] the N most frequent types are extracted (see figure 6.1a). For these N types, the number of occurrences in each of the two parts are extracted, resulting in two frequency lists with each N elements (figure 6.1b). These two lists are used as data sets for a χ^2 test, as described in 5.4.2 (figure 6.1c).

[3] Often only one text (text A) is used for extraction, and then compared to another text B. Because the two texts may contain different vocabularies, the comparison of A to B may lead to other results than the comparison of B to A: the method is not symmetrical. If the lists are extracted from both texts, as done here, the analysis becomes symmetrical.

6.3 Sliding window approach

In section 5.3 the sliding window approach was introduced. This section will present two basic concepts of the method, sliding window width and sliding window weight functions.

6.3.1 Sliding window width

Sliding windows consist of a central sentence and 0 or more surrounding sentences. In this thesis symmetric sliding windows are used, i.e. equally many sentences surround the central sentence on both sides. The number of sentences left and right of the central sentence is denoted with *halfRange*. Since *halfRange* is a natural number, the sliding window width is always a odd number ($SWW = 2 \cdot halfRange + 1 \cdot centralSentence$). If the sliding window width is 1, *halfRange* is 0. In that case the sliding window consists just of the central sentence, and sentences are not grouped together.

6.3.2 Sliding window weight functions

Three sliding window weight functions were implemented. Figure 6.2 shows these three weight functions.

The "constant values" function weighs all values in the sliding window equally,

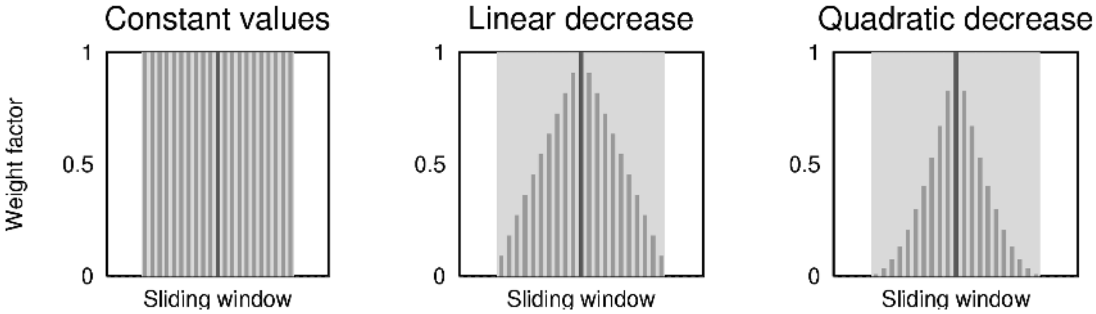


Figure 6.2: Three different sliding window weight functions.

therefore the weight W for a sentence with a distance of Δ from the central sentence is $W_\Delta = 1$.

The "linear decrease" function gives most weight to the central sentence, while the weight of the surrounding sentences in the sliding window decreases linearly with Δ. If *halfRange* is the number of the sentences in a sliding window which is left or right of the central sentence, the weight W for a sentence with distance Δ from the central sentence is defined by

$$W_\Delta = 1 - \frac{\Delta}{\mathit{halfRange} + 1}$$

The "quadratic decrease" weight function is similar to the constant decrease function, but the values decline quadratically from the mean:

$$W_\Delta = \left(1 - \frac{\Delta}{\mathit{halfRange} + 1}\right)^2$$

Apparently, these weight functions are only usable if the sliding window contains values which can be weighed. This is the case for simple ratio measures, readability scores, and vocabulary richness measures. If, however, the sliding window contains 'un-weighable' information, for example frequency lists, no weighing is performed.

6.4 Statistical tests

In this study four statistical tests are used: t test, Mann-Whitney U test, median test, and χ^2 test (see section 5.4.2).

For the experiments using the t test, the Perl CPAN module `Statistics::TTest` (Version 1.0. Yun-Fang Juan, Yahoo! Inc., Sunnyvale, CA)[4] was used. For the tests, the two samples (i.e. arrays of values representing the sample) are passed to the module, resulting in a statement whether the differences of the means are significant or not.

[4] Available at: http://search.cpan.org/author/YUNFANG/Statistics-TTest-1.0/TTest.pm

The existing Perl CPAN module for the χ^2 test (`Statistics::ChiSquare`) is not usable for this work since it just tests the randomness of a supplied dataset, but not whether two datasets are significantly different. Therefore a new module, also called `Statistics::ChiSquare`, was implemented. When speaking of the χ^2 implementation or the module `Statistics::ChiSquare`, it is referred to the newly developed module. After calculating the χ^2 value, the decision whether the difference is significant is made by looking into a hard-coded contingency table.

Since no Perl implementation for the Mann-Whitney *U* test was found, the module `Statistics::MannWhitneyUTest` was programmed. The implementation is based on the formulae from Oakes (1998, p. 17), which were directly converted into Perl source code. Similarly, the module `Statistics::MedianTest` was created, which is based on the description from Oakes (1998, p. 19).

Chapter 7

Pre-Study

As outlined in the background chapter, much work has been done on using style markers for attributing texts to authors. On the other hand, little attention has been paid to the question whether these style markers fulfil basic properties of a metric. Prerequisites like positivity, symmetry and triangle inequality are met by nearly every style marker,[1] while others, like the ability to fingerprint an author or the constancy with sample size, are less clear.

Especially the problem of a variable's constancy with sample length - which is a prerequisite for comparing texts of different lengths - is rarely addressed. Tweedie and Baayen (1998) is one of the few papers found discussing that issue for vocabulary richness measures; for frequency distributions and readability scores no such studies have been published. Furthermore, due to the lack of work that is available for intra-document analysis, it is unclear if the variables are generally applicable at a sentence or paragraph level. This pre-study will check whether the prerequisites of "fingerprintability" and constancy with sample size are met, and if the style markers can really be used for plagiarism detection.

[1] Some extraction methods, for example for extracting frequency lists, violate the symmetry axiom. This problem can be circumvented by varying the extraction method. See section 6.2 for more information.

For these reasons a pre-study is performed which should help to answer the following questions:

- Are the variables constant for different text lengths, so that sections of different size can be compared?
- Are the variables able to fingerprint an author, i.e. is the difference between sections of two authors significantly higher than the variations of sections written by one single author?

7.1 Data sets

Three novels from Project Gutenberg archives (PG)[2] were chosen as test sets for pre-study:

- Jane Austen (1775-1817): *Pride and Prejudice* (Austen 1813)
- Rudyard Kipling (1865-1936): *The Jungle Book* (Kipling 1894)
- Oscar Wilde (1854-1900): *The Picture of Dorian Gray* (Wilde 1891)

There were several reasons for choosing these three novels from PG. First of all, literary texts have been widely examined in authorship attribution studies (Baayen 1996; Burrows 1987; Holmes 1992; Hoover 2002; Khmelev and Tweedie 2001), and because these authors present encouraging results, it seems advantageous to use similar data sets. Secondly, prose is generally more related to scientific texts than drama or lyric and hence preferable. Project Gutenberg provides electronic versions of texts in ASCII format, so the data could be easily gathered. Furthermore, all PG texts are public domain and no legal issues had to be taken into account. Last, the

[2] Project Gutenberg web site: http://promo.net/pg

three novels appear to have different styles of writing,[3] and the style markers should - provided that they really work - detect changes.

All novels were downloaded from Project Gutenberg in ASCII text format, sections which are not part of the novels (like information about Project Gutenberg, copyright and legal information) were removed manually. Furthermore, the table of contents, chapter numbers and chapter headlines were removed.

7.2 Experimental setup

Two different experimental setups were chosen for the different types of style markers. Simple ratio measures, readability scores, and vocabulary richness measures produce one score for an analysed section. The experimental setup for these single-value measures will be described in the following section. The analysis of frequency distributions produces a list of results, hence the experimental setup differs from the single-value measures. The experimental setup for list-based measures will be described in section 7.2.2.

7.2.1 Single-value measures

In order to answer the questions of constancy and “fingerprintability”, two experiments are performed for single-value measures. In both experiments, the texts were split into portions of N sentences as shown in figure 7.1. In the following, this splitting into groups of N sentences is referred to as an analysis at a N-sentence level.

[3] *The Reference guide to English literature* characterizes *Pride and Prejudice* as a “romantic love story” (Kirkpatrick 1991, p. 1786), it is relatively easy to read. *The Jungle Book* as a children’s book is assumed to be easily readable and comprehensible. *The Picture of Dorian Gray* is the most demanding of the three novels, with its style appearing “artificially melodramatic” and “florid” (Radler and Jens 1988, p. 667, own translation).

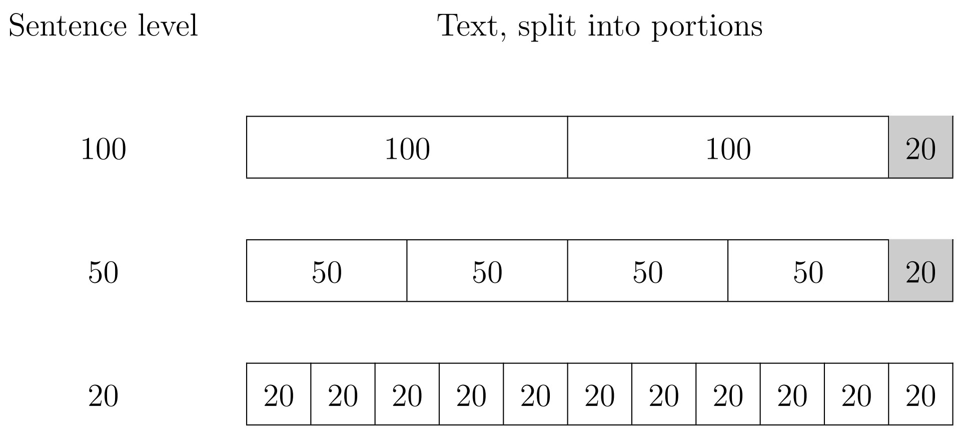


Figure 7.1: A text of 220 sentences, which is split into portions of 20, 50, and 100 sentences. The grey sections are residues, which are shorter than the sentence level.

Fingerprinting issue

For examining the fingerprinting issue, the three novels were split into portions of 500 sentences, i.e. the texts were analysed at a 500-sentence level. The last portion was omitted if it was shorter than 500 sentences, that means, the grey sections in figure 7.1 were ignored. The reason is that potential errors coming from different sample sizes shall be eliminated in this experiment.

This splitting lead to 17 portions of “Pride and Prejudice”, 6 for “The Jungle Book” and 8 for “The Picture of Dorian Gray”. For each style marker, a diagram is created which will be qualitatively evaluated in section 7.3.1.

Constancy with sample size

For testing the constancy of the style markers with sample size, the three novels were split into portions of 1, 2, 5, 10, 20, 50, 100, 200, 500, and 1000 sentences. The last portion is not omitted in this test, as a constant sample size is not required here. In other words, the grey sections in figure 7.1 are also included in the analysis. The results of each split are saved into a file, resulting in 10 result files for each novel.

The results are evaluated with three different statistical tests: t test, Mann-Whitney U test and median test. The null hypothesis for each of these tests is that there is no stylistic difference in the text. A 95% confidence interval is used. The results of the 465 tests[4] are grouped into correct acceptations and correct rejections. Low percentages of correct decisions might pinpoint problems with constancy, and will be further analysed by looking at visualizations.

In contrast to Mann-Whitney U test and median test, the t test is parametric and assumes normally distributed data. To test whether the style marker values are normally distributed, the texts are analysed at a 1-sentence level and at a 20-sentence level. All occurrences of values are grouped into 10 equally sized intervals, and intervals for $[0 \ldots min\_value[$ and $]max\_value \ldots \infty]$ are added. Following the suggestion by Woods, Fletcher and Hughes (1986, p. 144), intervals with less than 5 entries are merged with the previous interval. The resulting distribution (i.e. the observed data) is compared to normally distributed data (i.e. the expected data) by using a χ^2 test (see Oakes 1998 and Woods, Fletcher and Hughes 1986 for more detailed information). Besides that quantitative test, the distributions will be qualitatively evaluated by presenting graphs which show the groupings from above.

7.2.2 List-based measures

In the previous section, the t test, the Mann-Whitney U test and the median test were used for analysis. This is not possible for the analysis of list-based measures, since they do not result in one single value, but in a list of values. The frequency lists, however, can easily be analysed with a χ^2 test, as, among others, suggested by Oakes (1998, p. 28).

For this experiment, the three novels were split into chapters according to the chapter headlines or the chapter numbers. This results in 61 parts for “Pride and

[4] As 3 novels with each 10 result files exist, there are $\sum_{i=1}^{30} i = 465$ combinations (including the comparison of a section to itself) for each style marker.

Prejudice", 14 parts for "The Jungle Book" and 13 parts for "The Picture of Dorian Gray". For the Austen text only the first 10 chapters are analysed, for the other text all parts are used. Together with the complete texts (which are not split into chapters), this lead to 820 combinations of chapters or novels.[5]

For each of the 820 combinations, the 20 most frequent words, letter unigrams, letter bigrams and letter trigrams from both texts are extracted as described in section 6.1.4 and section 6.2.

The lists containing the 20 most frequent types are then used for performing a χ^2 test. The last step of the test, which compares the χ^2 value to a critical value and hence determines whether the differences are significant, is left out and the χ^2 values are compared to each other. This leads to finer grained results as the boolean output significant/not significant. The χ^2 values are a measure of similarity, where zero denotes equality. With increasing χ^2 values also the dissimilarity grows.

The χ^2 values are then assigned to one out of six groups representing the comparison it belongs to (e.g. Austen vs. Austen, Austen vs. Kipling etc.). It is expected that the groups comparing sections of one author have low values (denoting high similarity), while groups with sections from different authors have high values (denoting low similarity). The actual results are presented graphically and are qualitatively evaluated in section 7.3.1.

Relative entropy scores are evaluated similarly. Since the relative entropy measure itself is a measure of similarity and diversity (see section 3.2.1), a χ^2 test is not necessary. For testing the applicability of relative entropy, each of the 37 chapters described above is compared to the three complete novels, while the chapters are defined as the sample p and the complete texts as the population q. The resulting entropy scores are then assigned to one out of six groups described in the previous paragraph. Again, it is expected that groups comparing a chapter of one author

[5] In total, 10 Austen chapters +14 Kipling chapters +13 Wilde chapters +3 whole novels = 40 texts are compared. This leads to $\sum_{i=1}^{40} i = 820$ combinations (including the comparison of a text to itself).

to the complete novel of the same author have lower values than groups comparing chapters and novels of different authors.

7.3 Results

This section summarizes the results of the pre-study. Section 7.3.1 checks which style markers can fingerprint an author. Section 7.3.2 tests if the style markers are normally distributed, while section 7.3.3 addresses the issue of constancy.

7.3.1 Which style markers can distinguish authors?

In order to be able to detect changes of authorship, the style markers should be able to "fingerprint" an author. That means, the style markers should be more or less constant for different texts of one single author, but should show significantly different values for texts of two different authors. However, especially the second claim is not always fulfilled for every combination of authors. As McEnery and Oakes state, "discriminants that can distinguish Marlow and Shakespeare will not necessarily distinguish Goldsmith and Johnson" (McEnery and Oakes 2000, p. 550). However, even if one single style marker cannot distinguish two certain authors, a combination of various measures might still find differences.

Figure 7.2 shows the style marker values for the three novels by Austen, Kipling and Wilde. All graphs show that an intra-authorial constancy of style is not always given. In the second half of the Wilde text a clearly visible peak occurs. This is common to sentence length, Gunning FOG readability test, Flesch reading ease and Flesch-Kincaid grade level. The reason for that outlier is one single sentence in chapter 9 of "The Picture of Dorian Gray". It is 198 words long and was incorrectly split up: Though the sentence contains four semicolons (which usually delimit a sentence), it was not split because after each semicolon an *and* in lower case follows, which prohibits sentence splitting. If this sentence is left out, the outlier disappears.

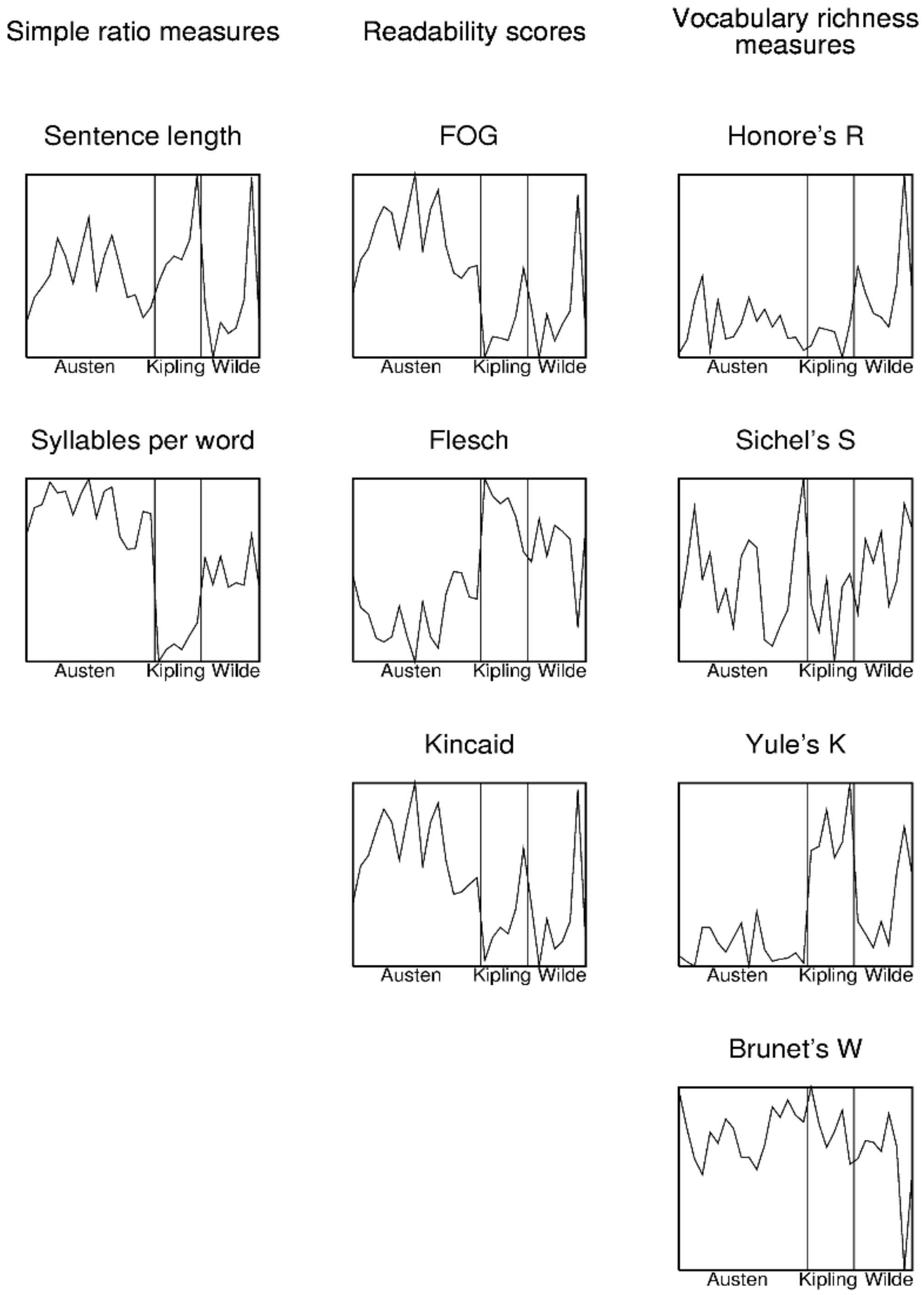


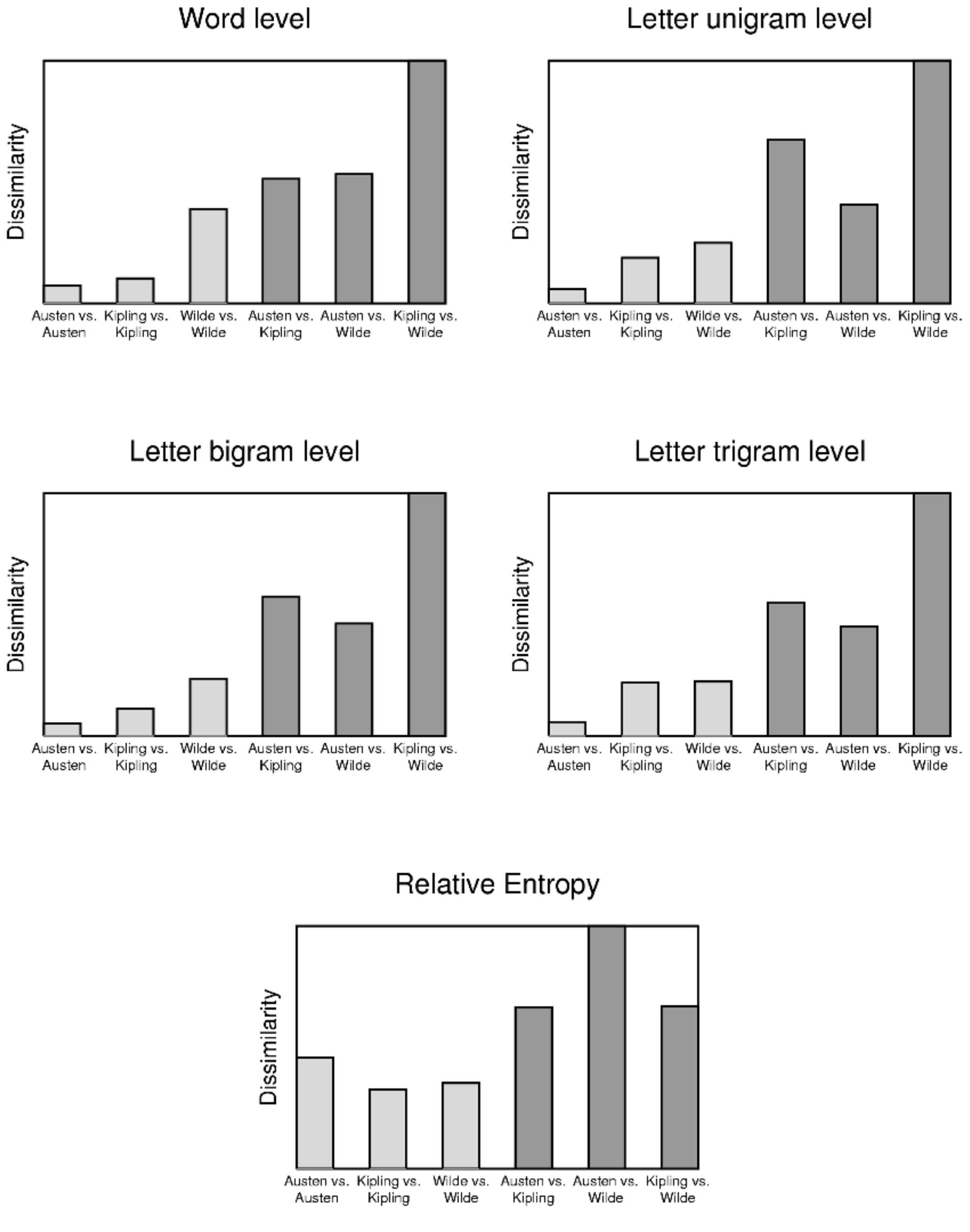
Figure 7.2: Simple ratio measures, readability scores and vocabulary richness measures for 500-word-long sections of the novels by Jane Austen, Rudyard Kipling and Oscar Wilde. The outlier at the end of the Wilde text comes from an incorrect sentence split. See the text for more information.

If the outliers coming from incorrect sentence splits are disregarded, most style markers can distinguish at least two authors. This is most clear for the syllables-per-word-measure, which has distinct values for all three authors. The three readability scores clearly separate Austen from the other two authors, and Wilde in general uses shorter sentences than Austen and Kipling.[6] The vocabulary richness measures separate the authors less well, especially Sichel's S and Brunet's W (even after disregarding the Wilde outlier) show intra-authorial variations which are of the same magnitude as the inter-authorial variations. Therefore these measures are not able to fingerprint an author.

In order to test if list-based measures can distinguish authors, χ^2 tests were performed and evaluated as described in section 7.2.2. The results from the tests are summarized in figure 7.3. The expected result that the three left bars in each graph (which represent comparisons of chapters of the same authors) have lower values than the three right bars (standing for chapters from different authors) is backed, although the difference is not very big. A more detailed analysis of the χ^2 values shows that the length of the sections does not affect the results. Therefore, frequency list measures seem to be constant for different text lengths. Since the two prerequisites of fingerprintability and constancy are fulfilled, frequency lists will be used in the main study.

Very similar results are observed for relative entropy (figure 7.3). Again, the predicted result that the comparison of chapters and novels of one author leads to lower values in the graph is confirmed. As in the analysis of frequency lists, the length of texts does not seem to affect the result. Therefore relative entropy will also be used as style marker in the main study.

[6] It is not clear why Jane Austen (and not Kipling or Wilde) is different from the other authors. One might claim that the variation is due to the fact that *Pride and Prejudice* was written 80 years earlier, or that Jane Austen is female. The analysis of Khmelev and Tweedie (2001), however, shows that epoch and gender do not affect the attribution results of texts.



Word level

Letter unigram level

Letter bigram level

Letter trigram level

Relative Entropy

Figure 7.3: Results for the chapterwise comparison of the three novels with a χ^2 test. All χ^2 values - which are measures of dissimilarity - of one author pair (e.g. Austen vs. Austen, Austen vs. Kipling) are grouped into one bar.

7.3.2 Are the style marker values normally distributed?

In section 7.3.3 statistical tests will be used to check whether the style marker values are independent of sample size. The t test has been used for similar tasks (Oakes 1998; Woods, Fletcher and Hughes 1986) and has the advantage that it is more powerful than non-parametric tests (Woods, Fletcher and Hughes 1986). The disadvantage is that the data must be normally distributed, otherwise "the reliability of the t test statistic may be compromised" (Sheskin 2000, p. 247). This section will check if the variables are normally distributed and if the t test can be used.

The Central Limit Theorem, which states that the distribution of N randomly selected values approximates normality for large N (Sheskin 2000, p. 58), cannot be applied here. The reason is that, depending on the skewness of the variable distribution, up to 100 samples are needed to ensure normality (Woods, Fletcher and Hughes 1986, p. 103). A granularity of 100 sentences, however, is without use for detecting plagiarism, since plagiarism occurs at a level of only a few sentences.

A χ^2 test as described in section 7.2.1 is used for checking if the data is normally distributed. The results are summarized in table 7.1.

At the 1-sentence level, the null hypothesis that the values of the style markers are normally distributed is always rejected. In other words, the values of the style markers are probably not normally distributed at the 1-sentence level.

At the 20-sentence level, this is not always the case. For the Gunning FOG readability test, the null hypothesis is rejected for none of the three novels, i.e. the data seems to be normally distributed. For four more style markers, the data seems to be normally distributed for at least one novel. On the other hand that means that the style marker values for 8 out of 9 measures are probably *not* normally distributed for at least one novel. The χ^2 test does not give hints how the data is actually distributed, it just shows that the distribution is not normal. The results did not improve when using alternative grouping methods, like using logarithmic instead of linear interval distances for grouping the values (Williams 1940). In any

		Simple ratio		Readability			Vocabulary richness			
		Sentence length	Syllables per word	FOG	Flesch	Kincaid	Honoré's R	Sichel's S	Yule's K	Brunet's W
1-sentence level	Austen	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Kipling	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Wilde	yes	yes	yes	yes	yes	yes	yes	yes	yes
20-sentence level	Austen	no	yes	no	yes	yes	yes	no	no	yes
	Kipling	no	no	no	yes	yes	no	yes	yes	yes
	Wilde	yes	no	no	yes	yes	yes	yes	yes	yes

Table 7.1: Is the null hypothesis that the values of the style markers are normally distributed, rejected at a 0.05 level? The results are calculated with a χ^2 test.

case, the χ^2 test shows that the data is not normally distributed, and therefore the t test cannot be used for evaluating the constancy issue.

Figure 7.4 gives an impression of how the style marker values are actually distributed for the three novels. Some graphs, especially the ones from the Austen novel, indeed have some resemblance to the normal distributions. Others, like Flesch Reading Ease and Brunet's W from "The Jungle Book", are certainly not normal distributions. These findings confirm the result of the χ^2 test that the t test cannot be used for evaluating if the style markers are constant for different sample lengths.

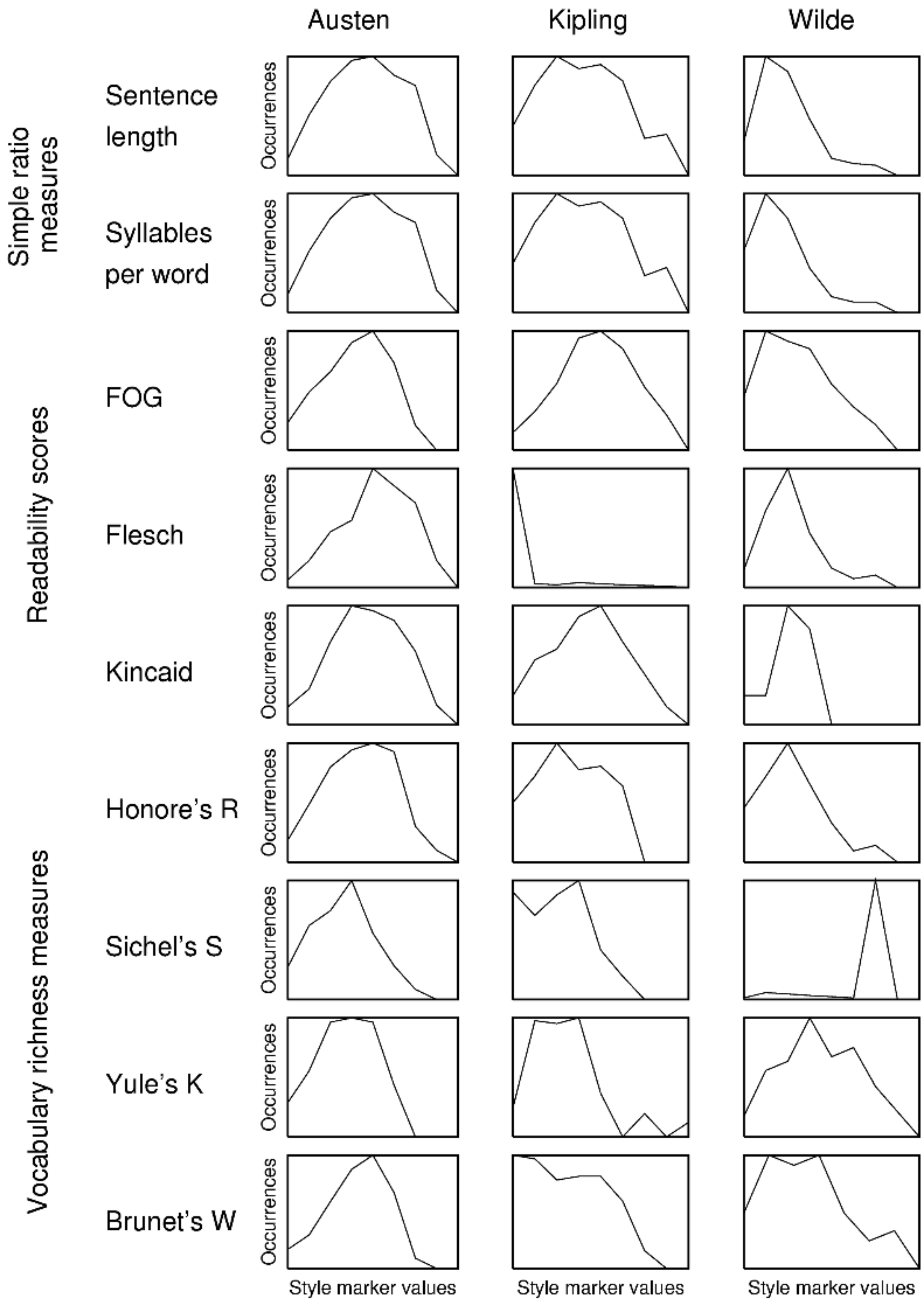


Figure 7.4: The distribution of style marker values for three authors. The texts are divided into portions of 20 sentences, the style marker values of these portions are grouped into one of 10 intervals. See the text for more information.

7.3.3 Which style markers are independent of sample size?

In the previous section it was shown that the style marker values are usually not normally distributed and that consequently the t test is not applicable to analyse texts. Therefore non-parametric statistical tests are used. These tests are not as powerful as the t test, but do not assume normally distributed data.

For this study the Mann-Whitney U test and the median test were used for comparing the literary texts at different levels. The null hypothesis is that there is no difference between the styles of the texts. For all the tests a 95% confidence interval was used.

Table 7.2 shows the results of the Mann-Whitney U test and the median test for the comparison of the 485 possible combinations of the novels at different levels. The table provides the percentages of correct acceptations (i.e. the two compared texts are the same texts, and the null hypothesis is not rejected) and correct rejections

		Mann-Whitney U test		Median test	
		Correct acceptations	Correct rejections	Correct acceptations	Correct rejections
Simple ratio	Sentence length	100.00%	40.33%	92.72%	74.00%
	Syllables per word	100.00%	58.33%	98.18%	66.00%
Readability	FOG	100.00%	43.00%	98.79%	66.67%
	Flesch	62.42%	44.67%	100.00%	21.00%
	Kincaid	100.00%	41.67%	100.00%	66.67%
Vocabulary richness	Honoré's R	62.42%	52.67%	83.03%	35.67%
	Sichel's S	42.42%	65.33%	64.24%	41.67%
	Yule's K	61.82%	37.67%	81.21%	24.00%
	Brunet's W	100.00%	6.33%	75.76%	50.67%

Table 7.2: Percentages of correct acceptations and rejections for the comparison of the three novels at different levels with two statistical tests.

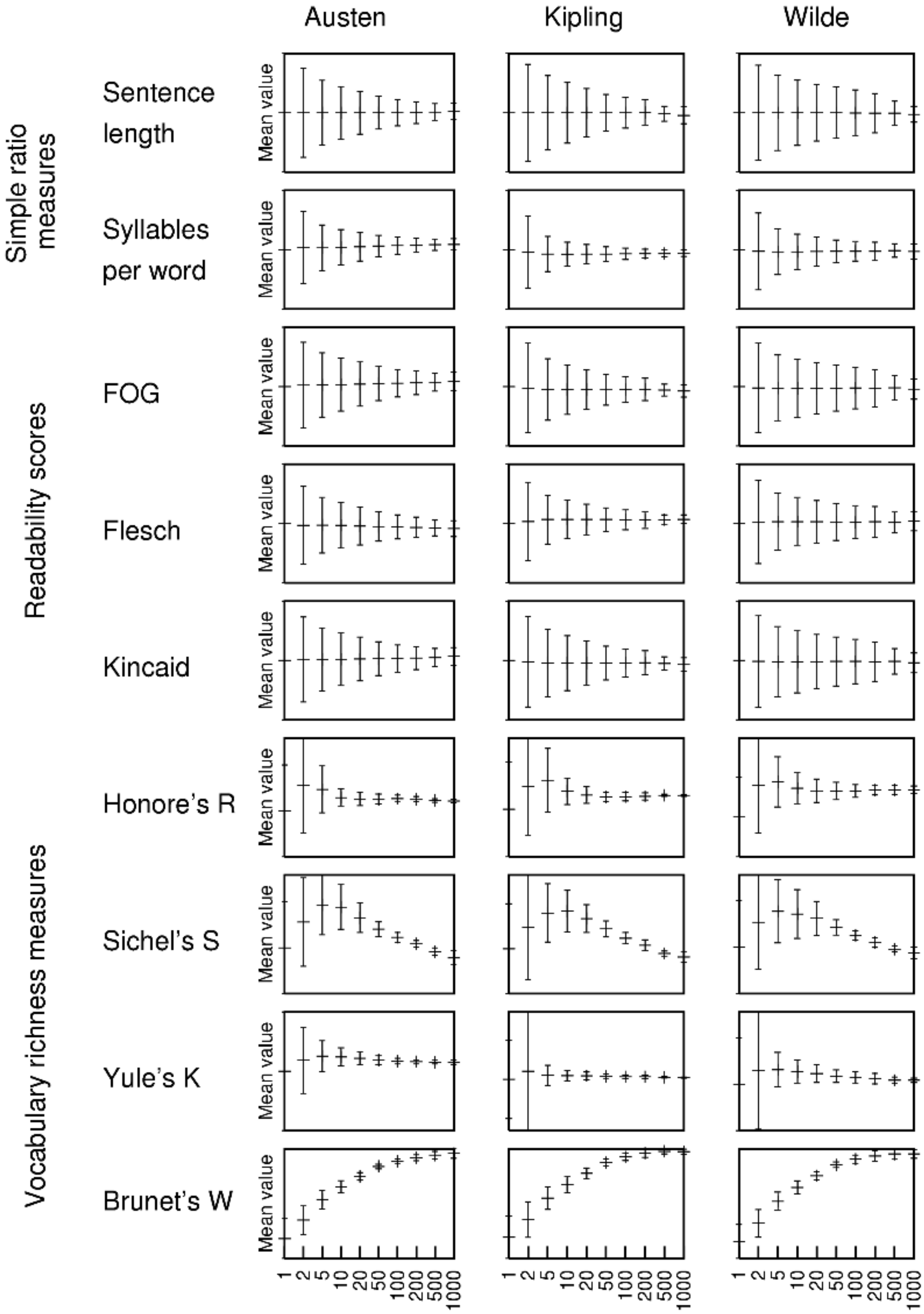


Figure 7.5: Means and standard deviations of style markers for three authors. The texts were split into portions of 1 to 1000 sentences, and the means and standard deviations of the style marker were calculated.

(i.e. the two compared texts are different novels, and the null hypothesis is rejected).

The table shows that both tests more often incorrectly reject the null hypothesis than incorrectly accept it. In other words, the tests make more type II errors than type I errors. If these detected changes really stood for plagiarism, this system could be called "student-friendly", as relatively few cases of cheating and plagiarism would be detected. In this case it is not a problem, since it is still possible to draw conclusions from the data.

Simple ratio measures and readability scores perform generally better than vocabulary richness measures, which especially produce more incorrect acceptations. This may be due to the relatively big intra-authorial variations, which has already been shown in section 7.3.1. Figure 7.5 shows another reason for the bad results: While the simple ratio measures and the readability scores show constant means for all sentence levels, the means for the vocabulary richness measures change with sentence level. This is most obvious for Sichel's S and Brunet's W, where the graphs clearly show that the mean varies with sample size. Therefore Sichel's S and Brunet's W do not fulfil the prerequisite of constancy and will not be considered further.

The mean also changes for Honoré's R and Yule's K, but the changes are overlapped by the high standard deviation in figure 7.5. Figure 7.6 clarifies that the

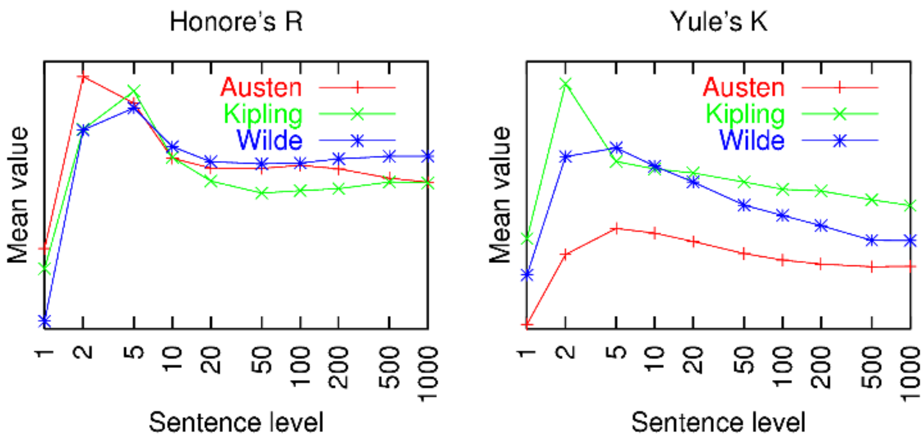


Figure 7.6: Means of Honoré's R and Yule's K for different authors.

mean of Honoré's R and Yule's K changes for different analysis levels. Furthermore, the values for different authors at the same level are very similar.

In section 7.3.1 it was shown that frequency list-based measures seem to be constant for different text lengths.

It can be concluded that all four vocabulary richness measures do - in contrast to simple ratio measures, readability scores and frequency lists - not fulfil the constancy prerequisite and therefore cannot be used for analysing text samples of different lengths.

7.4 Summary

In this chapter the style markers were analysed for their general applicability for plagiarism detection. It was checked if the style markers can fingerprint one author and if they are constant for different sample sizes. Three statistical tests (Mann-Whitney *U* test, median test and χ^2 test) were used. The *t* test could not be used since it was shown that the style marker values are not normally distributed. Three novels from Jane Austen, Rudyard Kipling and Oscar Wilde were used as test sets.

It was shown that simple ratio measures, readability scores, frequency lists, and relative entropy can distinguish different authors while being more or less constant for one author. Moreover, these measures are constant for different sample sizes. Therefore simple ratio measures, readability scores, frequency lists, and relative entropy will be used in the main study.

On the contrary, vocabulary richness measures do not fulfil basic properties of a variable. All four vocabulary richness style markers are not constant for analysis levels between 1 and 1000 sentences, making it impossible to analyse texts of different lengths. Apart from that, the intra-authorial variation is approximately as big as the inter-authorial variations, hence these measures do not have discriminatory power. Consequently, vocabulary richness measures will not be used in the main study.

Chapter 8

Main Study

In the previous chapter it was shown that vocabulary richness measures do not fulfil basic properties of a variable and hence cannot be used for plagiarism detection. On the other hand, simple ratio measures, readability scores, frequency lists, and relative entropy seem to be working since they can fingerprint authors and are constant for different sample sizes.

This chapter describes the main study, which uses the style markers that showed promising results in the pre-study, for text analysis. Furthermore, the newly developed specific words measure (see chapter 4) is used.

At first the data sets which are used are described. The following section explains how data is produced in a newly developed program called *JStynalyser* and how the results are loaded into two other analysis programs, *PAST* and *GeneCluster*. Section 8.3 summarizes the results of analysing artificially plagiarized texts with an unsupervised approach. The results will be more thoroughly discussed in chapter 9.

8.1 Data Sets

In chapter 7, three novels were used as data sets for checking whether different variables can be generally used for the detection of stylistic changes in texts. This thesis, however, focuses on shorter texts of a few thousand words, which is the approximate length of most student essays and assignments. Consequently, other data sets are needed for the main study.

8.1.1 Academic texts

Six texts with the topic "Censorship and Internet" or "Copyright and Internet" were chosen. Their length is between approximately 1000 and 4500 words, which is a typical length of papers submitted by students.[1] Three of the texts (Schmidt, Mboob, Anonymous) are papers written by students and were downloaded from Internet papermills, the other three (Stallman, Landier) were either published as essays on the Internet, or were published in journals or books.

The six texts which were chosen are:

- Peter Schmidt: Regulation of Internet Content (Schmidt 2001)
- Biram Mboob: Censorship and the Internet (Mboob 2001)
- Anonymous: Government Censorship and the Internet (Free Student Essays (Ed.) 1995)
- Richard Stallman: The Right To Read (Stallman 1997)
- Richard Stallman: Misinterpreting Copyright (Stallman 2002)
- Michael Landier: Internet Censorship is Absurd and Unconstitutional (Landier 1997)

[1] A spot check showed that around around 85% of the papers in http://a-termpapers.com are between 1000 to 4500 words (or 3 to 12 pages) long. Therefore this length can be called typical.

The texts were copied from the websites and saved in ASCII format. No further preprocessing was performed.

8.1.2 Plagiarized texts

The six academic texts were written by one author each and do - as far as we know - not contain cases of plagiarism. Therefore artificially plagiarized documents were created.

The algorithm for producing plagiarized documents is as follows:

- Out of the six academic texts, one document is randomly chosen,[2] the *main text*.
- Out of the remaining five texts, one is randomly chosen. This *insertion text* is the text from which is plagiarized.
- A user-specified number of sentences is extracted from the insertion text. The starting point is randomly selected.
- The extracted sentences from the insertion text are inserted at a random position of the main document.

In this manner, 200 artificially plagiarized documents with an inserted section of 1 to 200 sentences were created. The file `100_mboob_88_stallman1_-27.txt`, in which 100 sentences from Stallman's "Misinterpreting Copyright" were inserted into the Mboob text, was randomly chosen for a thorough analysis in section 8.3. In the following, this text will be referred to as the Mboob-Stallman text.

[2] The term *random* is problematic. Since a computer is completely deterministic, the results it produces are predictable and hence not random. Most of the random functions (like the Perl `rand` function which was used here) use *pseudorandom generators*: A sequence of resulting numbers is determined by an algorithm, but it appears to be random to an external observer who does not know that algorithm. This pseudo-randomness as the de facto standard for random number generation is sufficient for the purpose of creating plagiarized documents.

8.1.3 Random text

One problem in research is to prove that the observed effects are really caused by the variable under consideration and not by something else (internal validity) (Williamson 2002). In the case of internal plagiarism detection, for example, it must be precluded that measured changes are caused by other effects than stylistic changes. This is done by a "random text" which does not include plagiarized sections. If measured changes also occur with this random text, the hypothesis that these variations are caused by a stylistic change (which may pinpoint plagiarism) must be rejected.

For the "production" of a random text *Col's Random Sentence Generator* (Version 2002-05-09, Colin Frayn, Cambridge, Great Britain) (Frayn 2002), which creates random text in a two-step procedure, was used. In the first step the program extracts statistical information about an input text (like sentence length, word frequency, sentence ending characters, groups of words) and stores that information in a temporary file. The number of words to be grouped can be triggered by the correlation length variable; a smaller value groups fewer words and leads to more random sentences. In the second step the saved data from the temporary file is used to generate a text of a specified length.

For this thesis, the complete King James' Bible (Project Gutenberg (Ed.) 1989)[3] was used as input text. As being one of the longest books at Project Gutenberg, the Bible forms a very good data base for creating random texts. To further randomize the results, the correlation length was altered to 1. From the temporary file created from this data, a text of 250 random sentences was produced. 100 randomly selected adjacent sentences were marked to simulate a section of suspected plagiarism.

[3] Unlike the other texts obtained from Project Gutenberg, the Bible text was not further preprocessed. Sections like copyright and legal information were not removed, since these parts add more variety to the document - a desired property if a random text should be produced.

8.2 Experimental Setup

In the method chapter a couple of Perl scripts have been described, which, when put together, can stylistically analyse a text. Such scripts are sufficient for creating text-based result files without having to provide a complex user interface, and are ideal for performing hundreds of tests with varying texts or variables. Therefore the Perl scripts were well applicable for the analyses of the pre-study.

The disadvantage of these scripts is that they are invoked from the command line, and hence they are not very intuitive, especially for unexperienced users. Moreover, these tools do not allow the visual interpretation of the results.

In the following, several tools will be presented which overcome some of the shortcomings of the Perl scripts. In the next section, a newly developed Java program, *JStynalyser*, will be presented, which displays the results graphically and adds some new functionality. *JStynalyserBatch* omits the GUI and is optimized for exporting

	Program type		
	Perl scripts	JStynalyser	JStynalyserBatch
Programming language	Perl	Java	Java
Used in …	Pre-study	Main study	Main study
Triggerable from command line	partly[a]	no	yes
Batch processing possible	yes	no	yes
GUI	no	yes	no
Ease of use for novice users	low	high	low
Visualization of results	no	yes	no
Export into *PAST*, *GeneCluster*	no	yes	yes
Sections can be marked	no	yes	yes

[a] Usually small Perl scripts are written which start the analysis scripts.

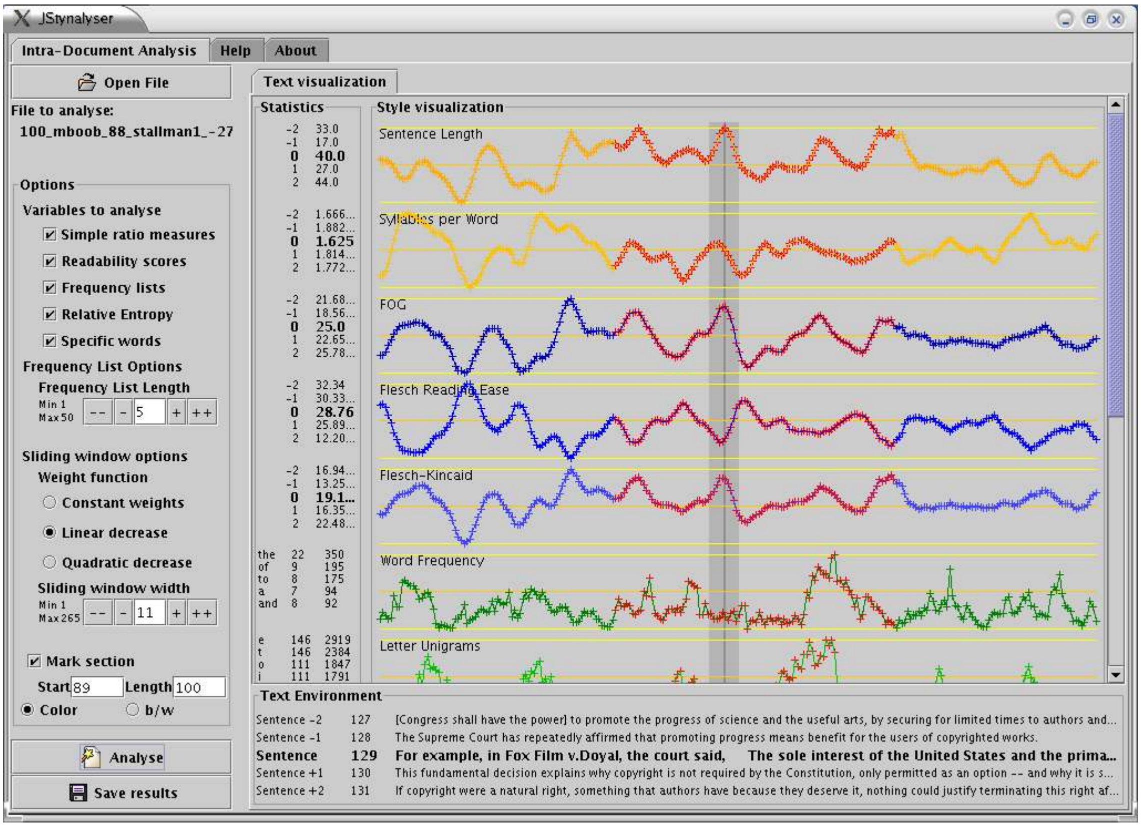
Table 8.1: Characteristics of Perl scripts, *JStynalyser* and *JStynalyserBatch*.

data for other programs. Table 8.1 summarizes the characteristics of these two Java programs and the Perl scripts. The subsequent sections explain how result files are created, and how they are handled in two further tools, *PAST* and *GeneCluster*.

8.2.1 *JStynalyser* and *JStynalyserBatch*

In order to simplify the analysis of texts, a Java program called *JStynalyser* (Version 1.0. Marco Kimler, Skövde, Sweden)[4] was written, which performs the same analyses as the Perl scripts, but adds an easy-to-use graphical interface and some more features. Figure 8.1 shows a screenshot of *JStynalyser*.

After opening an ASCII file, *JStynalyser* extracts the base measures from the text by using the original Perl scripts. The user can specify analysis options like



File to analyse: 100_mboob_88_stallman1_-27

Options

Variables to analyse
- ☑ Simple ratio measures
- ☑ Readability scores
- ☑ Frequency lists
- ☑ Relative Entropy
- ☑ Specific words

Frequency List Options
Frequency List Length (Min 1, Max 50): 5

Sliding window options
Weight function
- ○ Constant weights
- ● Linear decrease
- ○ Quadratic decrease

Sliding window width (Min 1, Max 265): 11

☑ Mark section
Start 89 Length 100
● Color ○ b/w

Statistics

```
-2   33.0
-1   17.0
 0   40.0
 1   27.0
 2   44.0

-2   1.666...
-1   1.882...
 0   1.625
 1   1.814...
 2   1.772...

-2   21.68...
-1   18.56...
 0   25.0
 1   22.65...
 2   25.78...

-2   32.34
-1   30.33...
 0   28.76
 1   25.89...
 2   12.20...

-2   16.94...
-1   13.25...
 0   19.1...
 1   16.35...
 2   22.48...

the  22   350
of   9    195
to   8    175
a    7    94
and  8    92

e    146  2919
t    146  2384
o    111  1847
i    111  1791
```

Text Environment

Sentence -2	127	[Congress shall have the power] to promote the progress of science and the useful arts, by securing for limited times to authors and...
Sentence -1	128	The Supreme Court has repeatedly affirmed that promoting progress means benefit for the users of copyrighted works.
Sentence	**129**	**For example, in Fox Film v.Doyal, the court said, The sole interest of the United States and the prima...**
Sentence +1	130	This fundamental decision explains why copyright is not required by the Constitution, only permitted as an option -- and why it is s...
Sentence +2	131	If copyright were a natural right, something that authors have because they deserve it, nothing could justify terminating this right af...

Figure 8.1: Screenshot from *JStynalyser*.

[4] See http://stynalyser.sourceforge.net for more information.

sliding window options and frequency list length. *JStynalyser* visualizes the data in different graphs, alternatively in colour for analysis on the screen (see figure 8.1) or in black-and-white for analysis on paper, as in figures 8.2 to 8.4. Since the focus is on detecting relative changes and not on absolute values, values are not displayed in the graph. However, the user can click on a suspicious data point and gets the absolute values of the style markers at this point and surrounding sentences. Furthermore a certain section in the text, for example suspected plagiarism, can be marked.

The user can export the results of the analysis into text files. Supported file formats are the PAST format used by *PAST* and the GCT format used by *GeneCluster*.

JStynalyserBatch is a command line tool which performs exactly the same tasks as *JStynalyser*, but takes its arguments from the command line. By writing a shell script, the generation of result files can be easily automated.

8.2.2 Producing data in *JStynalyserBatch*

By using *JStynalyserBatch*, results file sets consisting of each 96 result files (all combinations of 3 sliding window weight functions, 4 sliding window widths, 4 frequency list lengths and 2 file formats) are created. In this manner result file sets for the files `100_mboob_88_stallman1_-27.txt` and the random bible text were created.

8.2.3 Results in *PAST*

JStynalyser visualizes the results with coordinate systems, but does not support multivariate analyses, which are often used in authorship attribution studies and allow an easier interpretation of the data (see sections 3.2.2 and 5.4.1). Principal components analysis (PCA) is one of the most frequently used analysis methods and has been used in Burrows (1992), Holmes and Forsyth (1995), Baayen (1996), and Stamatatos, Fakotakis and Kokkinakis (1999). Hierarchical clustering has for example been used in Holmes (1992), Clough (2000a), and Hoover (2001).

The Windows program *PAST* (Version 1.08. Øyvind Hammer and Daniel A.T.

Harper, Oslo, Norway and Copenhagen, Denmark) (Hammer, Harper and Ryan 2001)[5] supports both hierarchical clustering and PCA. It is a handy tool for this purpose, as its interface is similar to common spreadsheet applications, the analyses are fast and the resulting graphics easily exportable.

The PAST-data which has been exported by *JStynalyserBatch* is loaded and "commatized".[6] Then the data to be analysed is selected. In case of PCA this is the whole sentence set with all variables. Due to limitations of the program, for hierarchical cluster analysis only the first 210 sentences are selected.[7]

For PCA analysis, the *Shape PCA* option is used. For cluster analysis, the paired groups algorithm with Euclidean distances is used.

8.2.4 Results in *GeneCluster*

GeneCluster (Version 2.13 Beta Build 20020926. The MIT Center for Genome Research, Cambridge, MA)[8] is a Java program built for gene expression analysis, but its good support of self-organizing maps makes it useful also for the analysis of texts.

The GCT-data which has been exported by *JStynalyserBatch* is loaded. The class finding algorithm using the standard values produces two SOM clusters, one with a 1x2 grid, the other one with a 1x3 grid. The data sets each cluster contains can be inspected by clicking on the clusters.

[5] See http://folk.uio.no/ohammer/past for more information

[6] Perl or Java represent floating point values with points (e.g. 3.1415), but *PAST* uses the continental European number format (e.g. 3,1415). The commatizing option in *PAST* converts the numbers.

[7] *PAST* version 1.08 can cluster a maximum of 210 datasets. Version 1.09 removes that limitation, but produces several graphics, which would have to be manually reassembled in a graphics program. Since tests with version 1.09 showed that the inclusion of the remaining sentences (around 30) of the Mboob-Stallman text does not significantly affect the clusterings, version 1.08 was used.

[8] See http://www-genome.wi.mit.edu/cancer/software/genecluster2/gc2.html for more information

8.3 Results

The last section described how results for the data sets from 8.1 are produced in *JStynalyser*, *JStynalyserBatch*, *PAST* and *GeneCluster*. This section will present the results from the analysis of the artificially plagiarized texts. The focus will be on the Mboob-Stallman text described in section 8.1.2, whereas the analysis of the other texts will be referred to if comparison is needed. Four evaluation methods will be used: visual inspection, hierarchical cluster analysis, principal components analysis (PCA) and self-organizing maps (SOMs).

8.3.1 Visual inspection

JStynalyser visualizes the results of the analysis of a text in one graph for each style marker. On the one hand, that means that several graphs have to be analysed in parallel and complex correlations between the variables could be missed. On the other hand, this more fine grained analysis allows a more detailed inspection and a statement *which* of the variables really vary and actually detect a difference. Therefore the rest of the section separately analyses simple ratio measures, readability scores, frequency measures, relative entropy, and specific words.

Simple ratio measures and readability scores

Figure 8.2 shows the visualization of simple ratio measures and readability scores for four different sliding window widths and the three sliding window weight functions. The section of 100 sentences by Stallman, which was inserted (i.e. plagiarized) into the original Mboob-text, is marked with a grey background.

If a sliding window of 1 is used, the intra-authorial stylistic changes (i.e. noise) is too big to be able to detect any inter-authorial differences. The bigger the sliding window gets, the more the noise is eliminated. On the other hand, bigger sliding windows blur the transition between stylistic changes, and it becomes hard to find

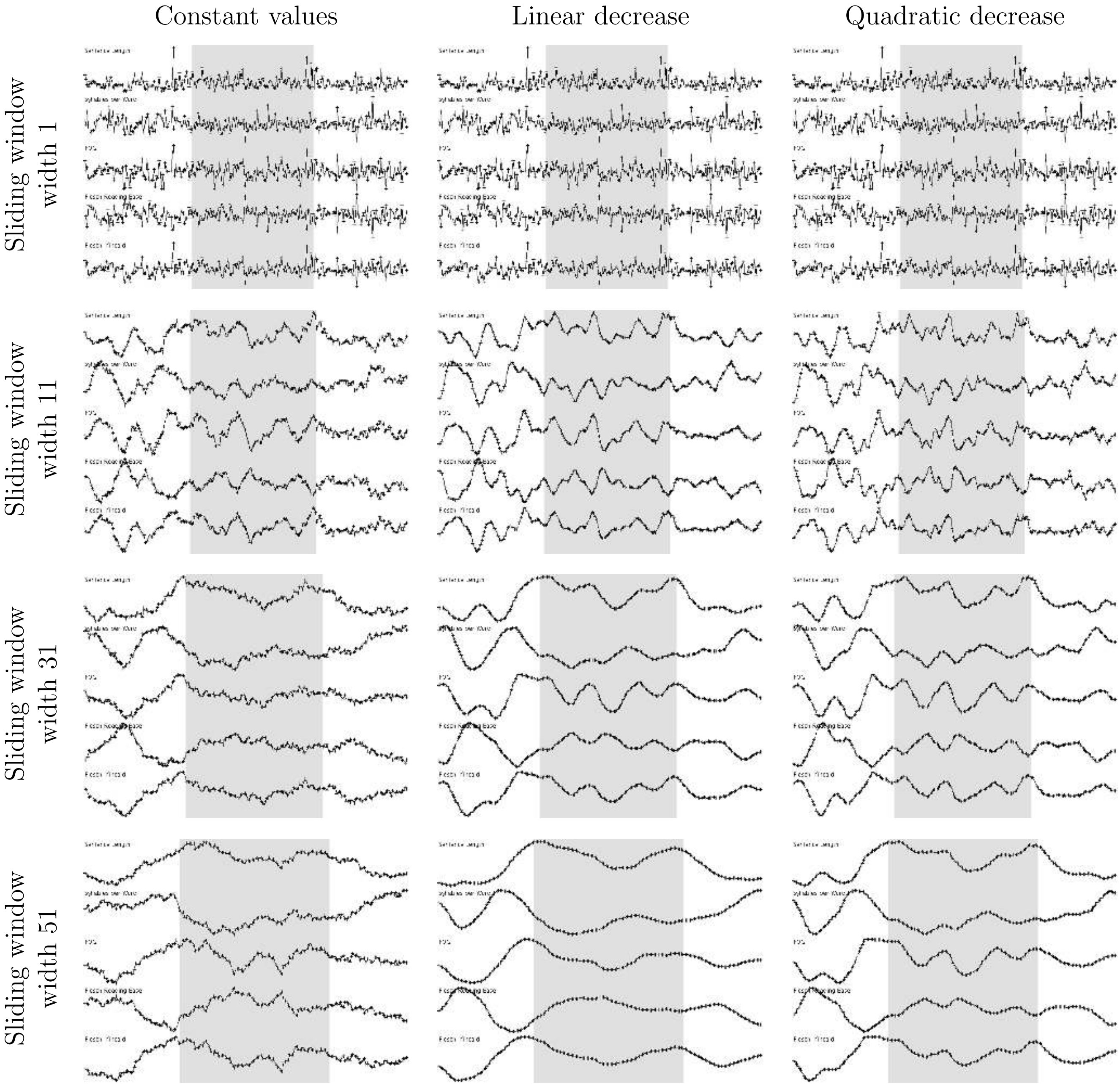
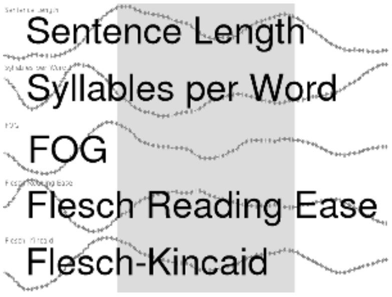


Figure 8.2: Simple ratio measures and vocabulary richness measures for different sliding window widths and weight functions for the Mboob-Stallman text. The section of 100 inserted (=plagiarized) sentences is marked grey. The figure on the right shows where in the graphs each style marker is to be found.



a threshold to distinguish authors.

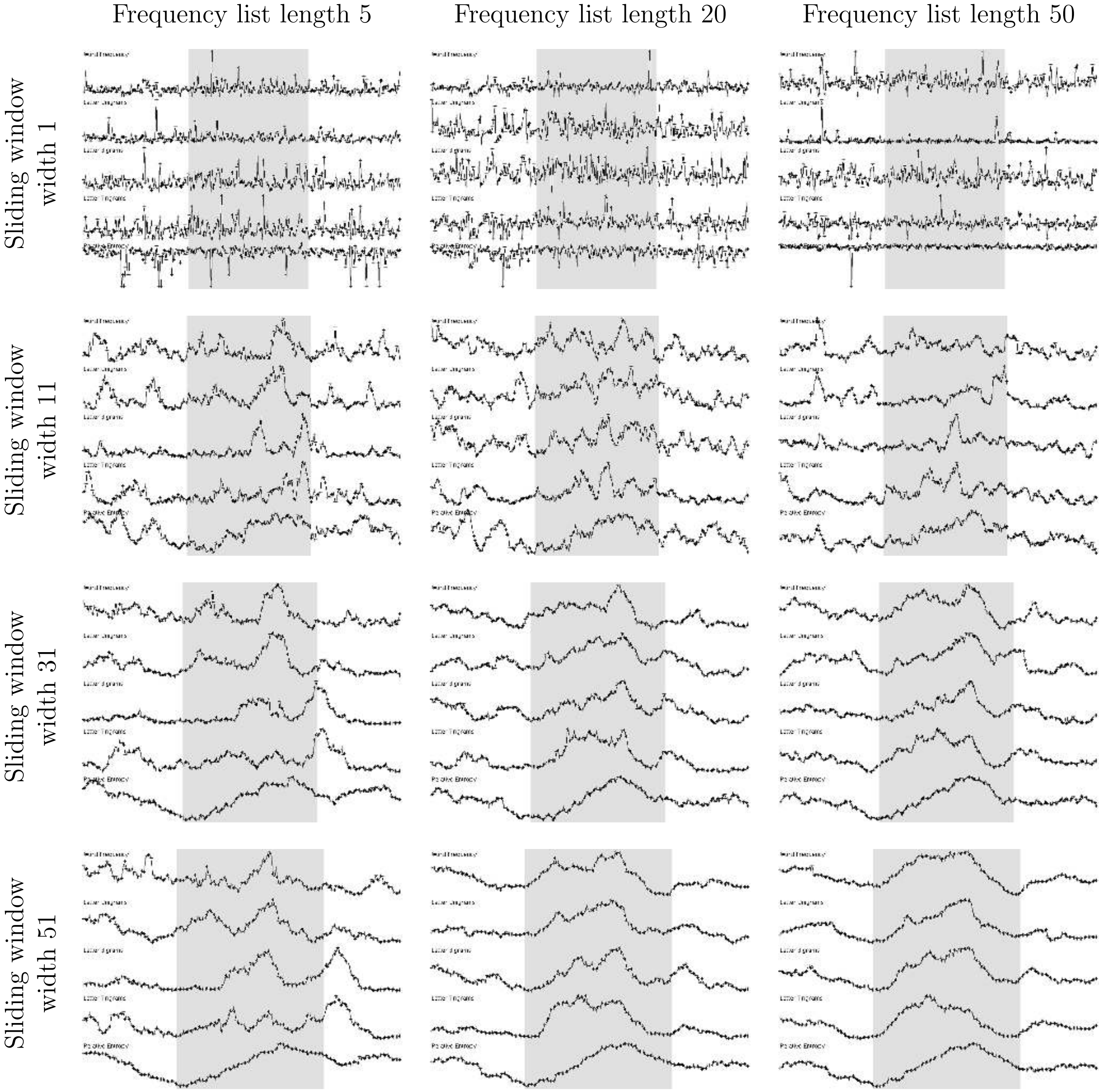
The sliding window weight function also has an effect on noise. The noise seems to be biggest if constant weights for all values in the sliding window are used. The graphs for linear and quadratic decrease are both much smoother. If the values in the sliding window decrease linearly from the mean, the noise seems to be eliminated best; if the values decline quadratically, the surrounding values are less emphasized as if linear decrease is used, and the effects of grouping are less strong. This results in slightly more noise.

The question whether simple ratio measures and readability scores can be used to distinguish authors in the case of the Mboob-Stallman text cannot be easily answered. Especially the graphs for linear decrease and sliding window widths 31/51 show that Richard Stallman tends to use longer sentences than Biram Mboob. However, it is hard to identify where exactly the plagiarized section starts and ends. The other measures cannot distinguish authors, as the first part of the Mboob text contains both the minimal and the maximal values of the *whole* text. Accordingly, the values for the Stallman part are in between these extreme values, and the two parts become indistinguishable for these variables.

Similar results were observed for the other artificially plagiarized texts: none of the style markers shows significantly different values for main and insertion text, so that the two parts cannot be distinguished by evaluating the style marker values.

Frequency lists and relative entropy

Figure 8.3 shows the visualization of frequency lists and relative entropy for three different sliding window widths. Since again the weight functions only affect the noise, but not the overall result, only the linear decrease weight function is used. Three different frequency list lengths are displayed. Again, the Stallman-part is marked grey. Concerning sliding window width, the same blurring-effect as described above occurs.



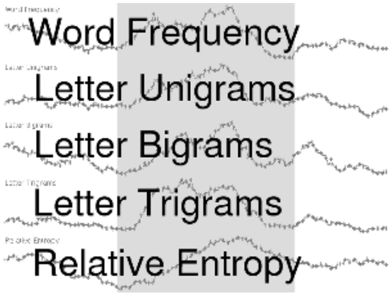


Figure 8.3: Frequency lists and relative entropy for different sliding window widths and frequency list lengths. The linear sliding window weight function is used. The Stallman section is again marked grey. The figure on the right shows where in the graphs each style marker is to be found.

Again, the question of whether authors are distinguishable is problematic. For small sliding windows, the noise is too big to draw conclusions; but even for big sliding windows no clear border can be found. For sliding window size 51 and frequency list length 50, Stallman seems to have higher values for the frequency measures, but this trend is unstable and not very clear. If longer frequency lists are used, the distinguishability seems to increase a bit, though the question of where to set the border remains. Especially the last sentences of the Stallman section show much lower values than the rest of the plagiarized section. These sentences much more resemble Mboob than Stallman. The relative entropy measure leads to results which are comparable to the frequency list measures.

What figure 8.3 does not show is whether the graphs are significant. The graphs of the frequency measures visualize the results of a χ^2 test, and one of the prerequisites of a χ^2 test is that all frequencies in the input data sets are "sufficiently large" (Woods, Fletcher and Hughes 1986, p. 144). Woods, Fletcher and Hughes (1986) note that frequencies below 5 lead to an increased likelihood of type I errors. That means that in order to get correct results, the least frequent token in the list should occur at least 5 times.

	Frequency list length N								
	5	10	15	20	30	40	50	70	100
Words	20	26	39	44	73	106	120	186	281
Letter unigrams	1	1	3	5	—[a]	—	—	—	—
Letter bigrams	5	8	9	10	12	16	19	25	34
Letter trigrams	19	25	32	35	45	56	62	74	90

[a] Because the alphabet has only 26 letters, it is not useful to use longer lists.

Table 8.2: How long must the sliding window be so that all tokens occur on average 5 times or more?

Table 8.2 shows how long the sliding window should be so that the Nth frequent type contains on average 5 tokens. Although that does not mean that all types occur 5 times or more, this allows an approximate assessment of how big the data sets should be to produce useful results. The table shows that letter unigrams are perfectly usable for intra-document analysis, since they require a sliding window width of just 1 to 5 sentences so that the least frequent type occurs 5 times. If short frequency lists (20 entries or less) are used, also letter bigrams are usable, since bigrams require a sliding window of 10 sentences or less at this level.

On the other side, the use of words and letter trigrams is problematic. Even for small frequency lists at least 20 sentences must be grouped together to find 5 occurrences of the least frequent word or letter trigram. Groups of 20 sentences, however, are longer than usual paragraphs, and therefore the analysis is not longer at a paragraph level. If shorter sliding windows than given in table 8.2 are used, the likelihood of type I errors in the χ^2 test increases.

Specific words

Figure 8.4 shows the visualization of specific words for three different sliding window widths and three different sliding window weight functions. Like for the other style markers, there is no detectable difference between the Mboob sections and the Stallman part. Therefore it seems impossible to use the specific words measure to pinpoint a stylistic change of a text of 100 sentences.

At a sentence level, however, specific words seem to be very well usable to detect stylistically different sentences. Figure 8.5 shows that the 6 most prominent outliers in the Mboob-Stallman text all represent sentences which are directly or indirectly cited. In other words, the stylistic change coming from different authorship can be detected in these 6 cases because extreme values (outliers) occur in the graph of the specific words measure. If more, less prominent data points are added in the analysis, false positives (i.e. sentences which are stylistically different although no change of

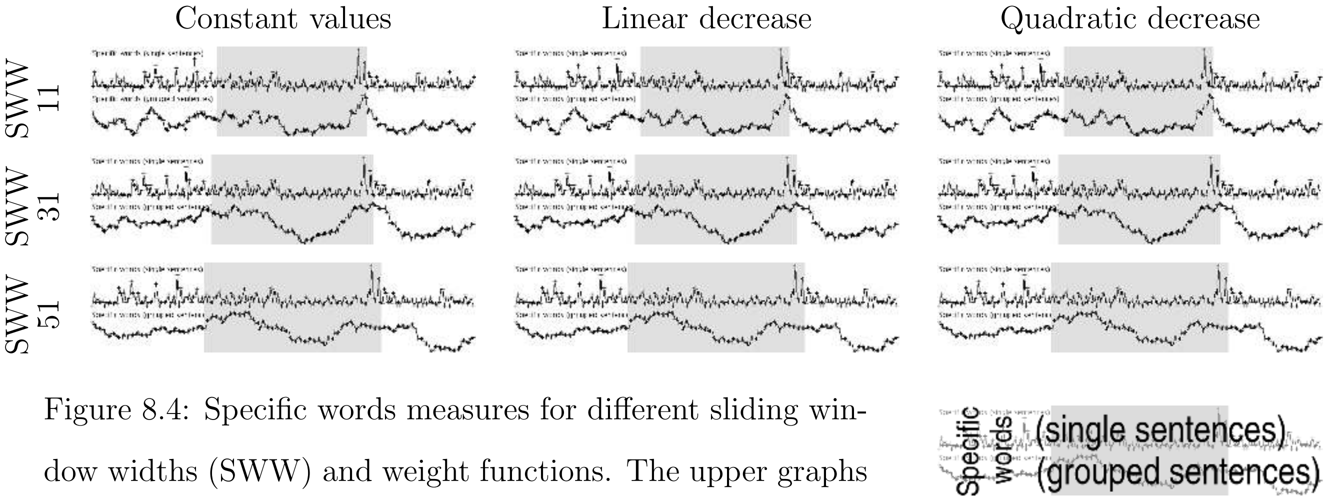
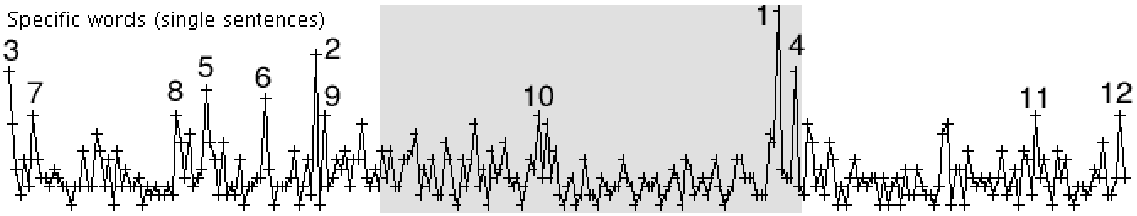


Figure 8.4: Specific words measures for different sliding window widths (SWW) and weight functions. The upper graphs denoted with *single sentences* visualize ungrouped sentences, i.e. SWW=1.

authorship occurs) appear. In the example of figure 8.5 the addition of further 6 data points leads to 3 false positives (25%). In other words, the precision[9] drops from 100% to 75%. The analysis of the other academic tests leads to corresponding results. An in-depth quantitative analysis, which might back these findings and might assess recall percentages, has not yet been performed due to the limited time.

As indicated above, the detection of different authorship by counting specific words works well for single sentences, but not for many sentences, like the inserted 100-sentence fragment in the Mboob-Stallman text. The reason is as follows: The specific words measure assumes that different authors use a different vocabulary, while the use of the vocabulary is similar for sentences of the same author. In other words: If a single sentence from another author is included in a text, it contains a lot of new words which do not occur in the rest of the text. The specific words measure will detect these words, and show an outlier. If, however, many sentences (let us say 50) are included, the new words that are used by the author of the

[9] Precision measures the proportion of true positives in the sample to all data sets in the considered sample. Recall is the proportion of positives in the sample to all positives in the population (in this case, all sentences of another style) (Oakes 1998, p. 176).



No.	Score	Sentence number	Section	Description	Change in style / Potential plagiarism
1	22	184	Stallman	Direct citation	yes
2	17	74	Mboob	Direct citation	yes
3	15	1	Mboob	Direct citation	yes
4	15	188	Stallman	Direct citation	yes
5	13	48	Mboob	Indirect citation†	yes
6	12	62	Mboob	Definition	yes
7	10	7	Mboob	Common knowledge	no
8	10	41	Mboob	Own formulation‡	no
9	10	76	Mboob	Indirect citation†	yes
10	10	127	Stallman	Direct citation	yes
11	10	245	Mboob	Indirect citation†	yes
12	10	265	Mboob	Own formulation‡	no

†A reference to the original work is not given. However, this sentence contains special knowledge which is probably copied from a source.

‡Subjective classification, since plagiarism can never be excluded.

Figure 8.5: The figure visualizes the ungrouped specific words scores for the Mboob-Stallman text. The 12 sentences with scores of 10 and above ("outliers") are numbered and analysed in the table below.

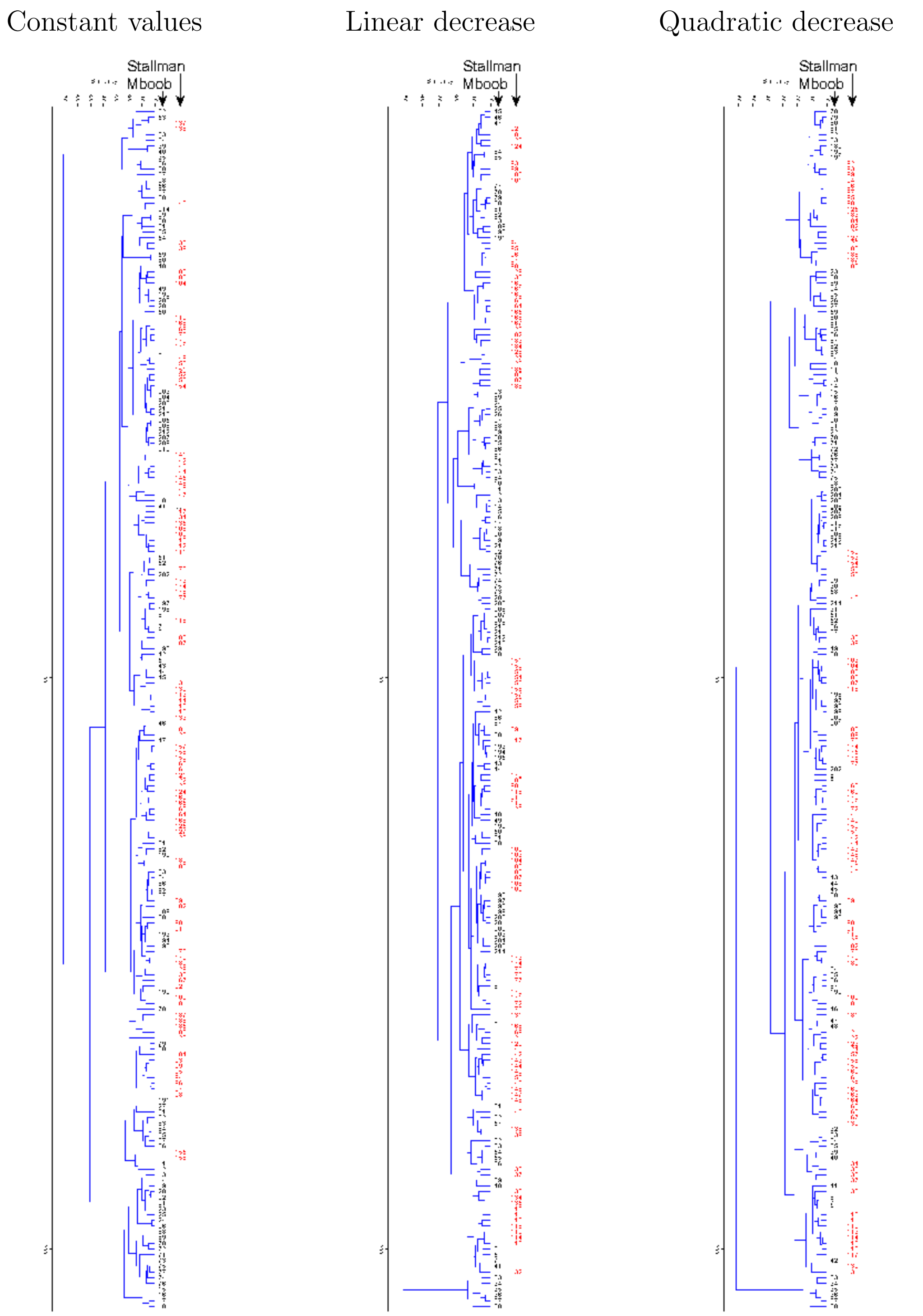
inserted passage spread over 50 sentences. The outliers from before become hardly noticeable. Still, the vocabulary from the 50 inserted sentences is different from the vocabulary used in the rest of the text, but the specific words measure cannot detect this change. Hence the use of the measure seems to be limited to pinpoint single inserted sentences. These single sentences, however, are very easily detectable.

8.3.2 Hierarchical cluster analysis

The problem with the visualization in *JStynalyser* is that several graphs have to be analysed in parallel and that the decision of how important changes of each style marker actually are, is subjective. Hierarchical cluster analysis, as well as PCA, evaluates all style markers together and visualizes them in one single graph. Although again the results are presented graphically and evaluated by a (subjective) user, the algorithms are objective and reproducible.

Hierarchical cluster analysis uses dendrograms for visualizing similarity (see section 5.4.1). Different dendrograms for the Mboob-Stallman text with all simple ratio measures, readability scores, frequency measures, relative entropy, and specific words are shown in figures 8.6, 8.7, and 8.8. Each figure shows a group of three dendrograms, in which *one* of the variables sliding window weight function, sliding window width and frequency list length is varied. The other two variables are not changed in this figure. The middle dendrogram of each figure is the same (linear decrease, sliding window width 11, frequency list length 20) and allows an easy comparison how the change of a variable affects the clustering. For better comprehensibility the sentence labels for Mboob and Stallman are written in different columns.

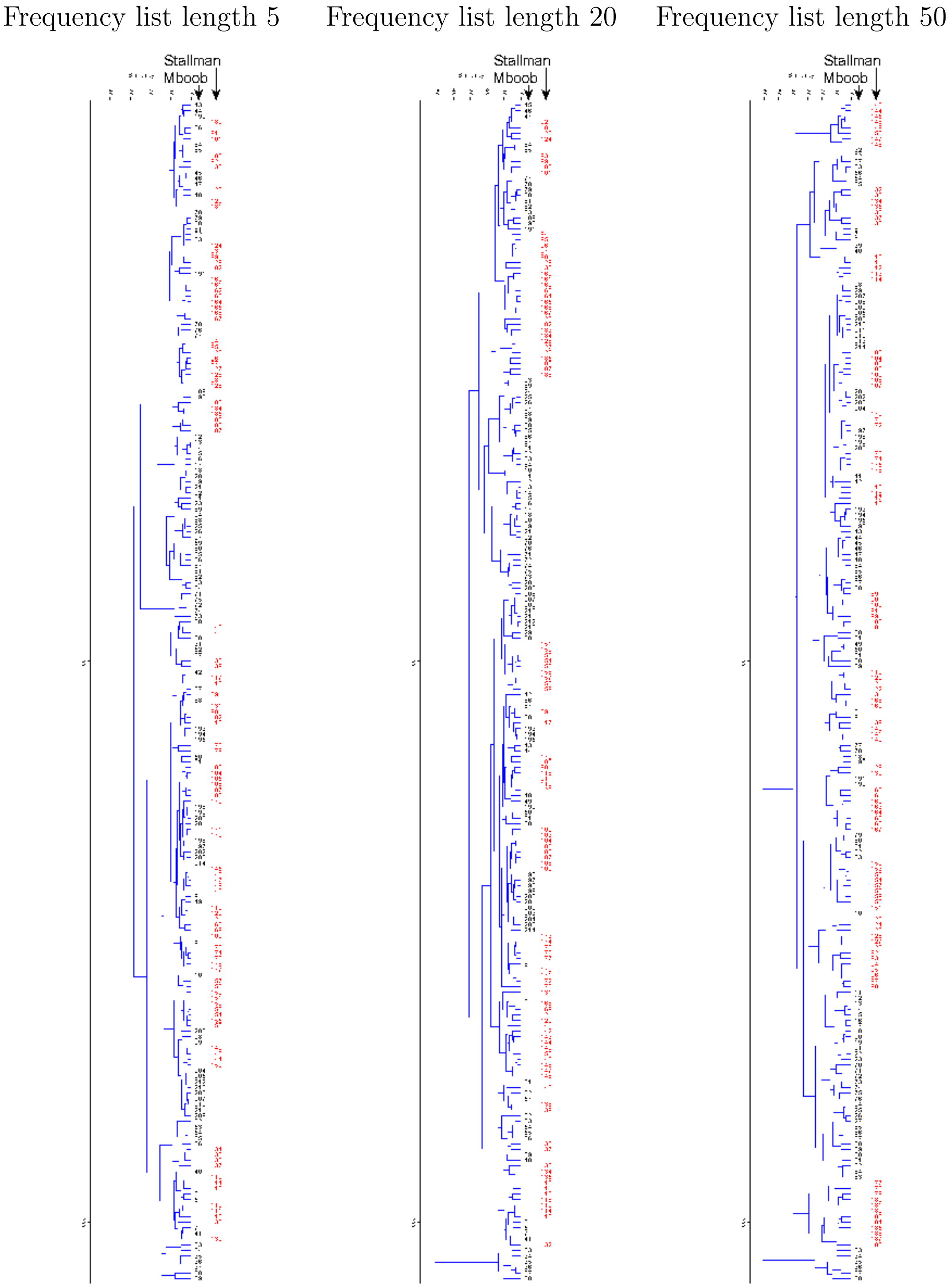
Figure 8.6 shows how the different sliding window weight functions affect the clustering. The linear and quadratic decrease functions show similar clustering results, as they have equally many cohesive clusters of comparable sizes. The constant value weight function performs less well; it produces smaller, discontinuous clusters.


Constant values

Linear decrease

Quadratic decrease

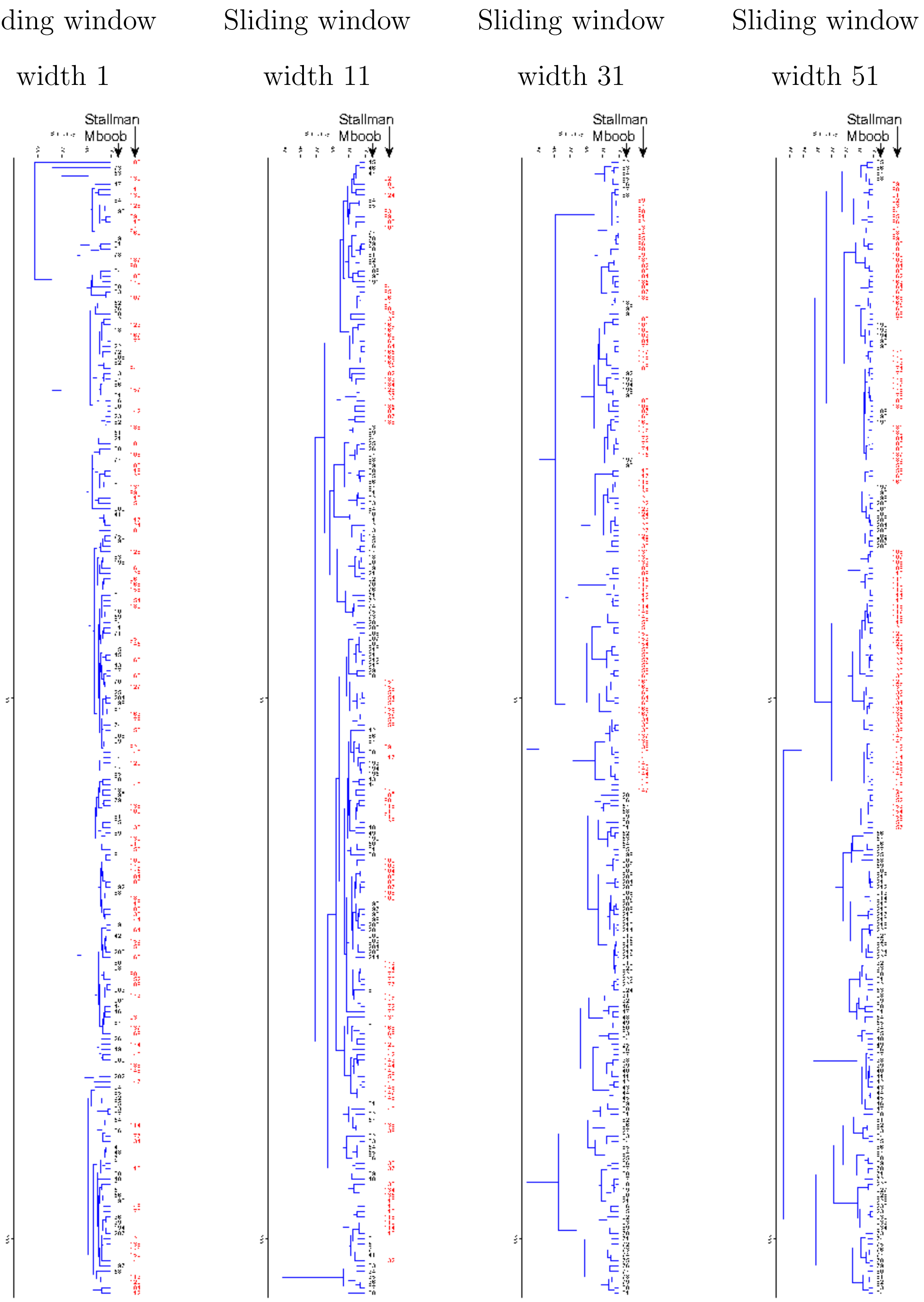
Figure 8.6: Dendrograms for the Mboob-Stallman text with different weight functions. Sliding window width is 11, frequency list length is 20 in all dendrograms. The paired groups algorithm with Euclidean distances is used.


Frequency list length 5

Frequency list length 20

Frequency list length 50

Figure 8.7: Dendrograms for the Mboob-Stallman text with different frequency list lengths. Sliding windows with width 11 and linear decrease are used in all dendrograms. The paired groups algorithm with Euclidean distances is used.



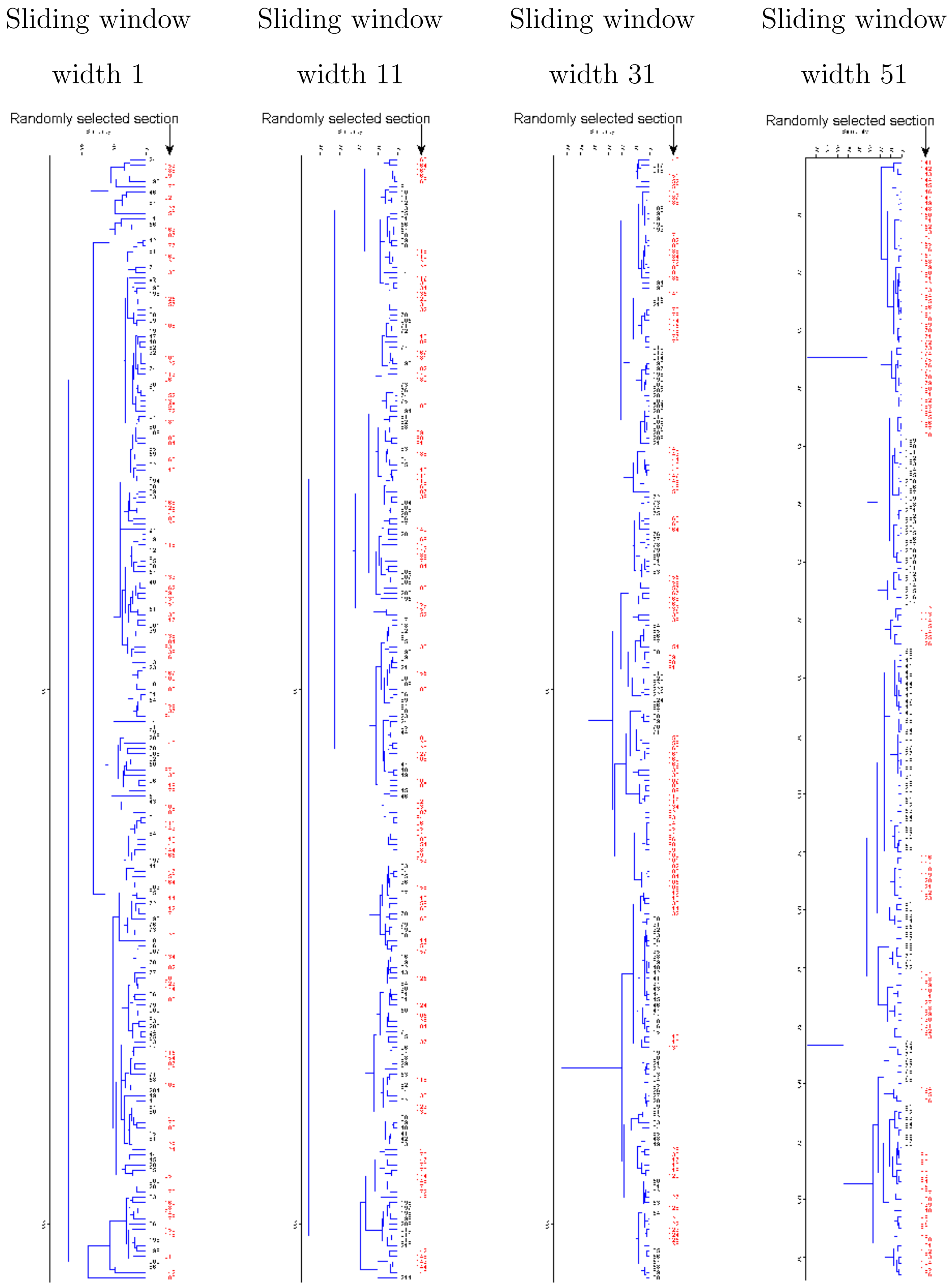
Sliding window width 1

Sliding window width 11

Sliding window width 31

Sliding window width 51

Figure 8.8: Dendrograms for the Mboob-Stallman text with different sliding window widths. The linear decrease weight function and a frequency list length of 20 are used in all dendrograms. The paired groups algorithm with Euclidean distances is used.



Sliding window width 1

Sliding window width 11

Sliding window width 31

Sliding window width 51

Figure 8.9: Dendrograms for the random bible text described in section 8.1.3, with different sliding window widths. The linear decrease sliding window weight function and a frequency list length of 20 are used in all dendrograms. The paired groups algorithm with Euclidean distances is used.

Increasing the frequency list length slightly favours the tendency to form author clusters. Figure 8.7 shows that longer frequency lists lead to slightly bigger and more continuous clusters. This trend is more distinct in some of the other artificially plagiarized texts, which are not visualized here. However, in none of the texts a clear discrimination into one cluster for each author is possible.

The enlargement of the sliding window strongly favours the formation of clusters (Figure 8.8). The left dendrogram in figure 8.8, in which the sentences are analysed independent of the surrounding sentences (sliding window width 1), shows no clustering of the two parts by Mboob and Stallman. The other dendrograms show that with growing sliding window width the sections of the two authors cluster better together, although again no perfect discrimination (Mboob in one subtree, Stallman in the other) is possible. This observation is backed by the results from the other plagiarized texts.

At first sight this improved clustering seems to be due to the large intervals under examination, which level out much of the noise of shorter intervals. Closer inspection of the two right graphs in figure 8.8, however, shows that especially those sentences by Mboob which are close to the Stallman part (sentences 85-88 and 189-208) are in the Stallman cluster. So possibly the clustering is affected by the sliding window approach.

To test whether the clustering described above comes from the stylistic similarity or from the sliding window method, the random bible text described in section 8.1.3 was analysed with hierarchical cluster analysis. The dendrograms visualizing the effect of changing the sliding window width are shown in figure 8.9. The dendrograms are very alike to the ones in figure 8.8: when sliding window width is increased, the clustering of the two parts of the text improves significantly. Because the text is assumed to be random, two adjacent sentences are, from a statistical point of view, as similar or dissimilar as two distant sentences. Consequently such a clustering cannot be due to the stylistic similarity of sentences.

The conclusion is that it must be the sliding window method that strongly affects the clustering and eclipses possible clusterings coming from stylistic similarities.

8.3.3 Principal components analysis (PCA)

Like hierarchical cluster analysis, PCA is a technique that visualizes all style markers in one graph, the *scatter plot*. The scatter plots for the Mboob-Stallman text are shown in figures 8.10 and 8.11. Again, simple ratio measures, readability scores, frequency measures, relative entropy, and specific words are used. In each figure, two of the variables sliding window weight function, sliding window width and frequency list length are changed, while the third variable is constant in that figure. Both the Mboob section (black dots) and the Stallman part (red crosses) are encircled by ellipses that enclose 95% of the data points.

The effect of changing the sliding window weight function is visualized in figure 8.10. The left column of the scatter plots shows that the constant weight function leads to noisier and less ordered results than the other weight functions in the middle and right. This is most clear for sliding window weights 31 and 51. The use of the linear and quadratic decrease function leads to similar clusterings, none of these two functions is clearly superior. These findings confirm the results of the visual inspection of the coordinate systems (section 8.3.1) and hierarchical cluster analysis (section 8.3.2).

Changing the length of frequency lists does not significantly affect the results (figure 8.11). The general pattern of each scatter plot does - apart from small variations in scaling - not change when altering the frequency list length. It should be noted that the scatter plot for sliding window width 1 and frequency list length 50 is largely affected by one outlier with a value of around 900 for component 1.

The variable that most strongly affects the results is sliding window width. This can be seen both in figure 8.10 and in figure 8.11. While for a sliding window width

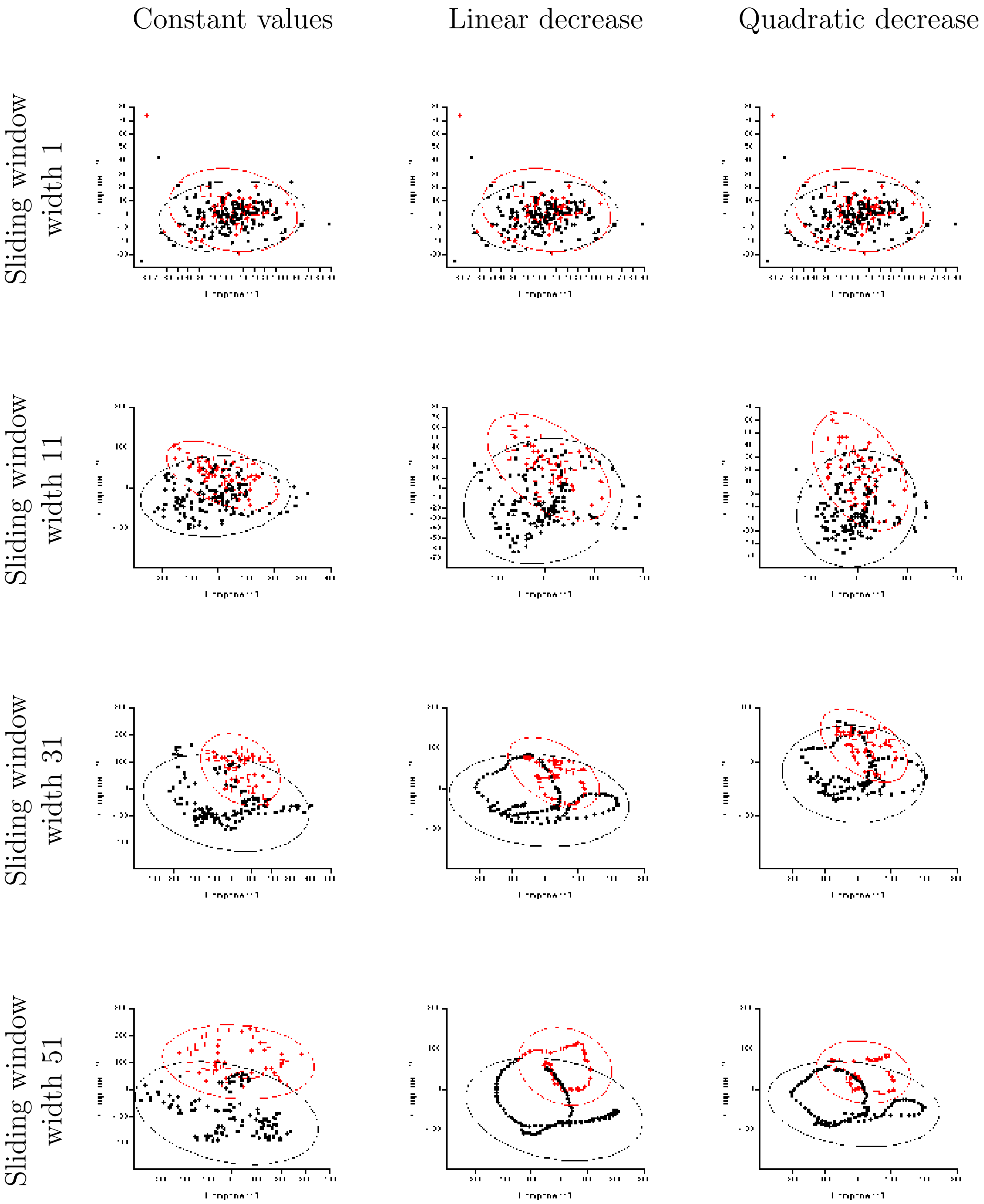


Figure 8.10: PCA scatter plots for the Mboob-Stallman text with frequency list length 20. The Shape PCA option of *PAST* is used. Sentences by Mboob are demarcated by black dots, sentences by Stallman by red crosses.

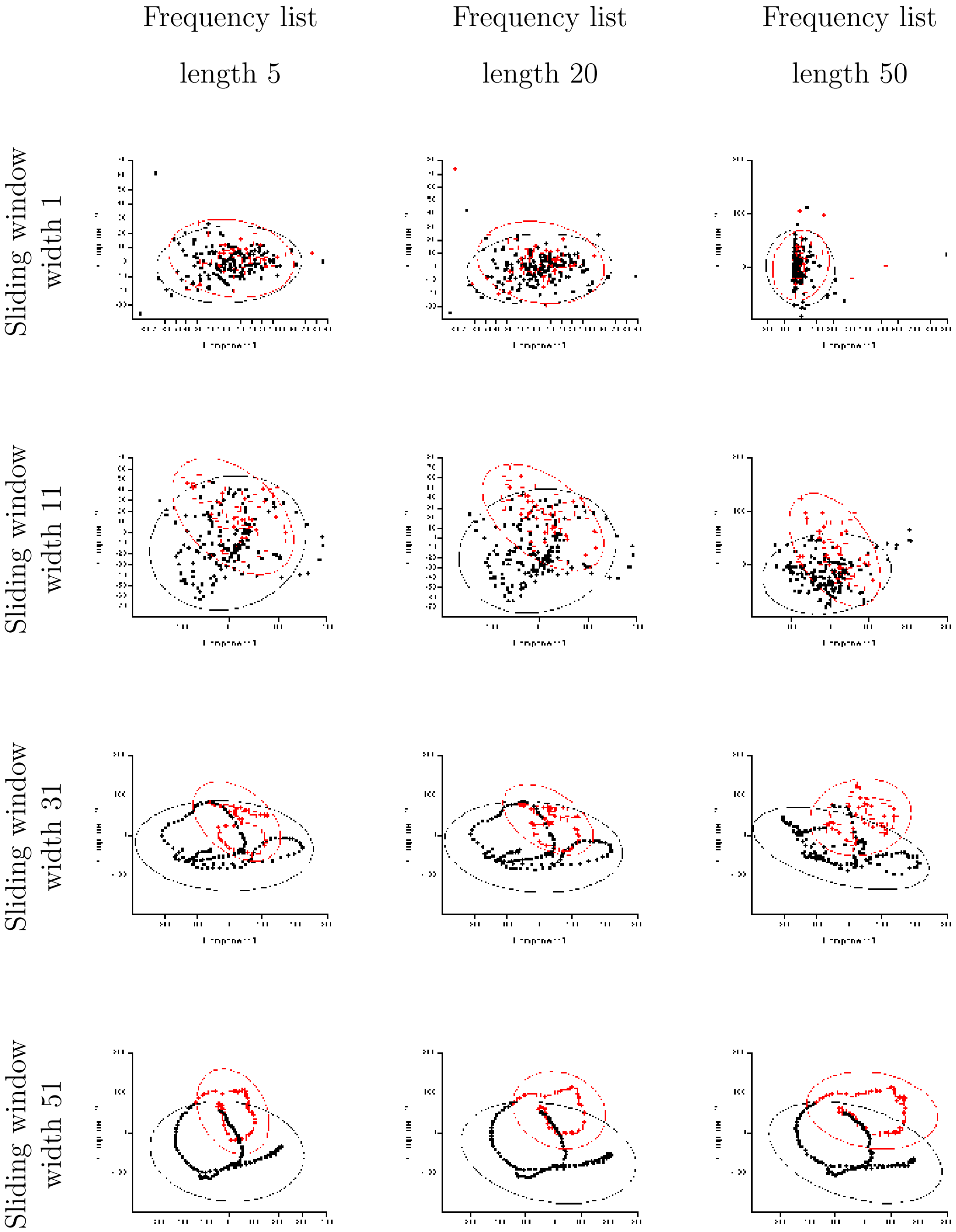
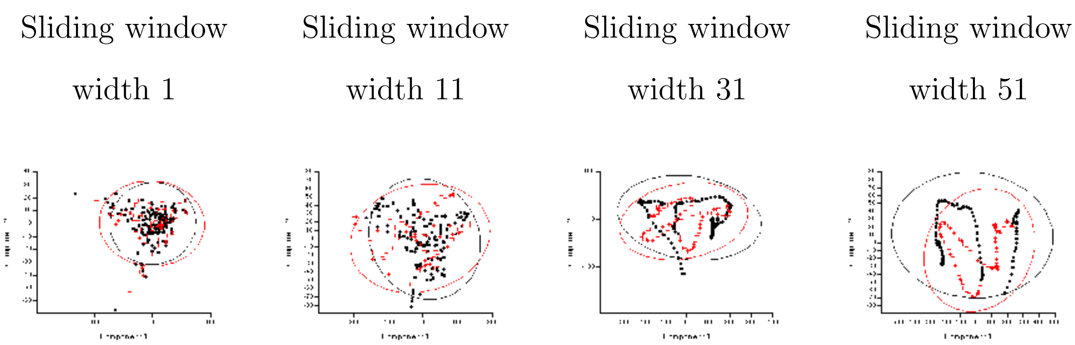


Figure 8.11: PCA scatter plots for the Mboob-Stallman text with linear decrease weight function. The Shape PCA option of *PAST* is used.

of 1 both the Mboob and the Stallman part strongly overlap and the points seem to be distributed randomly, a clear pattern emerges with increasing sliding window size. For sliding window width 51 the ellipses of the two authors still overlap, but to a lower degree as for smaller windows.

As in section 8.3.2 the question arises whether these changes come from the grouping of sentences or from the sliding window method. Therefore the random bible text from section 8.1.3 was analysed with PCA. As it was the case for hierarchical cluster analysis, the random text itself cannot lead to a pattern as in the Mboob-Stallman text, because all sentences are random. Provided that no other variable affects the results, the results will also be random.

Figure 8.12 shows the scatter plots for the random bible text (see section 8.1.3), in which 100 randomly selected adjacent sentences are marked. As in figure 8.10 and figure 8.11, a pattern arises for big sliding windows. Since the text itself is random, this pattern must be due to the effect of the sliding window analysis. As in section 8.3.2 it must be concluded that the sliding window approach strongly affects the results and eclipses possible clusterings coming from stylistic similarities of the sentences.



Sliding window width 1

Sliding window width 11

Sliding window width 31

Sliding window width 51

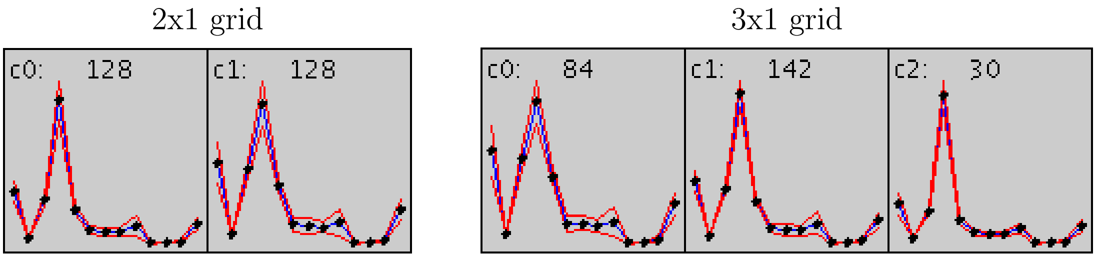
Figure 8.12: PCA scatter plots for the random bible text with linear decrease weight function and frequency list length 20. The Shape PCA option of *PAST* is used. The red crosses are randomly marked sentences.

8.3.4 Self-organizing maps (SOMs)

The analysis of the plagiarized texts with self-organizing maps (SOMs) did not lead to better results than hierarchical clustering or PCA. Therefore the results from the analysis with SOMs will be presented rather shortly.

Figure 8.13 shows two different SOM clusterings for the Mboob-Stallman text. The graphs of all clusters are very alike, with a clear maximum for Flesch-Kincaid formula (4th data point), a local minimum at the 2nd data point, and a rather long tail. This general form is preserved for all combinations of parameters (sliding window widths, weight functions and frequency list lengths) and all texts (other academic texts, random text).

All of the clusters contain sentences of both authors, i.e. SOM analysis cannot identify stylistic characteristics which discriminate the authors. Further analysis of the clusters in *GeneCluster* showed that with increased sliding window size adjacent sentences tend to cluster together. This effect has also been observed for hierarchical clustering and PCA (see sections 8.3.2 and 8.3.3).



Mboob: 72	Mboob: 84
Stallman: 56	Stallman: 44
Total: 128	Total: 128

Mboob: 60	Mboob: 78	Mboob: 18
Stallman: 24	Stallman: 64	Stallman: 12
Total: 84	Total: 142	Total: 30

Figure 8.13: SOMs for the Mboob-Stallman text with sliding window width 11, linear decrease and frequency list length 20. Each data point represents one style marker. The numbers below indicate how many sentences of each author the cluster contains.

8.4 Summary

In this chapter the main study of this work was outlined. Six texts with the topic “Censorship and Internet” or “Copyright and Internet” served as a basis for creating artificially plagiarized documents, which are analysed in this chapter. One text, called Mboob-Stallman text, was randomly chosen for thorough analysis. For reasons of comparison, a random text was created.

A section about the experimental setup explains how data was created and analysed; it should contain enough information for being able to replicate the experiments.

Section 8.3 presents the results from the application of style markers to plagiarized documents. It was shown that in most cases style markers cannot be used for plagiarism detection, because the inter-authorial stylistic changes are usually too small to stand out from the intra-authorial noise. If sections of one author cluster together, this is mainly due to the effects of the sliding window approach and not because sections of one author are very similar in style. The specific words measure, however, produces impressively good results at a 1-sentence level and pinpoints cited or plagiarized sentences.

The results of the main study will be more thoroughly discussed in the next chapter.

Chapter 9

Discussion

In chapter 7 a pre-study checked whether commonly used style markers fulfil basic prerequisites of a variable. In the main study (chapter 8), the style markers that showed promising results in the pre-study were applied to artificially plagiarized documents. This chapter will assess the results from both pre-study and main study and check whether the aim if style markers can detect stylistic changes at a sentence or paragraph level is achieved.

Style markers that are used to analyse texts at a sentence or paragraph level must fulfil two prerequisites: First, they must be able to discriminate authors, otherwise a distinction of authors due to stylistic differences is not possible. Second, they must be constant with sample size, otherwise sentences and paragraphs - which are not of constant length - cannot be compared to each other. The comparison of three 19th century novels by Jane Austen, Rudyard Kipling and Oscar Wilde in the pre-study showed that all of the four vocabulary richness measures violate the second prerequisite, and only two fulfil the first one (section 7.3, figures 7.2, 7.5 and 7.6). Therefore it is not possible to use vocabulary richness measures for detecting changes of style in a text. On the contrary, simple ratio measures, readability scores, frequency lists, and relative entropy fulfil these two prerequisites, and hence could be used in the main study.

The analysis of the graphs visualizing each individual style marker in a coordinate system shows that it is generally not possible to detect plagiarized sections (section 8.3.1, figures 8.2 and 8.3). At least at levels which are interesting for plagiarism detection, the stylistic differences between authors are too small to be recognizable when compared to the intra-authorial changes (i.e. noise). For large sliding windows some style markers seem to produce different values for different authors, but this tendency is weak, and it is not possible to draw conclusions from these little changes.

However, very promising results were observed for the specific words measure, a newly developed style marker which is optimized for detecting stylistic changes at a very small level. If single sentences by another author are inserted into a text, the experiments performed backed the hypothesis that the specific words measure can be used for detecting changes of authorship. In this case, the stylistic changes are indicated by outliers in the specific-words-graph, and hence easy to discover (figure 8.5). Precision values between 75% and 100% were observed for the most extreme outliers, i.e. some stylistically different sentences can be discovered with high probability. Recall values have to be assessed in future work, since an in-depth quantitative analysis was not performed. It should be noted that the sentences which were detected in the experiments are not plagiarized, but more or less correctly referenced. Still, the detected sentences were written by another author with another style, and the resulting stylistic change was detected by the specific words measure.

This detection of a change in authorship is only possible if one or only a few sentences are grouped together. If many sentences by another author are inserted into a text, the number of specific words disperses over the whole section, and the expected peak becomes a small bump, which is hardly distinguishable from noise. In that case, the hypothesis that the specific words measure can be used for detecting changes of authorship is falsified. On that account, the specific words measure behaves significantly different from the other style markers, which can only detect changes at large, but not at small levels.

Multivariate analyses, where all style markers are evaluated together, could not detect stylistic changes (see sections 8.3.2 to 8.3.4). On the one hand, the artificially plagiarized texts show a clear clustering of sections by the two authors if large sliding windows are used (see figures 8.8, 8.10 and 8.11). On the other hand, since a random text - which should also have random data points - shows a similar pattern (figures 8.9 and 8.12), the clustering cannot come from the detected stylistic change. It is rather due to the sliding window approach: The fact that all but two elements of a sliding window have the same (for constant weights) or at least similar (for linear and quadratic decrease) values affects the results strongly, and variations which might come from stylistic changes are obscured. Possibly other grouping methods like cumulative sums might circumvent these shortcomings.

The multivariate approaches evaluate all style markers together. That implicates that single style markers can affect the overall result largely. If these style markers have high noise ratios, the overall result will be noisy as well. So, possibly, other combinations of style markers as input variables for the multivariate analyses might have lead to improved, less noisy results.

Changing the sliding window parameters does not lead to improved results. On the one hand it is necessary to use a large sliding window so that the noise coming from intra-authorial variations in style is reduced, but this leads to the adulteration of results described above. Small sliding windows, which do not affect the results in that way, are also not able to fingerprint an author's style because the noise level is too high (figures 8.2, 8.8 and 8.10). Sliding window weight functions with linear or quadratic decrease can reduce the noise and lead to smoother graphs, but the basic problem of adulteration still remains (figures 8.2, 8.6 and 8.10).

If style markers are analysed individually, longer frequency lists seem to improve the distinguishability of authors (figure 8.3). The change of the list length, however, affects the results by far less significantly than the change of sliding window size, and not in all plagiarized texts the differences were as distinct as in the Mboob-Stallman text in figure 8.3. For multivariate analyses, the change of list length leads to slightly different clusterings, but the overall results neither improve nor deteriorate (see figures 8.7 and 8.11). If long word lists or letter trigram lists are analysed at a small level, the results might be inaccurate, since the expected frequencies of occurrences are too small to reliably use statistical tests (see section 8.3.1).

To put it in a nutshell: The results of this study suggest that the style markers used can generally *not* detect stylistic changes at a sentence or paragraph level, at least not with the unsupervised method used here. How surprising is this result?

Maybe it was foreseeable. Very little has been published in the field of internal plagiarism detection. Maybe researchers have an intuitive feeling that stylometry does not work at such small levels, and maybe that is the reason why they did not invest time in that topic. The only known related paper (Hersee 2001) is “only” a BSc thesis, and it fails to detect plagiarism.

On the other hand, stylistic changes *are* detectable by internal approaches. Lecturers can identify changes in an author’s style when reading a text. The specific words measure pinpoints sentences which were written by another author. Therefore there must be a detectable stylistic change in the text. The problem is that most style markers are not able to detect them. This maybe due to the style markers, which simply are not optimized for small analysis levels. Possibly the methods used to detect changes are not good enough, and approaches mimicking human pattern detection (like neural networks) or supervised methods would lead to better results. Maybe the stylometric assumptions do not hold for the academic texts used in this study, and other data sets would have lead to other conclusions. Other issues concern the quality of the preprocessed data (e.g. incorrect sentence splits) or the grouping method (influence of the sliding window method on results).

In short, other style markers, other methods, and other data sets might have led to different, possibly better results. See section 10.2 for ideas how the results could be improved in future work.

Chapter 10

Conclusion

The aim of this thesis was to apply style markers from the field of authorship attribution to a single document and to investigate, if these style markers are also applicable at a sentence and paragraph level to detect stylistic changes. This work has shown that some style markers are not applicable because they do not fulfil necessary prerequisites of a metric (e.g. vocabulary richness measures). Other style markers are applicable but cannot detect stylistic changes at a sentence or paragraph level (e.g. simple ratio measures, readability scores, frequency measures, and relative entropy). The experiments performed showed that the newly developed style marker specific words *can* pinpoint a change of authorship, but is limited to small analysis levels; if many sentences of different authorship are inserted into a text, the specific words measure cannot dectect a change of authorship.

Therefore, based on the evidence presented in this thesis, the hypothesis that lexical style markers are also usable with an unsupervised approach at a sentence and paragraph level to detect stylistic changes which might pinpoint plagiarism must be rejected.

10.1 Contributions

Simple ratio measures, readability scores, frequency lists, and entropy measures have been frequently used at a text level (see chapter 3 for examples). In this thesis, these style markers were - as far as we know - for the first time applied at a much smaller sentence and paragraph level.

In order to find out if the style markers are applicable at a sentence and paragraph level, a pre-study checked whether the style markers are constant with sample size, and whether they have discriminatory power for different authors ("fingerprintability"). Studies like this have been performed for some style markers, e.g. for vocabulary richness measures (Tweedie and Baayen 1998). The pre-study of this thesis, however, is the first known study which provides a concise analysis of sample size independence and fingerprintability for simple ratio measures, readability scores, vocabulary richness measures, frequency lists, and relative entropy.

The computer programs used in other work about stylometry were not applicable for intra-document analysis because they were either not available[1] or were not usable because they do not support intra-document analysis. Therefore a set of Perl scripts has been programmed. These scripts combine existing modules from the Perl CPAN archive with newly written modules (in case no implementation existed) and can stylistically analyse a text. For easy handling and visualization, a GUI-based Java program called *JStynalyser* was also implemented. Both the Perl scripts and the Java tool are freely available[2] and can be used or improved for future work on stylometry.

[1] Some of the programs are expensive commercial tools, or were not found via the Internet because they are too old and not longer maintained. Most authors, however, have not referenced which programs they have used.

[2] http://stynalyser.sourceforge.net

10.2 Future work

Although this work has evaluated six types of style markers under different conditions, the analysis was not exhaustive. A whole set of experimental variations, which might improve the results, can be identified for future work.

Style markers have successfully been used at a level of several thousand words to attribute texts of unknown authorship (see chapter 3 for examples). This work shows that these style markers are not usable at small levels of a few dozen words. It would be interesting to find a threshold down to which vocabulary richness measures, readability scores, frequency lists, and relative entropy can detect stylistic changes, and what this threshold depends on.

The specific words measure showed promising results at a sentence level. What needs to be done is an in-depth quantitative analysis with more data to back the findings, and to provide a threshold which can discriminate plagiarized from original sentences. Once such a threshold is found, the specific words measure can for example be used in a tool which checks a text for correct referencing.

The experiments showed that the analysis results are affected by incorrect preprocessing, for example erroneous sentence splitting (see especially the effects of the Wilde outlier in figure 7.2, section 7.3.1). The sentence and word tokenization algorithms both use simple regular expressions, and do not take into account the context of the sentence or word to be split. Certain subtleties of tokenization (see Gibaldi 1999, pp. 50-65 and Palmer 2000, pp. 18-22 for some examples) are not recognized with these expressions; the result is an incorrect split. Trainable or adaptive approaches might circumvent these problems and lead to superior tokenization results.

All style markers which have been used work on a lexical level. Part-of-speech tagging and syntactic annotation have shown good results for authorship attribution (Baayen 1996; Stamatatos, Fakotakis and Kokkinakis 2001) and might also be usable for intra-document analysis.

Stylistic changes could not be detected in this work. On the other hand, human readers, for example lecturers, can detect changes in style when reading a student's paper. Possibly more "human-like" pattern detection systems, like neural networks, will lead to improved results.

Bibliography

Austen, J. (1813). Pride and Prejudice [online]. Available from: http://www.ibiblio.org/gutenberg/etext98/pandp12.txt [Accessed 2003-08-06].

Baayen, R. H. (1996). Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing 11*(3), 121–132.

Binongo, J. N. and M. W. A. Smith (1999). The Application of Principal Component Analysis to Stylometry. *Literary and Linguistic Computing 14*(4), 445–466.

Brunet, E. (1978). *Le vocabulaire de Jean Giraudoux: structure et évolution: Statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la Langue Française.* Genève: Éditions Slatkine.

Bruno, A. M. (1974). *Toward a Quantitative Methodology for Stylistic Analyses*, Volume 109 of *University of California Publications in Modern Philology.* Berkeley: University of California Press.

Bull, J., C. Collins, E. Coughlin, and D. Sharp (2001). Technical Review of Plagiarism Detection Software Report. Technical report, University of Luton. Available from: http://www.jisc.ac.uk/uploaded_documents/luton.pdf [Accessed 2003-08-06].

Burrows, J. F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method.* Oxford: Clarendon Press.

Burrows, J. F. (1992). Not Unless You Ask Nicely: The Interpretative Nexus between Analysis and Information. *Literary and Linguistic Computing 7*(2), 91–109.

Carroll, J. and J. Appleton (2001). Plagiarism - A Good Practice Guide. Technical report, Oxford Brookes University. Available from: `http://www.jisc.ac.uk/uploaded_documents/brookes.pdf` [Accessed 2003-08-06].

Carroll, R. T. (2001). Cryptomnesia. The Skeptics Dictionary [online]. Available from: `http://www.skepdic.com/cryptomn.html` [Accessed 2003-08-06].

Chaski, C. E. (1997). Who Wrote It? Steps Toward a Science of Authorship Identification. *National Institute of Justice Journal 233*, 15–22. Available from: `http://ncjrs.org/pdffiles/jr000233.pdf` [Accessed 2003-08-06].

Clough, P. (2000a). Analysing style - Readability. Technical report, University of Sheffield. Available from: `http://ir.shef.ac.uk/cloughie/papers/readability.pdf` [Accessed 2003-08-06].

Clough, P. (2000b). Plagiarism in natural and programming languages: an overview of current tools and technologies. Technical report, University of Sheffield. Available from: `http://ir.shef.ac.uk/cloughie/papers/plagiarism2000.pdf` [Accessed 2003-08-06].

Clough, P. (2001). A Perl program for sentence splitting using rules. Internal report, University of Sheffield. Available from: `http://www.dcs.shef.ac.uk/~cloughie/papers/sentences.ps` [Accessed 2003-08-06].

Clough, P. (2003). *Measuring Text Reuse*. PhD thesis, University of Sheffield. Available from: `http://ir.shef.ac.uk/cloughie/papers/thesis.pdf` [Accessed 2003-08-06].

Corns, T. N. (1990). *Milton's Language*. Oxford: Basil Blackwell.

Culwin, F. and T. Lancaster (2000). A Review of Electronic Services for Plagiarism Detection in Student Submissions. In *Proceedings of the 1st LTSN-ICS*

Conference, Edinburgh, 54–61. Available from: `http://www.ics.ltsn.ac.uk/pub/conf2000/Papers/Culwin.pdf` [Accessed 2003-08-06].

de Haan, P. (1998). Review of Farringdon: Analysing Authorship: A Guide to the Cusum Technique. *Forensic Linguistics 5*, 69–76. Available from: `http://www.let.kun.nl/p.dehaan/reviews/CUSUM_book_review.php` [Accessed 2003-08-06].

Evans, J. (2000). The New Plagiarism in Higher Education: From Selection to Reflection [online]. Available from: `http://www.warwick.ac.uk/ETS/interactions/vol4no2/evans.htm` [Accessed 2003-08-06].

Farringdon, J. M. (1996). *Analysing for Authorship: A Guide to the Cusum Technique*. Cardiff: University of Wales Press.

Frayn, C. (2002). Col's Random Sentence Generator [online]. Available from: `http://www.ast.cam.ac.uk/~cmf/generate` [Accessed 2003-08-06].

Free Student Essays (Ed.) (1995). Government Censorship on the Internet [online]. Available from: `http://freestudentessays.com/get_essays/computer/20_1_12.shtml` [Accessed 2003-08-06].

Fucks, W. (1952). On Mathematical Analysis of Style. *Biometrika 39*(1/2), 122–129.

Gibaldi, J. (1999). *MLA Handbook for Writers of Research Papers* (5 ed.). New York: The Modern Language Association of America.

Glover, A. D. (1996). Automatically detecting stylistic inconsistencies in computer-supported collaborative writing. Masters Thesis, University of Toronto. Available from: `ftp://ftp.cs.toronto.edu/pub/gh/Glover-thesis.pdf` [Accessed 2003-08-06].

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. M. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and

E. S. Lander (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Hammer, Ø., D. A. Harper, and P. D. Ryan (2001). PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontologia Electronica 4*(1), 9pp. Available from: `http://palaeo-electronica.org/2001_1/past/issue1_01.htm` [Accessed 2003-08-06].

Hersee, M. S. (2001). *Automatic Detection of Plagiarism: An Approach Using the Qsum Method*. BSc Computer Science, University of Sheffield. Available from: `http://www.dcs.shef.ac.uk/teaching/eproj/ug2001/pdf/u8msh.pdf` [Accessed 2003-08-06].

Hoad, T. and J. Zobel (2002). Methods for Identifying Versioned and Plagiarised Documents. Technical report, RMIT University, Melbourne. Available from: `http://www.seg.rmit.edu.au/research/download.php?manuscript=52` [Accessed 2003-08-06].

Hockey, S. (2000). *Electronic Texts in the Humanities: Principles and Practice*. Oxford: Oxford University Press.

Holmes, D. I. (1992). A Stylometric Analysis of Mormon Scripture and Related Texts. *Journal of the Royal Statistical Society Series A 155*(1), 91–120.

Holmes, D. I. (1994). Authorship Attribution. *Computers and the Humanities 28*(2), 87–106.

Holmes, D. I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing 13*(3), 111–117.

Holmes, D. I. and R. S. Forsyth (1995). The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing 10*(2), 111–127.

Honoré, T. (1979). Some simple measures of richness of vocabulary. *Association for literary and linguistic computing bulletin 7*(2), 172–177.

Hoover, D. L. (2001). Statistical Stylistics and Authorship Attribution: an Empirical Investigation. *Literary and Linguistic Computing 16*(4), 421–444.

Hoover, D. L. (2002). Frequent Word Sequences and Statistical Stylistics. *Literary and Linguistic Computing 17*(2), 157–180.

Johnson, K. (1998). Readability [online]. Available from: `http://www.timetabler.com/readable.pdf` [Accessed 2003-08-06].

Khmelev, D. V. and F. J. Tweedie (2001). Using Markov Chains for Identification of Writers. *Literary and Linguistic Computing 16*(3), 299–307.

Kilgarriff, A. (1996). Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora. Technical report, Information Technology Research Institute, University of Brighton. Available from: `http://acl.ldc.upenn.edu/W/W97/W97-0122.pdf` [Accessed 2003-08-06].

Kipling, R. (1894). The Jungle Book [online]. Available from: `http://www.ibiblio.org/gutenberg/etext95/jnglb10.txt` [Accessed 2003-08-06].

Kirkpatrick, P. (1991). *Reference guide to English literature: Works. Title Index* (2 ed.), Volume 2. Chicago: St. James Press.

Kjell, B. (1994). Authorship Determination Using Letter Pair Frequency Features with Neural Network Classifiers. *Literary and Linguistic Computing 9*(2), 119–124.

Kohonen, T. (2000). The Self-Organizing Map. *Proceedings of the IEEE 78*, 1464–1480.

Kukushkina, O. V., A. A. Polikarpov, and D. V. Khmelev (2000). Using Literal and Grammatical Statistics for Authorship Attribution. *Problems of Information Transmission 37*(2), 172–184. Available from: `http://www.ma.hw.ac.uk/~dmitri/PAPERS/published/gramcodes/gramcodeseng.pdf` [Accessed 2003-08-06].

Landier, M. (1997). Internet Censorship is Absurd and Unconstitutional [online]. Available from: http://www.landier.com/michael/essays/censorship/fulltext.htm [Accessed 2003-08-06].

Lesko, J. P. (1996). Plagiarism and questionable appropriation of text by non-native speaker students in taught postgraduate courses: views and experiences of postgraduate staff. In *Proceedings of the Edinburgh Linguistics Department Conference '96*, 142–149. Available from: http://www.ling.ed.ac.uk/~pgc/archive/lesko.ps [Accessed 2003-08-06].

Lindey, A. (1952). *Plagiarism and Originality.* New York: Harper.

Lubovac, Z. (2000). Evaluation of clusterings of gene expression data. Master Thesis, University College of Skövde. Available from: http://www.ida.his.se/ida/htbin/exjobb/2000/HS-IDA-MD-00-007 [Accessed 2003-08-06].

Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press.

Martin, B. (1994). Plagiarism: a misplaced emphasis. *Journal of Information Ethics 3*(2), 36–47. Available from: http://www.uow.edu.au/arts/sts/bmartin/pubs/94jie.html [Accessed 2003-08-06].

Matthews, R. A. and T. V. Merriam (1993). Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher. *Literary and Linguistic Computing 8*(4), 203–209.

Mboob, B. (2001). Censorship and the Internet [online]. Available from: http://www.oppapers.com/read.php?id=30214&idenc=KxyHiuJa [Accessed 2003-08-06].

McCombe, N. (2002). *Methods of Author Identification.* B.A. (Mod.) CSLL, Trinity College Dublin. Available from: http://www.cs.tcd.ie/courses/csll/mccombe0102.pdf [Accessed 2003-08-06].

McEnery, T. and M. Oakes (2000). Authorship Identification and Computational Stylometry. In R. Dale, H. Moisl, and H. Somers (Eds.), *Handbook of Natural Language Processing*, Chapter Authorship Identification and Computational Stylometry, 545–562. New York: Marcel Dekker.

Merriam, T. V. and R. A. Matthews (1994). Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe. *Literary and Linguistic Computing 9*(1), 1–6.

Morton, A. Q. (1978). *Literary Detection: How to prove authorship and fraud in literature and documents*. Epping: Bowker Publishing Company.

Mosteller, F. and D. Wallace (1964). *Inference and Disputed Authorship: The Federalist*. Reading: Addison-Wesley.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh University Press.

Palmer, D. D. (2000). Tokenisation and Sentence Segmentation. In R. Dale, H. Moisl, and H. Somers (Eds.), *Handbook of Natural Language Processing*, 11–35. New York: Marcel Dekker.

Pierce, J. R. (1980). *An Introduction to Information Theory - Symbols, Signal and Noise* (2 ed.). New York: Dover Publications.

Project Gutenberg (Ed.) (1989). The King James Bible. The Old and New Testaments King James Version [online]. Available from: `http://www.ibiblio.org/gutenberg/etext90/kjv10.txt` [Accessed 2003-08-06].

Radler, R. and W. Jens (1988). *Kindlers neues Literaturlexikon*. München: Kindler.

Rudman, J. (1998). The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities 31*(4), 351–365.

Schmidt, P. (2001). Regulation of Internet Content [online]. Available from:

http://www.hausarbeiten.de/rd/faecher/hausarbeit/inh/13107.html [Accessed 2002-08-06].

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures* (2 ed.). Boca Raton: Chapman & Hall/CRC.

Shivakumar, N. and H. Garcia-Molina (1995). The SCAM Approach to Copy Detection in Digital Libraries. Available from: http://citeseer.nj.nec.com/64547.html [Accessed 2003-08-06].

Si, A., H. V. Leong, and R. W. H. Lau (1997). CHECK: A Document Plagiarism Detection System. In *Proceedings of ACM Symposium for Applied Computing*, 70–77. Available from: http://citeseer.nj.nec.com/si97check.html [Accessed 2003-08-06].

Sichel, H. S. (1975). On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association 70*(351), 542–547.

Stallman, R. (1997). The Right To Read. *Communications of the ACM 40*(2), 85–87. Available from: http://www.gnu.org/philosophy/right-to-read.html [Accessed 2002-08-06].

Stallman, R. (2002). Misinterpreting Copyright. In J. Gay (Ed.), *Free Software, Free Society: The Selected Essays of Richard M. Stallman*. Boston: GNU Press. Available from: http://www.gnu.org/philosophy/misinterpreting-copyright.html [Accessed 2003-08-06].

Stamatatos, E., N. Fakotakis, and G. Kokkinakis (1999). Automatic Authorship Attribution. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL '99)*, Bergen, 158–164. Available from: http://slt.wcl.ee.upatras.gr/papers/stamatatos1.pdf [Accessed 2003-08-06].

Stamatatos, E., N. Fakotakis, and G. Kokkinakis (2000). Text Genre Detection Using Common Word Frequencies. In *Proceedings of the 18th Int. Conference* *on Computational Linguistics (COLING2000)*, Saarbrücken. Available from: `http://slt.wcl.ee.upatras.gr/papers/stamatatos2.pdf` [Accessed 2003-08-06].

Stamatatos, E., N. Fakotakis, and G. Kokkinakis (2001). Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities 35*(2), 193–214.

Stephens, C. (2000). All About Readability [online]. Available from: `http://www.plainlanguagenetwork.org/stephens/readability.html` [Accessed 2003-08-06].

Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub (1999). Interpreting patterns of gene expression with self-organising maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the USA 96*, 2907–2912.

Tirvengadum, V. (1996). Two Methods of Author Identification: the Gary/Ajar Case. Technical report, University of Manitoba. Available from: `http://www.hit.uib.no/allc/tirvenga.pdf` [Accessed 2002-08-06].

Tweedie, F. J. and R. H. Baayen (1998). How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities 32*(5), 323–352.

Verco, K. L. and M. J. Wise (1996). Software for Detecting Suspected Plagiarism: Comparing Structure and Attribute-Counting Systems. In J. Rosenberg (Ed.), *First Australian Conference on Computer Science Education*, Sydney. Available from: `ftp://ftp.cs.su.oz.au/michaelw/yap_vs_acm.ps` [Accessed 2003-08-06].

Wilde, O. (1891). The Picture of Dorian Gray [online]. Available from: `http://www.ibiblio.org/gutenberg/etext03/8dgry10.txt` [Accessed 2003-08-06].

Williams, C. B. (1940). A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style. *Biometrika 31*(3/4), 356–361.

Williamson, K. (2002). *Research methods for students, academics and professionals: Information management and systems* (2nd ed.). Number 22 in Topics in Australasian library and information studies. Wagga Wagga: Center for Information Studies.

Woods, A., P. Fletcher, and A. Hughes (1986). *Statistics in language studies.* Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society 104*, 444–466.

Yule, G. U. (1939). On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to two Cases of Disputed Authorship. *Biometrika 30*(3/4), 363–390.

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary.* Cambridge: Cambridge University Press.