jcls. Journal of Computational Literary Studies

OPEN ACCESS

Article

Why the Daisy Sisters are Different

A Stylometric Study on the Oeuvre of Swedish Author Henning Mankell and the Dutch Translations of His Work

Martje Wijers[1]

1. Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, The Netherlands.

Citation
Martje Wijers (2023). "Why the Daisy Sisters are Different. A Stylometric Study on the Oeuvre of Swedish Author Henning Mankell and the Dutch Translations of His Work". In: *Journal of Computational Literary Studies* 2 (1). 10.48694/jcls.3585

Date published 2024-02-25
Date accepted 2024-01-21
Date received 2023-01-04

Keywords
stylometry, Cluster analysis, PCA, delta, zeta, Mankell, translation

Reviewers
Jan Rybicki, Mats Malm

Note
This paper has passed through the conference track of JCLS. In addition to being peer reviewed, it was presented and discussed at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. In this paper, 32 books by the Swedish writer Henning Mankell were investigated using stylometric methods, to find out whether his style varies in different genres, if his style changed measurably over time, or if his books differ from each other stylistically for other reasons. The results show that the time of publication can play a role, but that other factors, such as dominant verb tense used and narrative perspective, as well as register, are more important in determining whether and how the style of novels differs. This study also gives more insight into frequently used methods in stylometry, such as cluster analysis and PCA, that give little information about the stylistic features that differ between texts. For this purpose, the original Swedish texts were also compared to the Dutch translations of the same texts to determine how translation and language influence the results of stylometric analyses.

1. Introduction

In a conversation at the university of Tulsa in 2011, Swedish author Henning Mankell told his colleague Michael Ondaatje:

> I'm like the farmer, who knows, that the land shouldn't be used for the same crops many years in a row. I try to cultivate the land in my head in the same way... [...] That's why I switch between styles and between novels, essays and theater. One of the decisive things for me is, when I have an idea for a story, to decide what kind of story it is. Is it a theater play? Is it a film script? A novel? A crime novel? (Jacobsen 2012, 31)[1]

Although Henning Mankell is most known for his detective series Wallander, he indeed wrote a variety of genres during his 42 years long career as a writer. He wrote literary novels, crime novels, non-fiction, theatre plays, film scripts and children's literature.

In this paper the whole oeuvre of Mankell is scrutinized using stylometric analyses to see if his style changed measurably over time, or if some books deviate stylistically from his other works for other reasons. In this study, style is used in the definition by Herrmann et al. (2015, 44): "Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively." In stylometry,

[1] My translation.

quantitative features are investigated, such as word frequency, ngrams, clause or sentence length, word classes or punctuation marks (Herrmann et al. 2015). In the current study style is measured by word frequency patterns.

The original works by Mankell are also compared to their Dutch translations. The goal of this comparison is to investigate to what extent language and translation in general can influence the results of stylometric analyses. Apart from more insight into the styles in Mankell's oeuvre, this study will yield interesting observations about the selected methods, and it can give new insights about frequently used methods in computational literary studies, such as cluster analyses and principal components analyses. These methods are generally based on the Most Frequent Words (MFW) of a text, although little is known about which type of words are decisive and what factors should be taken into account in this type of analysis.

The current paper is inspired by the computational research project 'The Riddle of Literary Quality' (2012-2019). In this project, Karina van Dalen-Oskam and her colleagues at the Huygens Institute for the History of the Netherlands in collaboration with the Fryske Akademy and the Institute for Logic, Language and Computation at the University of Amsterdam investigated readers' perceptions of what (good) literature is and to what extend these perceptions can be linked to formal patterns in novels (Van Dalen-Oskam 2021, 15–16).[2] Five of the many novels that Van Dalen-Oskam investigated were written by Mankell and particularly one of them, the literary novel *Daisy Sisters*, stood out in several ways compared to the other novels written by translated male authors in the project. She looked into the novels by Mankell in more detail and the main conclusion was as follows:

> It seems, therefore, that although Mankell published books in two different genres, Suspense and Literary novel, his style as reflected in his use of words, perhaps is not very different. The fact that *De Daisy Sisters* was an outlier does not disprove this because the original is much older (1982) and it is known that an author's writing style may develop over time just like languages and the conventions that apply to different genres [...]. Further research into Mankell's complete oeuvre would be needed to confirm this. (cited after Van Dalen-Oskam 2023, 76)

So, Mankells style did not differ very much between genres when looking at word use compared in a corpus including books by other writers, but the *Daisy Sisters* deviated clearly from the other books, possibly because it was written much earlier. By looking at the broader oeuvre of Mankell, some of the questions that remained after the 'Riddle of Literary Quality' was finished can be answered.

The corpus compiled for this study consists of 32 Swedish books written by Mankell in four genres: crime-fiction (N=15), literary novels (N=11), children's books (N=4) and non-fiction (N=2). For comparison purposes, ten books by the following best-selling Swedish writers were added to the corpus: Johannes Anyuru, Majgull Axelsson, Marianne Fredriksson, Lars Kepler, John Ajvide Lindqvist, Camilla Läckberg, and Håkan Nesser. Six of the books by other Swedish writers are literary novels and four are

[2] An updated English version of this book has been published in English in June 2023 under the title *The Riddle of Literary Quality: A Computational Approach* (Van Dalen-Oskam 2023).

crime novels. The translation corpus contains 42 translations of all the above-mentioned works by Henning Mankell and other Swedish writers into Dutch.

2. Multi-Faceted Henning Mankell

Arvas and Nestingen (2011, 1) state that Mankell is the top selling Swedish crime-fiction author who, according to them "has sold 25 million copies, even outperforming Harry Potter in the German-language market." Mankell was certainly one of Sweden's most popular and well-read crime-fiction writers, although Berglund (2012, 10) puts these numbers in perspective. He shows that Henning Mankell was not necessarily the number one best-selling author in Sweden in the period 2004-2010, but that he indeed was among the top-sellers. He was in fact in fourth position after Camilla Läckberg, Stieg Larsson and Liza Marklund on the top 40 best-selling crime-fiction authors in Sweden (Berglund 2012, 81). Interestingly, compared to the even more popular authors Stieg Larsson, Camilla Läckberg and Liza Marklund, the books by Henning Mankell were borrowed much more frequently at libraries (Berglund 2012, 100–101).

Henning Mankell is an interesting author to investigate for multiple reasons. He modernized the already existing Swedish police novel that included criticism on modern society (started by Maj Sjöwall & Per Wahlöö) and he was the first Swedish author of crime novels to be published in many languages abroad with wide circulation. Therefore, Mankell played an important role in the rise of the Nordic noir genre (Berglund 2012, 114).

Mankell belongs to a group of authors that were already established writers of fiction before they started to write crime (in the late 90s) when there was a boom in crime-fiction in Sweden. His debut in the crime genre was in 1991 with *Mördare utan ansikte* (*Faceless Killers*), but his debut as a writer of fiction was much earlier: in 1973 with *Bergsprängaren* (*The Rock Blaster*).

The fact that he has a broad oeuvre covering four genres over a time span of 42 years (1973-2015) also makes Henning Mankell useful for a computational study. Furthermore, his novels are widely translated into other languages. Almost his entire oeuvre is translated into Dutch. There is a limited number of Dutch translators from Swedish which makes it possible to compare translations by different translators. As mentioned earlier, these translations were sometimes published much later in Dutch than the original. It is important to bear in mind that a writer's style can change over time, and so do ideas about translation (Can and Patton 2004; Hoover 2020; Ríos-Toledo et al. 2022).

2.1 Mankell and the Riddle of Literary Quality

In the 'Riddle of Literary Quality', Van Dalen-Oskam and her colleagues investigated if literary quality is measurable using stylometric methods. They selected 401 contemporary novels in Dutch published between 2007-2012 based on sales numbers and library borrowings in the three years prior to the survey, i.e. in 2009-2012 (Van Dalen-Oskam 2021, 44). The works included both novels originally written in Dutch as well as translated novels. These novels were rated for their literary quality on a scale from one (not literary at all) to seven (very literary) by almost 14,000 readers in 'Het Nationale Lezersonderzoek' (The National Reader Survey) in 2013 (Van Dalen-Oskam 2021, 40–43). The ratings were then linked to the formal aspects of the books, such as vocabulary and sentence length or contextual information, such as whether the author is male or female (Van Dalen-Oskam 2021).

One of the findings in 'The Riddle of Literary Quality' was that many readers seemed to be somewhat more critical towards translated literary fiction compared to literary novels originally written in Dutch. In other genres, such as crime novels, the bias was just the opposite: On average, translated works received higher scores on literary quality than original Dutch crime novels (Van Dalen-Oskam 2021, 104–105).

However, there was a clear difference between books translated from English and books translated from other languages. Translated books from other languages than English scored higher on literary quality than works translated from English and books originally written in Dutch in the category literary novels as well as the category crime novels (Van Dalen-Oskam 2021, 105). Van Dalen-Oskam suggests that readers are more critical toward translations from languages they know than from languages they are unfamiliar or much less familiar with (Van Dalen-Oskam 2021, 112).

In total there were 249 translated books in the survey. English was by far the language in which most of these books were written, namely 180. After English, the second most recurring original language was, somewhat surprisingly, Swedish (Van Dalen-Oskam 2021, 102). One Swedish author is represented with five books in the corpus: Henning Mankell. Three of these books are in the category crime. Remarkably, these three books end up relatively high in the ranking of literary quality among literary novels (Van Dalen-Oskam 2021, 178). The two literary novels, on the other hand, ended up among the lowest scoring literary novels, although the scores were still somewhat higher than his crime novels (Van Dalen-Oskam 2021, 193). The literary novel *Daisy Sisters* turned out to have different frequency patterns of MFWs compared to other translated novels written by male authors. However, the frequency patterns of this book were remarkably close to the frequency patterns of one of the highest scoring translations: *Norwegian Wood* by Haruki Murakami (Van Dalen-Oskam 2021, 190). Van Dalen-Oskam wonders whether this could have something to do with the fact that both works were translated into Dutch much later than they were published in the original languages Swedish and Japanese (both in the eighties) (Van Dalen-Oskam 2021, 189).

However, she did not have enough data in her corpus to investigate this assumption further. The corpus in the study I report on in this contribution, consisting of 32 books written by Mankell during his entire career, can confirm or reject this hypothesis. The following section reports on the results of the studies, and looks at genre differences, possible change over time and other factors that influence the clustering of texts.

3. Genre and Style Differences

When a book gets translated the genre classification chosen by the publisher could, at least theoretically be different in the source language. However, in the translations of the books by Henning Mankell into Dutch this is not the case. Squires (2007, 71–72) states that genre is a necessary part of book publishing. It is implemented in the whole publishing process, from cover design to advertising and what literary prizes the book qualifies for. The genre also determines on what shelf the book ends up in the bookstore or library. Because of this, Squires (2007, 71–72) concludes that genre classification is not so much a literary boundary, but rather a marketing tool. Although this might be true to some extent, multiple studies in computational literary studies have shown that it is possible to distinguish genres based on style, measured by high frequency words (e.g. Dalen-Oskam 2021; Jautze 2014; Jautze et al. 2013; Jockers 2013).

Jockers (2013, 68–70) showed that genre and style are closely linked. Jockers and his colleagues looked at various subgenres in nineteenth-century English novels. They divided the text into samples of 1,000 words and performed an unsupervised clustering using the most frequent words (MFW). The high-frequency words turned out to not only be highly successful in distinguishing samples from the same author and novel, but also placed text samples that belonged to the same genre closely together. Jockers concluded that (sub)genres have a stylistic fingerprint that can be detected by looking at high-frequency words.

Jautze (2014) investigated whether the MFWs can distinguish chick lit from literary novels. She performed a stylometric analysis using the R package stylo (Eder et al. 2016) and found that chick lit was stylistically different from high literature. High literature turned out to have a more descriptive style, whereas chick lit seemed to be more informal.

In an earlier study, Jautze et al. (2013) compared high literature and chick lit syntactically and found that novels that are classified as high literature contain more complex sentences than chick lit. High literature was also found to be richer in prepositional phrases than chick lit.

To my knowledge, there are no studies that compare the style of high literature and crime novels, the genre that Henning Mankell is most known for. The genre is sometimes even referred to as literary crime novel, indicating that it has a higher literary quality than regular crime novels or thrillers. One might expect that it is harder to distinguish between high literature and 'literary' crime novels, especially if they are written by the same author.

To find out if an analysis of the MFWs can make this distinction, I performed a stylometric analysis on the Mankell corpus using the R package Stylo (Eder et al. 2016). The Stylo package automatically compiles a list of MFWs in the entire corpus and can check which words occur relatively frequently in the various texts, based on the Delta procedure for authorship attribution (Burrows 2002). Burrow's delta looks at texts as a collection of data or 'bag of words' and disregards the context of sentences. The frequency of each word in the corpus is counted and the separate texts are compared to each other based on their frequency lists (MFWs). For this comparison, the relative, normalized z-scores are used, so differences in text length or the high impact of a small number of high-frequency words on the total outcome are ruled out (Eder et al. 2016). The distances between texts can, for instance, be visualized in a dendrogram representing the results of a cluster analysis, grouping texts that are similar to each other.

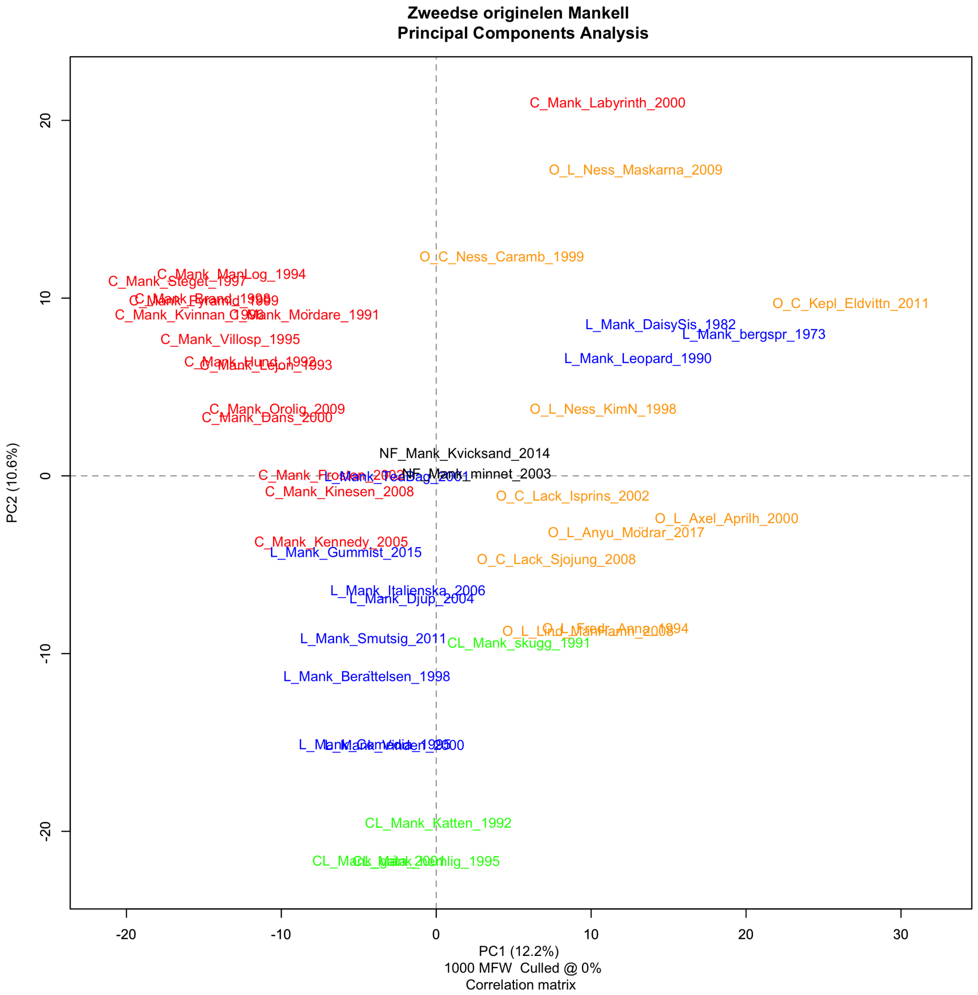

Figure 1: Cluster analysis of the Swedish books in the corpus based on the 1,000 most frequent words (culling 0, classic delta).

A cluster analysis was first performed on the Swedish corpus, to see if there are clear stylistic differences between various genres. The analysis is based on the 1,000 most frequent words in the books. The results are visualized in Figure 1: Most books are neatly clustered by genre, where L stands for literary novel; C for Crime novel; NF for Non-Fiction and CL for Children's literature, although some crime novels appear among literary novels or vice versa. This seems to be the case for crime novels written from the year 2000 onwards. It is unclear from this analysis to what extent the clustering by genre is mainly caused by genre-specific words or by other stylistic features too.

The earlier crime novels: From the 1991 *Mördare utan ansikte* (*Faceless Killers*) until *Pyramiden* (*The Pyramid*) from 1999, all belonging to the Wallander series, are in a separate cluster. This cluster has two subclusters: one for the books written in the first half of the 1990s (1991-1994), and one for the Wallander books published in the second half of the 1990s (1995-1999). Remarkably, Mankell's last Wallander book *Den oroliga mannen* (*The Troubled Man*), that was published in 2009, falls outside of this cluster. This could again be explained by the fact that this last book of the series was written ten years after the previous Wallander book, and that Mankell's style changed over time. The crime novel *Innan frosten* (*Before the Frost*) from 2002, which is written from the perspective of detective Wallander's daughter, does not belong to the Wallander cluster either. However, this book is closer in time to the other Wallander books, indicating that there are other factors that weigh in.

The books by other authors than Mankell are also clearly different from the books by Mankell. The crime novel *Carambole*, from the Van Veeteren series by Håkan Nesser, for instance, is very comparable genre-wise to Mankells Wallander series. However, author seems to be a stronger factor in the clustering of the text than genre, because *Carambole* ends up in a separate cluster and clusters with other literary novels by Nesser. Genre, in its turn, plays a more important role than time overall. If we for instance look at the two non-fiction books by Mankell, they clearly form a cluster, even though they were published eleven years apart.

However, something remarkable is going on with the three oldest books by Mankell in the corpus. These three literary novels: Mankell's debut, *Bergsprängaren* (*The Rock Blaster*) from 1973, *Daisy Sisters* from 1982 and *Leopardens öga* (*The Eye of the Leopard*) from 1990 appear closer to other authors and even cluster with the 2011 crime novel *Eldvittnet* (*The Fire Witness*) by Lars Kepler. The same is true for the crime novel *Labyrinten* (*The Labyrinth*) from a much later period (2000), that clusters with Mankell's early literary novels. textitLabyrinten is different from other works, because it was originally written as a film script and later turned into a novel. This might have influenced the style of this particular novel.

The fact that the three early Mankell novels are stylistically different from his later works seems to indicate that Mankell's writing style and word choice indeed has changed over time and confirms the findings by Van Dalen-Oskam that *Daisy Sisters* is different from other novels by Mankell. However, the texts and their MFWs have to be investigated in more detail to see how his style has changed and to ensure there are no other factors at play.

The Dutch translation corpus was analyzed using the same procedure as shown for the Swedish corpus to see if the texts appear in different clusters when they are translated. The Dutch corpus consists of the same books by Mankell and by the ten books by the aforementioned Swedish authors (Johannes Anyuru, Majgull Axelsson, Marianne Fredriksson, Lars Kepler, John Ajvide Lindqvist, Camilla Läckberg, and Håkan Nesser). This comparison can give important information about what type of MFWs influence the clustering of texts in stylometric analyses. The results are shown in Figure 2. The different titles were all labeled by genre first (L for literary novel; C for Crime novel; NF for Non-Fiction, CL for Children's literature and O for different author than Mankell). The second tag is the translator's initials, followed by the author's last name and two years, the first one indicates the year the original novel was published, the second one stands for the year the translation was published.

Overall, the results are similar to the results of the cluster analysis of the Swedish corpus. However, the genre differences seem to be slightly bigger in the translated works. Unlike the results in the Swedish corpus, all the non-Wallander crime novels end up in one cluster together. Two novels stand out in particular: the literary novel *Tea bag* from 2001, that appears close to Mankell's later crime novels and *Labyrinten*, which just like in the Swedish novels clusters with the three early literary novels by Mankell.

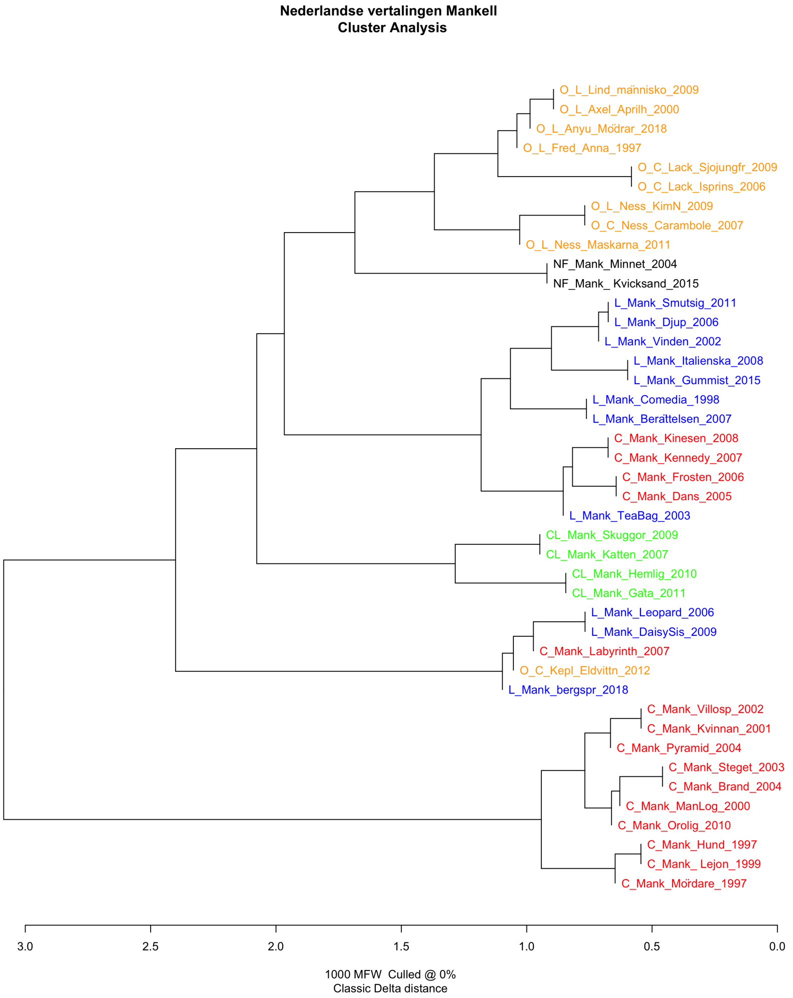

Figure 2: Cluster analysis of the Dutch translation corpus based on the 1000 most frequent words (culling 0, classic delta).

Another noticeable difference between the Swedish and the Dutch cluster analysis, is that unlike in the Swedish originals, the early literary novels in the translations are more similar to other works by Mankell than to the other Swedish authors, although Lars Kepler's *Eldvittnet* shows up in this cluster again.

4. Network Analysis of Mankell's Oeuvre

As pointed out by Eder (2015), there are a couple of problems with cluster analysis pertaining to the distance, linkage and number of features (MFWs) used for analysis. Outcomes can differ depending on the MFWs used and there is no real consensus about what the optimal number of MFWs is. These problems can partially be overcome by using a bootstrap consensus tree, because it repeats measurements for multiple numbers of MFWs, and looks for the most robust groupings across different measurements.

However, Eder (2015, 55–56) notes there is still some arbitrariness involved in the production of a consensus tree, such as how many time the analysis should be repeated, for how many words in total are considered and the underlying algorithm used for linkage. A bigger caveat for the current study, however, is that a consensus tree only looks for the closest ranking text, which means it mainly looks at the strongest similarities. In most cases, this is the authorial fingerprint.

In this paper, the central question is rather why some works within one oeuvre deviate from the majority of works and what other factor beside the author are decisive for clustering of texts. These weaker patterns might better be detected by producing a network analysis as proposed by Eder (2015). In a network analysis, not only the closest text in rank is taken into account, but also the second and third closest neighbours. These links are visualized in a network in which close similarities are shown with thicker lines and weaker links with thinner lines.

I performed a bootstrap consensus tree in Stylo and used the CSV output to create a network analysis in the open-source tool GEPHI (Bastian et al. 2009) using the ForceAtlas2 algorithm. I ran a Modularity Analysis (resolution 0.6) in GEPHI which detects communities in the network, helping to distinguish closely related topological subgroups of nodes from each other and to make clusters more visible in the network. Finally, I applied eigenvector centrality, to measure the influence of nodes in the network. Ranking the function size of nodes indicates the centrality of a work for the cluster it is in.

The results of the network visualizations are shown in Figure 3 and Figure 4. A short description of the clusters is given in the titles in red for ease of interpretation.

In general, the results shown earlier in the cluster analyses are confirmed by the consensus networks. Works cluster mainly by author and genre, although there is some overlap between crime novels and literary novels. There is a separate cluster for Mankell's Wallander series and the older literary novels are in a separate cluster.

However, there are some remarkable differences between the Swedish consensus network and the translated Dutch one. In the Swedish network the older Mankell novels cluster with two literary novels and a crime novel by Nesser as well as a crime novel

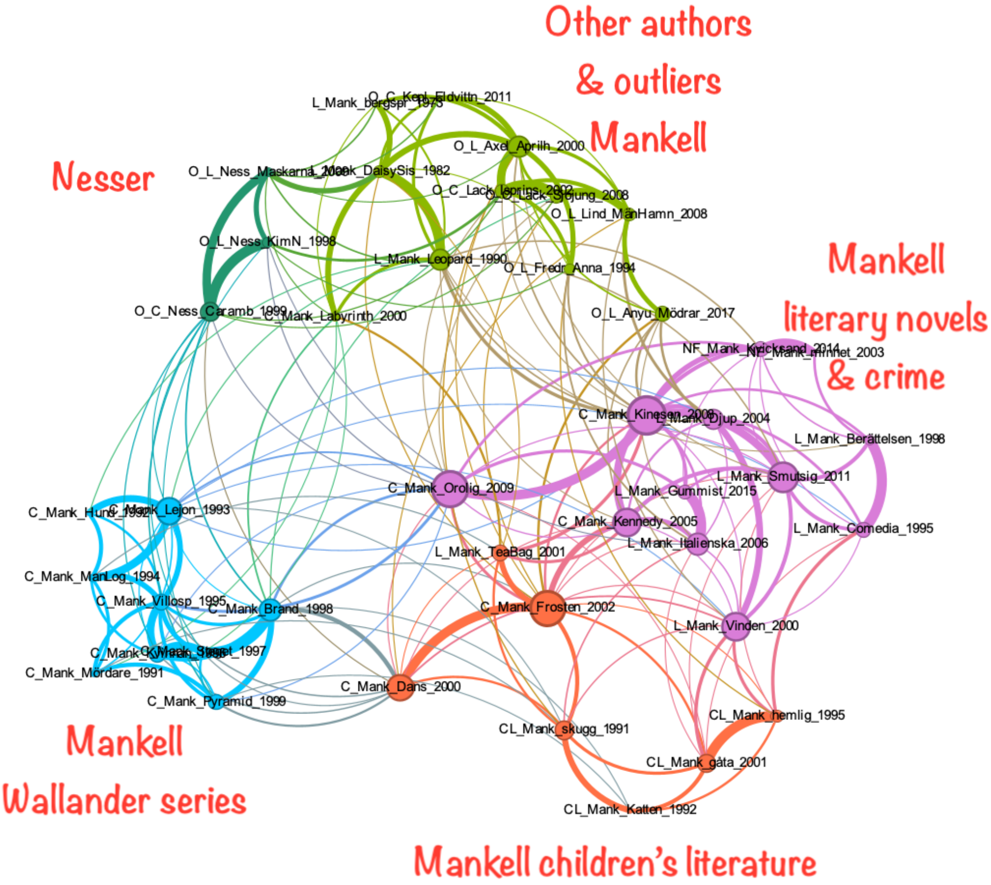

Figure 3: Consensus network of the Swedish corpus: classic Delta distance, 100–1,000 MFWs, modularity 0.6.

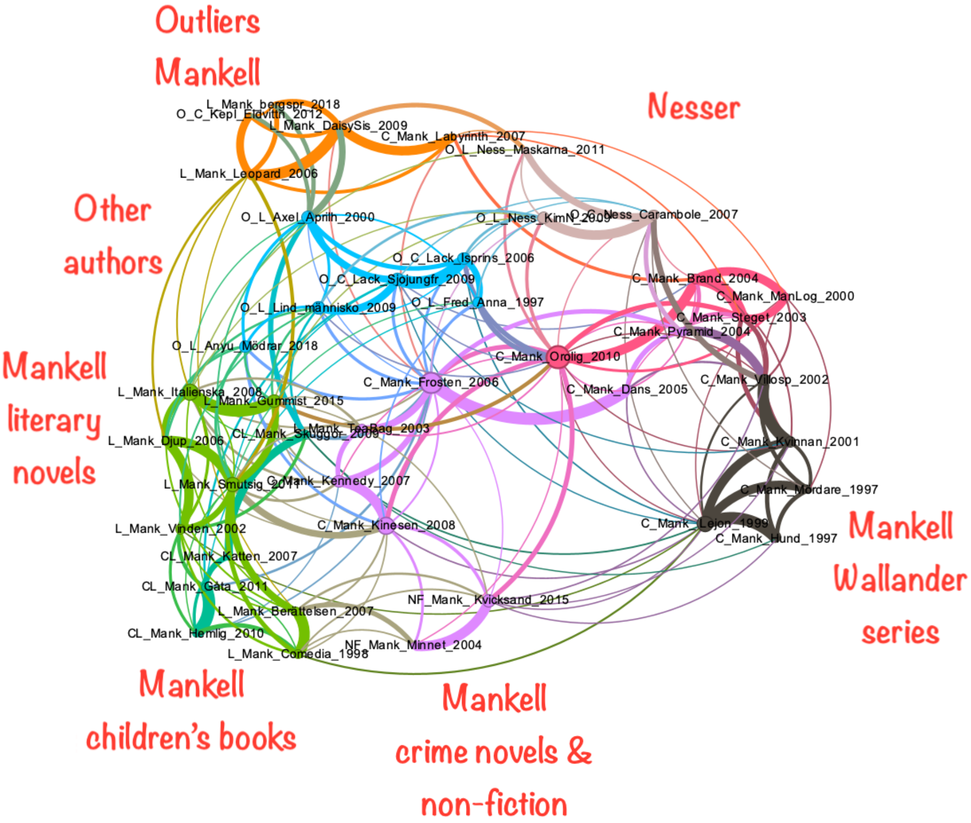

Figure 4: Consensus network of the translated Dutch corpus: classic Delta distance, 100–1,000 MFWs, modularity 0.6.

by Kepler. In the Dutch network, they are only grouped together with the crime novel by Kepler only. The consensus network of the original Swedish corpus distinguishes six clusters, whereas the consensus network of the translated corpus contains eight. Unlike the cluster analysis, where the texts were clustered more clearly by genre in the translated corpus, the network looks somewhat more messy in the translated corpus with smaller and less clearly defined clusters.

Most importantly for the current study, the same four novels that showed up as outliers in the cluster analyses: *Bergsprängaren* (*The Rock Blaster*) from 1973, *Daisy Sisters* from 1982, *Leopardens öga* (*The Eye of the Leopard*) from 1990 *Labyrinten* (*The Labyrinth*), again form a separate cluster. In the following section, the MFWs associated with these works are analysed to see why these particular novels stand out from the rest of Mankell's novels.

5. A Closer Look into the MFWs

To look at more dimensions in the data, a Principal Components Analysis (PCA) was performed on the Swedish corpus. Like a cluster analysis, a PCA also analyzes the MFWs in the dataset, but they are visualized in a scatterplot instead of a dendrogram. In a PCA multiple features are combined in an artificial variable, the so-called principal component that explains the largest proportion of the variance in the data (Jockers 2013, 65–67). On the x-axis, the first principal component is shown. The first principal component is often related to the author (Hoover 2020). The y-axis shows the second principal component. The second principal component is less obvious to interpret, it could be explained by variables like chronology or genre (Hoover 2020). These two principal components are unrelated.

I performed a classic PCA on the Swedish data in Stylo, with the Classic Delta and the correlation option, analyzing the 1,000 MFWs. The results of the PCA of the Swedish corpus are presented in Figure 5 below. The x-axis, showing the first principal component, explaining 12.2% of the variance in the data, can clearly be linked to author and is in line with the findings in the cluster analysis. The same four books that were mentioned earlier are deviant from Mankell's other works and more similar to the other Swedish writers in the corpus. The books in question are the crime novel *Labyrinten* from 2000, the two oldest books in the corpus, namely *Bergsprängaren* (*The Rock Blaster*) from 1973 and *Daisy Sisters* from 1982, and *Leopardens öga* (*The Eye of the Leopard*) from 1990. Unlike in the cluster analysis, we can now see that Lars Kepler's *Eldvittnet* is further away on the x-axis and probably clustered with these books because of the variance in the data that is represented on the y-axis.

Figure 5 shows that author and genre are still the most important factors in distinguishing between texts. However, there are a few books by Mankell that clearly behave differently and that end up closer to books by other authors. What makes these four stand out from the rest of Mankell's works?

If we perform the same PCA again, but with the option 'loadings' in Stylo, showing which words occur significantly more frequently in the texts they are close to in the graph, we might get a first impression about an important difference between the four

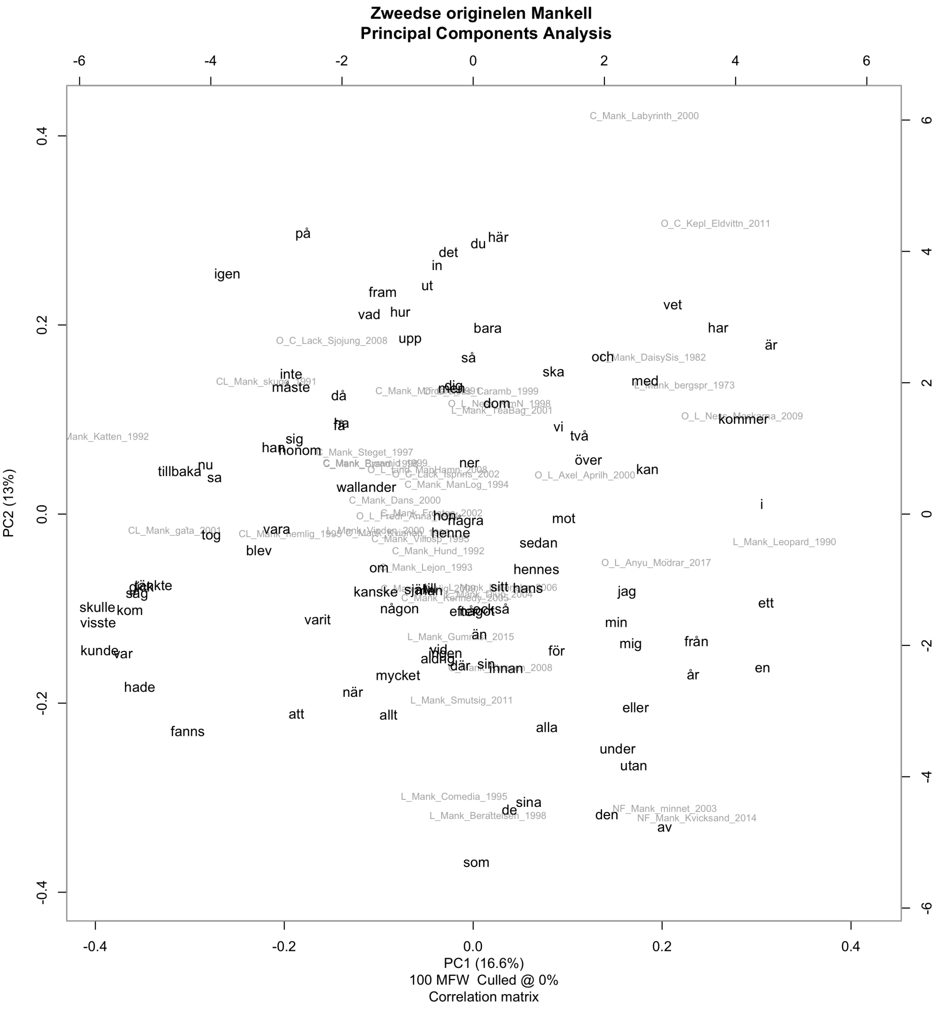

Figure 5: Principal component analysis of the Swedish corpus (1,000 MFW, Classic Delta correlation, culling 0)

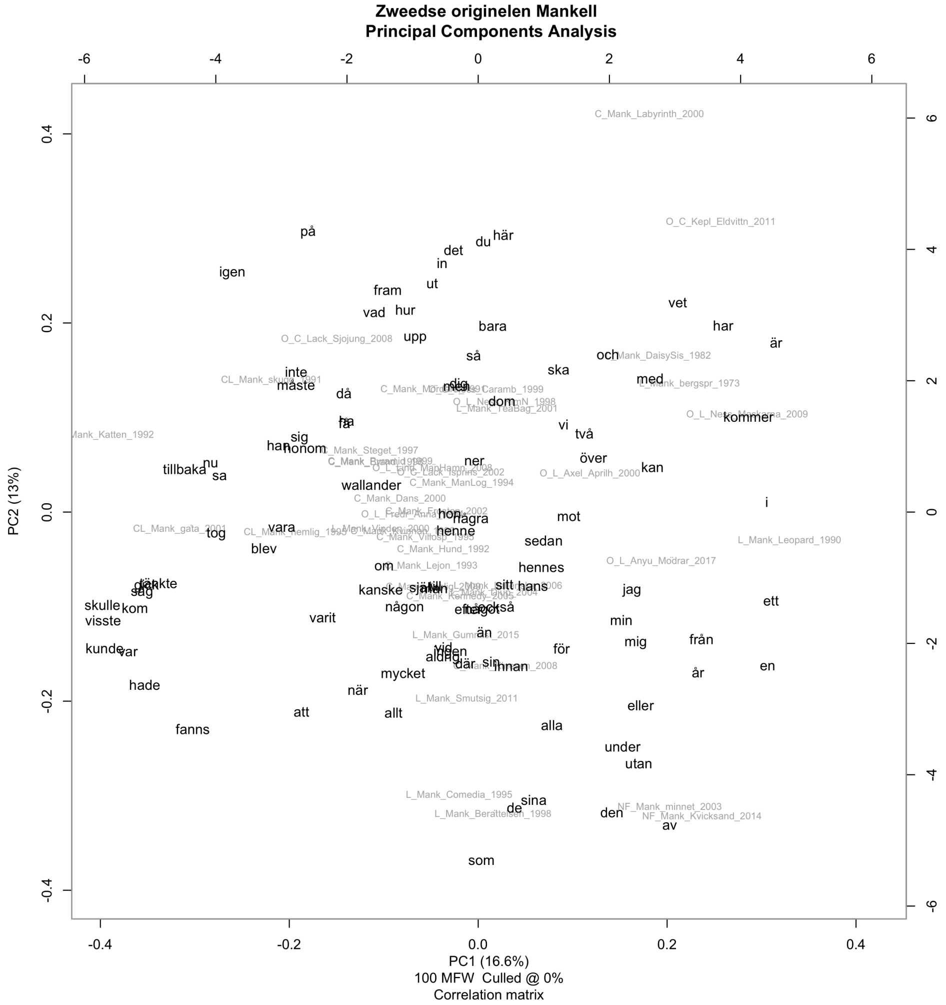

Figure 6: Principal Components Analysis showing the 100 MFW in the Swedish corpus

atypical books and the rest of Mankell's work. In Figure 6, the results of the PCA with the option loadings are shown. This analysis was performed on the 100 MFW, because a figure with 1,000 words would become illegible. This also means that the distribution of the novels on the graph is somewhat different. For instance, Kepler's novel *Eldvittnet* is now closer to Mankell's *Labyrinten*, whereas Nesser's *Maskarna på Carmine street* (2009) appears close to Mankell's older novels. Importantly, Mankell's four diverging novels still stand apart from his other novels. In Figure 6, they are shown a bit below the upper right corner. The words that are associated with these novels, and less with the other books, are: *kommer* 'come', *vet* 'know', *har* 'have', *är* 'is/are', *och* 'and' and *med* 'with'. The first four are verbs in the present tense, whereas the verbs associated with other works are all in past tense or past participles.

This indicates that rather than the chronology, the tense primarily used in the narrative, established by verb tense, might be a decisive factor in why the four mentioned books are different from other Mankell books. On closer inspection, these books as well as *Eldvittnet* by Lars Kepler are primarily written in the present tense, whereas the other works by Mankell are primarily written in the past tense. Of course, this may be related to a chronological development: Over time a writer can also change their preference for which tense to narrate a story in.

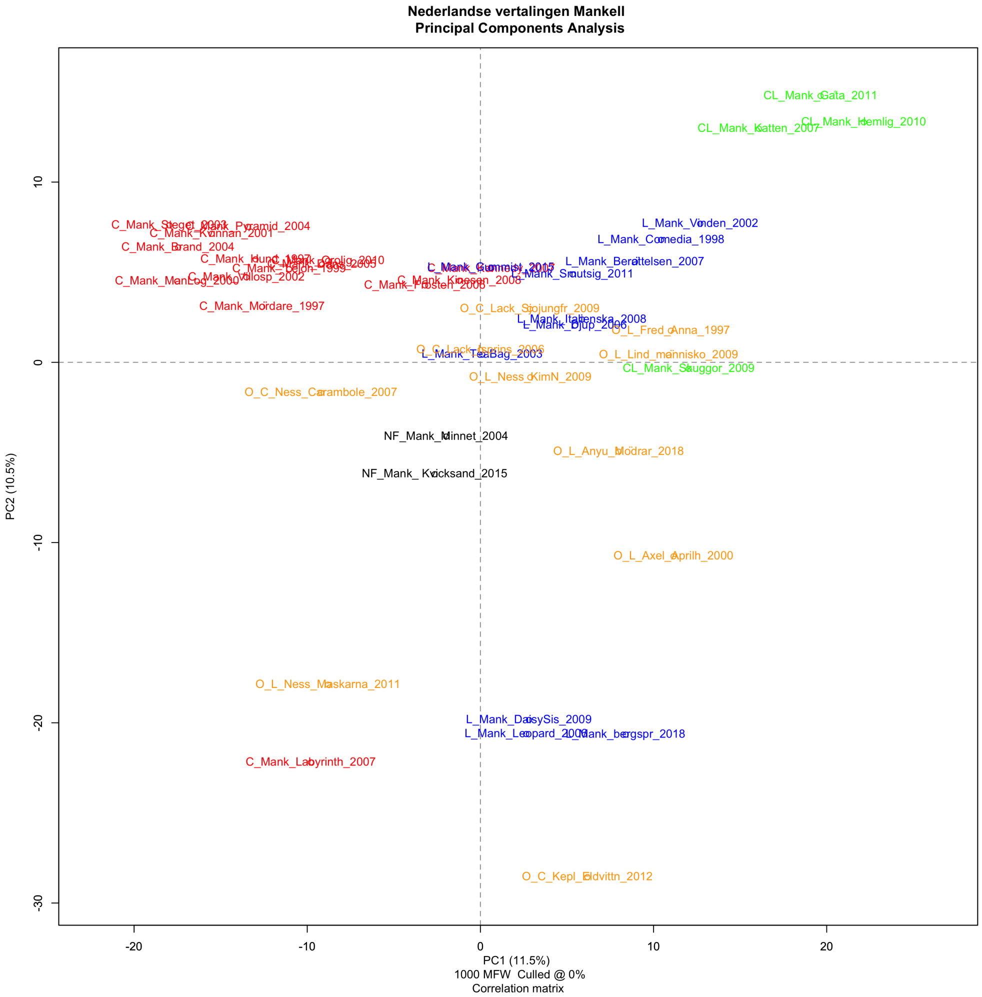

Figure 7: PCA of the translated Dutch corpus based on the 1,000 MFWs

The same procedure was followed for the translated Dutch corpus. The results are shown in Figure 7 and Figure 8. In Figure 7 the PCA for the translated Dutch corpus is shown, which is in many ways comparable to the results of the Swedish PCA. One remarkable outcome is that Mankell's Children's books are quite different on the x-axis, where this was not the case at all in the Swedish results. Another remarkable finding is that some novels by other writers in the corpus, namely Håkan Nesser, Camilla Läckberg and Marianne Fredriksson, end up very close to the literary novels by Mankell and in between books by Mankell in different genres.

Otherwise, the same four books (*Leopardens öga*, *Bergsprängaren*, *Labyrinten*, and *Daisy Sisters*) diverge in the translation corpus. The PCA with the loadings function (Figure 8) clearly shows that this is likely caused by the narrative tense again. Words that occur more frequently in these books are: *moet* 'has to', *kan* 'can', *is* 'is', *heeft* 'have' and *weet* 'know' whereas past tense verbs occur more frequently in other works. An important difference between Swedish and Dutch is that Swedish only has tense marking on verbs whereas Dutch has tense and person marking. This also means that Swedish verbs probably tend to end up higher in the list of MFWs because there are fewer possible forms compared to Dutch where the same verb is spread out over more possible forms.

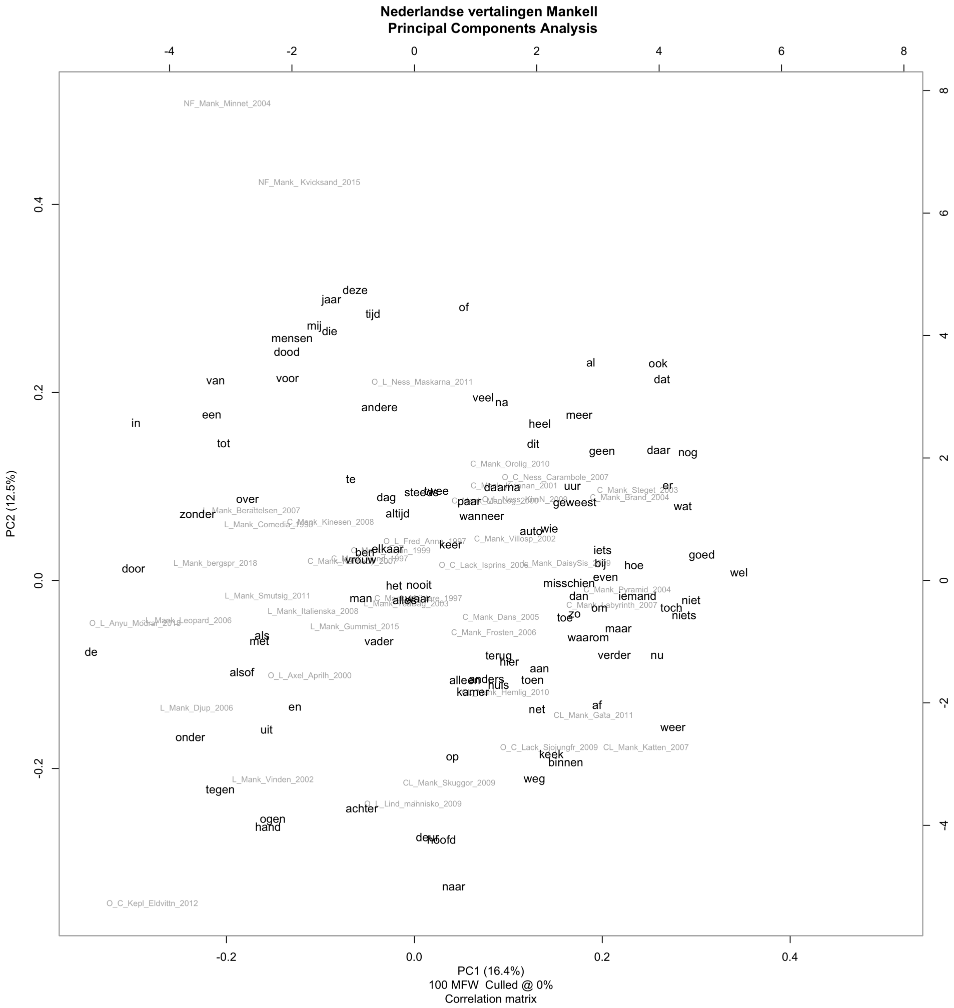

Figure 8: PCA (loadings) of the translated Dutch corpus based on the 100 MFWs

5.1 Verb Tense and Perspective

In the remaining part of this section, I will elaborate on the results of the study so far, to get more insight into the stylometric methods used and the studied texts. It is important to look beyond the analyses of Delta distances to see what is behind the measurements and which words are decisive in the clustering of texts.

The results so far show that verb tense is an important factor for the outcome in stylometric methods based on the MFWs, because there are many verbs (with tense marking) among the MFWs. Along similar lines, narrative perspective might also play an important role, because pronouns are very frequent words. In order to get a good indication of the predominant narrative perspective in the books in the current corpus, I applied Van Rossum's I-index to the data (Van Rossum et al. 2020). The authors applied both a machine learning and a narratology-based approach in which they computed the ratio of pronouns. Both methods turned out successful in determining the narrative perspective of texts, although the second approach was slightly more robust and yielded a perfect 1.00 score. This perfect score was possible, because Rossum et al. (Van 2020) cleaned the data from dialogue. The narrative perspective was already known, so the predictions could be tested for their accuracy. For now, this is not possible in the Mankell corpus, but the ratio of pronouns can still give a good indication of a book's narrative perspective.

Van Rossum's I-index is focused on the first person narrative perspective but can be applied to other perspectives as well. I computed the I-index and the he-index, she-index and (singular) you-index (du-index) for both the Swedish originals and the Dutch translations. For the he-index (han-index), for instance, I did this by adding the relative frequency scores of *han* 'he', *honom* 'him' and *hans* 'his' as calculated in Stylo and divided this number by 1 + the relative frequency scores for all the pronouns in the text. The reflexive possessive pronouns *sin*, *sitt* and *sina* 'his/her/their' were left out of the equation, because they are used to refer to both male and female antecedents. For the she-index (hon-index) I followed the same procedure counting the pronouns *hon* 'she', *henne* 'her' (object form) and *hennes* 'her' (possessive). Finally for the singular you-index (du-index) I divided the sum of the relative frequencies of *du* 'you' (singular), *dig* and *dej* 'you' (object form in two spelling variants), *din*/*ditt*/*dina* 'your' (singular in three inflection forms) by the relative frequencies of all pronouns combined.

Figure 9 shows the results of the indexes in a graph. The ratio of pronouns gives a good indication of the narrative perspective(s) in the texts. Only the results of the Swedish corpus are shown here, because I observed no big differences between the Swedish and the Dutch ratios. Mankell's texts are ordered chronologically from oldest to most recent. The other authors are in random order. The first part of the bars on the bottom left side shows the I-index. In most texts, this index is between 0.10 and 0.30. Clear peaks in the I-index can be detected for the two non-fiction books *Jag dör, men minnet lever* (*I Die, but the Memory Lives*) from 2003 and *Kvicksand* (*Quicksand: What It Means to Be a Human Being*) (2014), which indeed are mainly written from first person perspective.

Peaks in the I-index can also be observed for *Italienska skor* (*Italian Shoes*)(2006) and *Svenska gummistövlar* (*After the Fire*) (2015) which are both literary novels with the same main character, Fredrik Welin, written from an I perspective. Two books by Håkan

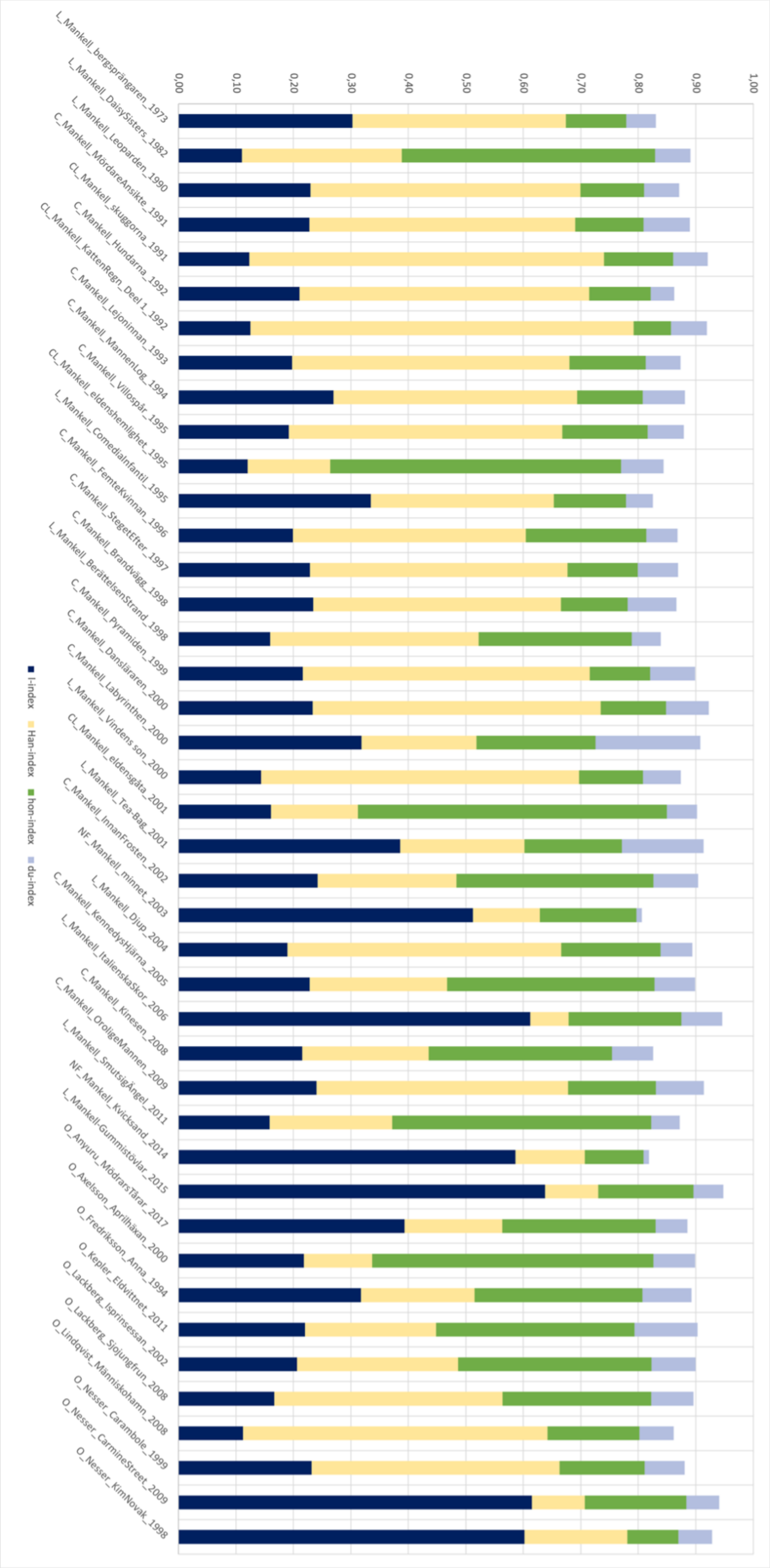

Figure 9: Indication of narrative perspective in the books in the Swedish corpus, measured by I-index (dark blue), Han-index (yellow), Hon-index (green) and du-index (light blue)

Nesser: *Maskarna på Carmine street* (2009) and *Kim Novak badade aldrig i Genesarets sjö* (1998) also score high on the I-index. This might explain why Nesser's books also appeared close to Mankell's outliers in the PCA.

The second part of the bars in Figure 9 show the han-index (he-index). Most books by Mankell score high on the han-index (he-index) and are indeed mainly written from a third person male perspective. The third part of the bar show the she-index (hon-index) and it is most interesting to compare these two indexes directly. *Daisy Sisters* (1982) is one of the books that score very high on the she-index (hon-index), which makes sense, because it is a novel about three generations of women. The combination of a deviant verb tense (present tense) and a third person female perspective could very well explain why this particular book appears to be an outlier in the cluster analyses and the PCAs. Two of the children's books (*Eldens gåta* and *Eldens hemlighet*) also score relatively high on the she-index. Some books have a more evenly divided ratio between pronouns. This is especially the case in *Labyrinten* (2000) which also was one of the clear outliers in the PCA and cluster analysis together with the earliest Mankell novels. Narrative perspective thus seems to be an important explanatory factor. The analysis of narrative perspective and narration tense leads to useful new observations about what can influence MFW scores for Swedish and Dutch and shows how a novel like *Daisy Sisters* differs from other novels by Henning Mankell.

To get more insight into how much of the outcome was influenced by narration tense and narrative perspective, we should only look at the words that are not clearly linked to verb tense and narrative perspective. Stylo has the option to analyze the corpus using an 'existing word list' which enables the researcher to look at specific sets of words. I excluded all verbs marked for tense and all personal pronouns to better determine how big their influence is on the analyses. I then ran another PCA with the 'loadings' function. The resulting PCA without personal pronouns and verbs indicating tense is shown in Figure 10. This figure clearly shows that now three of the four deviant books are much closer to Mankell's other works, at least on the x-axis, and they no longer form a separate cluster. *Leopardens öga* is also closer to other books in the same genre, but especially *Bergsprängaren* and *Labyrinten*, and to a lesser extent also *Daisy Sisters*, are still more distant from other works by Mankell.

Figure 11 shows the Dutch PCA excluding tense-marked verbs and personal pronouns. In this graph it becomes clear that *Daisy Sisters* (together with *Labyrinten*) diverges more from other Mankell novels on the y-axis than *Bergsprängaren*. So, the translations and the original Swedish texts are different in this perspective. The word frequency patterns of the translations of the other Swedish authors are also much harder to distinguish from the frequency patterns in Mankell's books compared to the results of the analysis of the Swedish texts. This implies that some aspects of style get lost in translation.

5.2 The Influence of Register

Due to space limitations, I will exclusively focus on one of the earliest Mankell novels *Daisy Sisters* in this last section. We have seen that this novel is deviant in style, partly because of the use of the present tense and because it is one of the relatively few books by Mankell written from a female third person perspective. A third reason for why *Daisy Sisters* has deviant word frequency patterns compared to Mankell novels that were

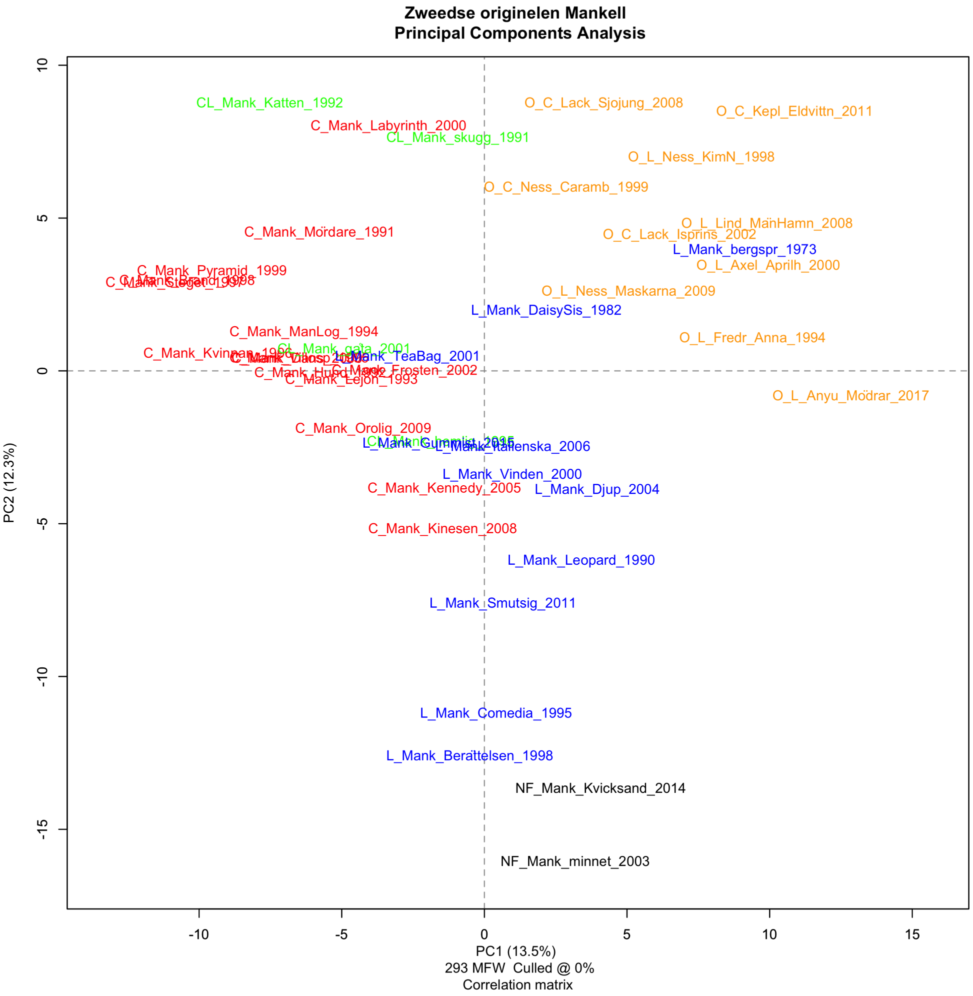

Figure 10: PCA (classic) of the Swedish corpus excluding tense-marked verbs and personal pronouns based on the 1,000 MFWs

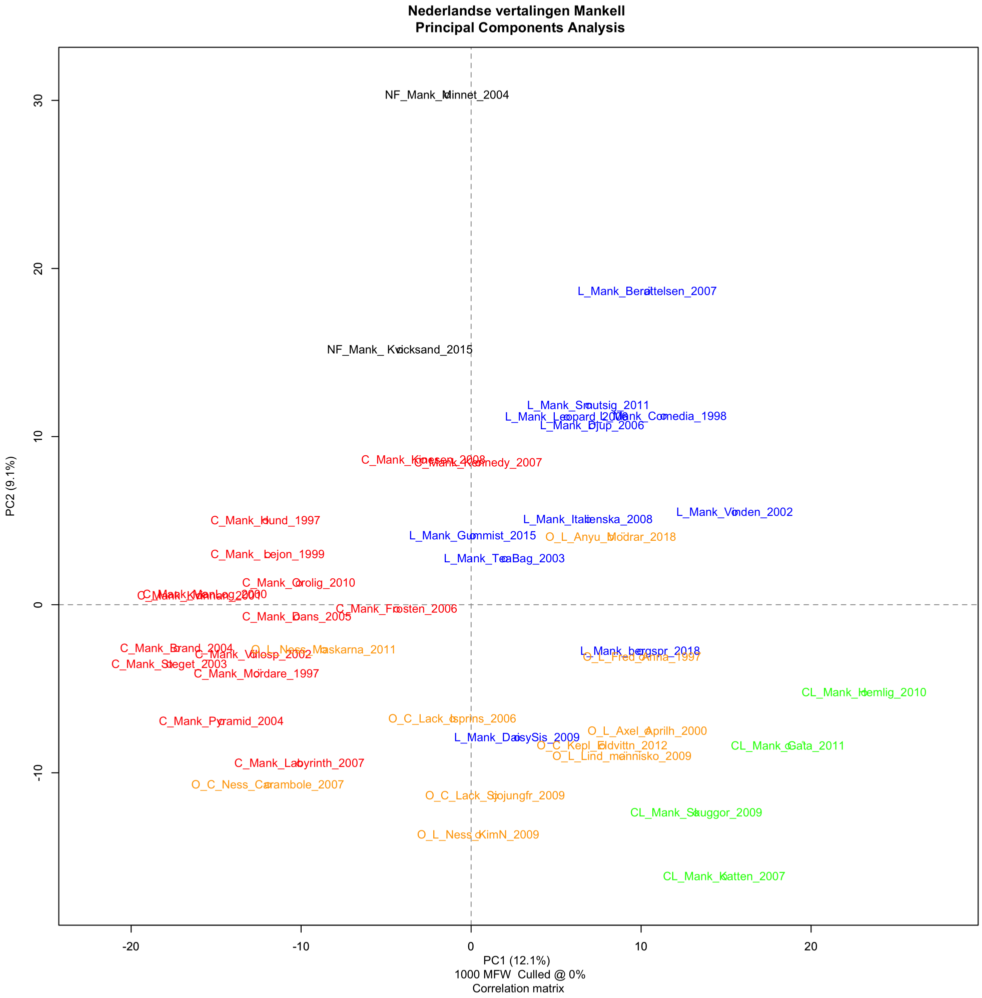

Figure 11: PCA (classic) of the translated Dutch corpus excluding tense-marked verbs and personal pronouns based on the 1,000 MFWs

published later, can be detected if we look at Zeta scores.

Zeta was initially introduced by Burrows (2007), and later on improved by Hugh Craig (Craig and Kinney 2009). Burrows's Delta, the method used in this study so far and the method most often used in stylometry, relies on high frequency words (MFWs). Burrows's Zeta and Craig's Zeta, on the other hand, analyze the middle frequency words and measure distinctiveness or keyness of keywords in a corpus relative to a reference corpus (Hoover 2010). Middle frequency words are usually more meaningful than high frequency words, because high frequency words generally are function words (Rybicki 2016, 751). In a Zeta analysis, the texts to be analyzed are first divided into equal segments, then the dispersion of each word in the two separate corpora is registered, by counting how many segments it occurs in at least once (Craig and Kinney 2009).

Stylo can generate wordlists containing the most distinctive keywords in two opposing texts or corpora (Eder et al. 2016). I compiled a primary corpus, consisting of *Daisy Sisters* and a secondary, reference corpus containing all other literary novels by Mankell. I did this for both the Swedish corpus and the Dutch translation corpus. I only selected literary novels in order to avoid getting lists with genre specific words. I then performed the command oppose() in Stylo to analyze *Daisy Sisters* and the reference corpus using Craig's Zeta. Zeta is the sum of the proportions of sections from Daisy Sisters in which each word occurs and the sections of other works in the corpus in which it does not (Hoover 2010). This can point out stylistically interesting characteristics of a text or a corpus. I used samples of 3,000 words.

This generates two word frequency lists: one list with words that are relatively more frequent compared to the other books in the initial corpus in *Daisy Sisters* and a list with words that are used relatively less frequent compared to the other books in the initial corpus. From these lists I selected the twenty most distinctive words, excluding names and verbs, because I was interested in whether there were other stylistic differences besides tense and narrative perspective. The results for the Swedish data are shown in Table 1.

The first obvious difference has to do with spelling conventions and register. *Mej* and *mig* are spelling variants of the same word: 'me'. The variant *mej*, occuring relatively more frequent, in *Daisy Sisters*, is the less formal variant which is closer to speech, whereas *mig* is the official variant. The same is true for *dig* and *dej* and *sej* and *sig*. This pattern could also be detected in the spelling of certain verbs, like *säga* 'write', which occured relatively more frequent in the alternative, informal spelling variant *säja* in *Daisy Sisters*. There are other words on the keyness list that confirm the idea that *Daisy Sisters* is written in a more speech-like, colloquial style. Examples are *jo* 'yes' used after a negation and *ju*, a discourse particle that is especially frequent in spoken language. Similarly, *visst* is a colloquial form for 'of course' and *vadå* a colloquial form for *vad* 'what'. The keyness list also contains swear words and curse words, which are clearly associated with everyday, informal language, *jävla* 'fucking', *herregud* 'lord' and *fan* 'damn'.

The following example from *Daisy Sisters* contains three keywords from the (longer) keyness list:

> Men vad spelar det för roll att **morsan** är här och säger att hon skäms? Hon

more frequent		less frequent	
mej	‘me’	mig	‘me’
dej	‘you’(object)	dig	‘you’(object)
jo	‘yes’(after negation)	mina	‘my’ (plural)
fan	‘damn’	sist	‘last’
ju	‘of course’	havet	‘the sea’
herregud	‘lord’)	genast	‘immediately’
ja	‘yes’	eftersom	‘because’
sej	(reflexive pronoun)	min	‘my’ (singular)
visst	‘certainly’	död	‘dead/death’
sen	‘then/late’	land	‘country/land’
jävla	‘fucking’	vatten	‘water’
vadå	‘what’	fartyg	‘ship’
nej	‘no’	bland	‘among’
alltså	‘so’	våra	‘our’ (plural)
omedelbart	‘immediately’	oss	‘us’
lust	‘desire’	vattnet	‘the water’
världen	‘the world’	gryningen	‘the dawn’
helst	‘preferably’	människor	‘people’
värre	‘worse’	långsamt	‘slowly’
väl	‘well/surely (discourse particle)’	djur	‘animal(s)’

Table 1: 20 most distinctive keywords based on Craig’s Zeta in *Daisy Sisters* compared to other literary novels in the Swedish corpus, excluding verbs and names

> kan **ju** inte veta något. Mer än... Ja, **vadå**? Så minns hon allt blod och förstår att det var därför hon måste gå till sjukhuset.

The English translation of this passage is as follows: [3]

> What does it matter that **mom** is here saying she’s ashamed? She can’t know anything, **right**? More than.. well **what**? Then she remembers all the blood and realizes that’s why she had to go to the hospital.

In both the English translation and the official Dutch translation of this passage the colloquial style is at least partially lost:

> Maar wat maakt het uit dat **haar moeder** hier is en zegt dat het een schande is? Ze weet **toch** nergens van. Alleen dat ... Ja, **wat**? Dan herinnert ze zich al het bloed en ze begrijpt dat ze daarom naar het ziekenhuis moest.

Discourse particles in general are very hard to translate, because they can have various meanings depending on context (Aijmer 2008). Here *ju* is translated, but there is no Dutch or English equivalent that is equally frequent and associated with speech as much as the Swedish word. The two other colloquial words in this short passage are translated into standard Dutch, which leads to a loss of this style feature.

A final result from the Zeta analysis is that different synonyms are used in the primary corpus and the reference corpus. In the list of distinctive words, *omedelbart* is preferred in *Daisy Sisters* and *genast* is avoided. These words are synonyms and both mean ‘immediately’ with no difference in register.

[3] My translation.

Table 2 shows the list with distinctive words based on Craig's zeta in the Dutch translation corpus.

more frequent		less frequent	
nou	'well'	ineens	'suddenly'
immers	'after all'	onze	'our'
ja	'yes'	zee	'sea'
opeens	'suddenly'	water	'water'
verdomme	'damn'	ons	'our'
best	'best/okay'	iedere	'every'
nee	'no'	vervolgens	'then'
fabriek	'factory'	vlug	'fast'
kennelijk	'apparently'	hoewel	'although'
allemaal	'all'	mij	'me'
wanneer	'when'	dood	'dead/death'
dus	'so'	lichaam	'body'
flat	'flat, appartment'	slechts	'only'
minder	'less'	eiland	'island'
natuurlijk	'of course'	boot	'boat'
zin	'desire'	zwart	'black'
eens	'once/agreed/discourse particle'	diep	'deep'
weleens	'sometimes'	amper	'barely'
niks	'nothing'	haven	'harbour'
zomaar	'just (like that)'	hierheen	'this way, here'

Table 2: 20 most distinctive keywords in *Daisy Sisters* based on Craig's Zeta, compared to other literary novels by Mankell in the Dutch translation corpus, excluding verbs and names

The register difference is not as obvious as in the Swedish list, although words like *nou* 'well', *nee* 'no', *ja* 'yes', *verdomme* 'damn' do point in the direction of register and speech-like language or dialogue-driven text. *Immers*, which is on top of the list of distinctive words in the Dutch translated corpus, is a good example of translationese. It is the translation of the previously mentioned discourse particle *ju*. In terms of meaning, this translation is accurate, but *immers* does not at all belong to the same register. While *ju* is associated with spoken language, *immers* is almost exclusively used in written language and has a somewhat archaic connotation. Again, this indicates that the speech-like, informal style gets partially lost in the Dutch translation. In the Dutch list with distinctive keywords there are also two synonyms both meaning 'suddenly': *opeens* is preferred in *Daisy Sisters* whereas *ineens* is preferred in the other books in the corpus. This can likely be explained by the individual preference of the translator. However, more research about the influence of the translator on style is necessary to confirm this.

6. Conclusion

In this paper, 32 books by the Swedish writer Henning Mankell were investigated using stylometric methods, to find out whether his style changed measurably over time, or if some of his books deviate stylistically from his other works for other reasons. 10 books by other Swedish authors were added to the corpus as a reference. The study also gives more insight into the methods that are frequently used in stylometry, such as cluster analysis and PCA, that basically are black boxes, because they give little information about the stylistic features that differ between texts. For this purpose, the original

Swedish texts were also compared to the Dutch translations of the same 42 texts to determine how translation and language influence the results of stylometric analyses.

Cluster analyses and PCAs of the data showed that works were clustered by author in the first place and secondly by genre, although there were a few exceptions. The division into genre was somewhat stronger in the translated corpus. The analyses also seemed to indicate that the factor time explains part of the variance. However, on closer inspection, verb tense rather than year of publication turned out to be the decisive factor: The most deviant books in the corpus were primarily written in the present tense, whereas most other books were predominantly written in the past tense. Moreover, narrative perspective also influenced the results noticeably. An analysis of the pronoun ratios in the works in the corpora indicated that the majority of the novels in the corpus had a dominant third person male perspective. Books that mainly had a first person perspective tended to cluster together, just like books with a third person female perspective. After leaving out pronouns and verbs marked for tense the deviant works appeared considerably closer to Mankell's other novels.

Finally, an analysis of the data based on Craig's Zeta (Craig and Kinney 2009) showed that words most distinctively used in the original Swedish *Daisy Sisters* were often colloquial words with a speech-like connotation. However, the most distinctive words in the Zeta analysis for the translated Dutch corpus, were not as clearly related to register. This can be due to the different language and language specific features or due to inherent characteristics of translated texts in general. More research on different languages and translations would be useful to get a better understanding of this process. In a follow-up study I intend to investigate the style differences between individual translators and how they can be detected and measured. Since this paper only considered word frequency patterns, future work could also look at syntactic measures for style. Furthermore, as pointed out by one of the reviewers, it would be interesting to lemmatize all verbs to view their importance regardless of their inflection.

This study has shown that Zeta analysis and a closer look at word lists in stylometric studies can give useful insights into the specific style features that make texts different from each other instead of focusing on the fact that they differ alone.

7. Data Availability

Data can be found here: `https://doi.org/10.5281/zenodo.10362679`.

8. Software Availability

Software parameters can be found here: `https://doi.org/10.5281/zenodo.10362679`.

9. Acknowledgements

I would like to thank Karina van Dalen-Oskam for our fruitful conversations and for her valuable feedback on earlier versions of this article.

10. Author Contributions

Martje Wijers: Conceptualization, Writing – original draft, Formal analysis, Investigation

References

Aijmer, Karin (2008). "Translating Discourse Particles: A Case of Complex Translation". In: *Incorporating Corpora. The Linguist and the Translator*. Ed. by Gunilla Anderman and Margaret Rogers. Multilingual Matters, 95–116.

Arvas, Paula and Andrew Nestingen (2011). *Scandinavian Crime Fiction*. University of Wales Press.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy (2009). "Gephi: An Open Source Software for Exploring and Manipulating Networks". In: *International AAAI Conference on Weblogs and Social Media*, 361–362. 10.1609/icwsm.v3i1.13937.

Berglund, Karl (2012). *Deckarboomen under lupp: Statistiska perspektiv på svensk kriminallitteratur 1977–2010*. Avdelningen för litteratursociologi, Uppsala universitet.

Burrows, John F. (2002). "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship". In: *Literary and Linguistic Computing* 17 (3), 267–287. 10.1093/llc/17.3.267.

— (2007). "All the Way through: Testing for Authorship in Different Frequency Strata". In: *Literary and Linguistic Computing* 22 (1), 27–47. 10.1093/llc/fqi067.

Can, Fazli and Jon M. Patton (2004). "Change of Writing Style with Time". In: *Computers and the Humanities* 38, 61–82. http://www.jstor.org/stable/30204925 (visited on 01/22/2024).

Craig, Hugh and Arthur F. Kinney (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press.

Dalen-Oskam, Karina van (2021). *Het raadsel literatuur: Is literaire kwaliteit meetbaar?* Amsterdam University Press.

— (2023). *The riddle of literary quality : a computational approach*. Amsterdam University Press.

Eder, Maciej (2015). "Visualization in Stylometry: Cluster Analysis Using Networks". In: *Digital Scholarship in the Humanities* 32 (1), 50–64. 10.1093/llc/fqv061.

Eder, Maciej, Jan Rybicki, and Mike Kestemont (2016). "Stylometry with R: A Package for Computational Text Analysis". In: *The R journal* 8 (1), 107–121. 10.32614/RJ-2016-007.

Herrmann, J. Berenike, Karina van Dalen-Oskam, and Christof Schöch (2015). "Revisiting Style, a Key Concept in Literary Studies". In: *Journal of Literary Theory* 9 (1). 10.1515/jlt-2015-0003.

Hoover, David L. (2010). "Teasing Out Authorship and Style with T-tests and Zeta". In: *Digital Humanities*, 168–170. https://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-658.html (visited on 01/22/2024).

— (2020). *Modes of Composition and the Durability of Style in Literature*. Routledge.

Jacobsen, Kirsten (2012). *Mankell om Mankell*. Leopard Förlag.

Jautze, Kim (2014). “Measuring the Style of Chick Lit and Literature”. In: https://pure.knaw.nl/ws/portalfiles/portal/742412/Jautze_Kim_Measuring_the_style_of_chick_lit_and_literature.pdf (visited on 01/24/2024).

Jautze, Kim, Corina Koolen, Andreas van Cranenburgh, and Hayco de Jong (2013). “From High Heels to Weed Attics: A Syntactic Investigation of Chick Lit and Literature”. In: *Proceedings of the Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, 72–81. https://aclanthology.org/W13-1410 (visited on 01/22/2024).

Jockers, Matthew L. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Ríos-Toledo, Germán, Juan Pablo Francisco Posadas-Durán, Grigori Sidorov, and Noé Alejandro Castro-Sánchez (2022). “Detection of Changes in Literary Writing Style Using N-grams as Style Markers and Supervised Machine Learning”. In: *Plos one* 17 (7). 10.1371/journal.pone.0267590.

Rossum, Lisanne van, Joris van Zundert, and Karina van Dalen-Oskam (2020). “I Catching: Computationally Operationalising Narrative Perspective for Stylometric Analysis”. In: DH Benelux. 10.5281/zenodo.3855652.

Rybicki, Jan (2016). “Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies”. In: *Digital Scholarship in the Humanities* 31 (4), 746–761. 10.1093/llc/fqv023.

Squires, Claire (2007). *Marketing Literature: The Making of Contemporary Writing in Britain*. Palgrave Macmillan.