Karina van Dalen-Oskam
The secret life of scribes. Exploring fifteen manuscripts of Jacob van Maerlant's *Scolastica* (1271)

Abstract -- The paper explores whether stylometric methods used in non-traditional authorship attribution can help to gain more insight in how medieval scribes dealt with the text they were copying. The research corpus consists of transcriptions from all more or less complete manuscripts of Jacob van Maerlant's *Scolastica*, a Middle Dutch translation/adaptation of Peter Comestor's Medieval Latin *Historia scholastica*. Five episodes were selected totalling 70 samples of around 1200 tokens each. Cluster analysis and principal component analysis are used to compare the copies per episode. The results based on the Middle Dutch texts were highly influenced by irrelevant spelling variation. Since the aim was to explore differences on the content level of the copies, the analyses were extended to lemmatized versions of the texts. The results based on the lemmatized texts are a good starting point for the exploration of the ways in which *Scolastica* scribes dealt with the work they were copying. The main results of the exploration are that most copies of an episode cluster closely together, with only occasional outliers. The outliers are not the same for each episode, which suggests that a separate analysis and explanation per episode is needed - scribes may have had reasons to elaborate on one topic and to leave another untouched, for instance under the instruction of a patron. The paper closes with some pointers for the next steps in the research, which are expected to need other kinds of methods for analysis.

1. Introduction

The aim of this paper is to test whether stylometric methods used in non-traditional authorship attribution can help to gain more insight in how medieval scribes dealt with the text they were copying. What we would like to know is what the range of freedom was that scribes allowed themselves, or were allowed to (or ordered to) by others, in copying a text. In our research we are primarily interested in content differences. It is well-known that medieval scribes easily and unsystematically adapted the spelling of the text they copied, but these changes are not expected to reflect any fundamental differences in view on the text itself. Variation on all levels can be seen as a key characteristic of medieval texts (Cerquiglini 1999), but for this research we want to exclude mere spelling idiosychrasies.

Earlier research (see below) suggested that the application of non-traditional authorship attribution methods can be of use for finding out more about medieval authors and scribes, although several uncertainties seem to hamper the use of medieval texts. Often, we do not know the name of an author. The text written by the original author(s) usually is not available anymore but only handed down in copies (of copies (of copies))). Not all intermediate copies will have survived, which means that the relation between the copies is even more difficult to establish. The dates of the copies usually are uncertain. And, last but not least, the known freedom that scribes allowed themselves makes it

uncertain whether authorship attribution methods would identify the author, a scribe or combination of successive scribes, or - even - a modern editor of the text.

Bernard Cerquiglini's *Éloge de la variante* (1989, translated into English in 1999 as *In praise of the variant*) eloquently pointed out the necessity to study all versions of a medieval text as individual texts. Cerquiglini also knew that this was easier said than done. A manual approach is almost impossible, so he hoped the computer would be able to help (Cerquiglini 1999: 79-80). Greco and Schoemaker (1993) succinctly described the problems in these areas and outlined some of the tools needed to explore text versions and their similarities and differences. Brefeld (1994) was one of the first to apply cluster analysis and factor analysis on a corpus of related medieval texts. Spencer and Howe (2001, 2002) developed a mathematical model for scribal accuracy and for estimating differences between manuscripts. In stemmatological research, building a family tree of text representations using e.g. cladistic methods from biology, certain scribal differences are a means to draw up the stemma but these stemmas usually are not studied on their own as a source of knowledge about the (history of the) text (Spencer c.s. 2006, Windram c.s. 2008).

Slowly, more empirical and computational research is shaping up and being published. In his research on different copies of the Canterbury Tales, Thaisen (in press) focusses on spelling differences in order to get insight in manuscript dependencies, developing a method to pinpoint exemplar changes (where a scribe stopped copying one manuscript and continued copying from another manuscript). In an authorship analysis of the Middle Dutch *Walewein* written by two medieval authors, van Dalen-Oskam and van Zundert (2007) found that in the only manuscipt of this text, copied by two scribes, the change of scribes was more sharply visible than the change of authors, especially in the 50 highest frequency words, when they applied a windowing version of Burrows's Delta. Kestemont and van Dalen-Oskam (2009) used a lazy machine learning technique on *Walewein* and on the corpus to be introduced in the next section of the current paper to get more insight in the nature of scribal differences. They found that scribes only edited texts in a shallow and superficial way, leaving authorial features generally intact on deeper levels.

2. Methods and Corpus

The non-traditional authorship attribution methods chosen for this research are cluster analysis and principal component analysis. In cluster analysis, objects are grouped in a tree-like visualization according to their level of similarity, in this case based on the vocabulary (words and frequencies) of the texts (Hoover 2010: 252-254). Principal component analysis represents objects in a scatterplot according to different components which are mathematically calculated. The first component reflects the characteristic in which the objects differ most from each other, etc. (Holmes 1994: 99; Craig and Kinney 2009: 30-31). Both present visualizations in a way that helps to think about the data and both lead to new ideas as to what the next steps in the research could be. Our area of research is Middle Dutch literature. A corpus containing transcriptions of copies of the same text was not available yet, so we had to create one first.

'Alphabetical' overview			**Chronological overview**		
A	Berlin	Dated 1331	**C**	Brussels	Around 1285
B	Brussels	Around 1300	**B**	Brussels	Around 1300
C	Brussels	Around 1285	**M**	The Hague	Around 1330
D	The Hague	Mid 14th century	**A**	Berlin	Dated 1331
E	The Hague	Around 1400	**G**	Groningen	Around 1339
F	The Hague	Around 1400	**D**	The Hague	Mid 14th century
G	Groningen	Around 1339	**K**	London	1370-1385
H	Leiden; OT+NT	Around 1465	**L**	London	Dated 1393
I	Brussels; OT	Around 1450	**E**	The Hague	Around 1400
J	Leiden	Around 1451	**F**	The Hague	Around 1400
K	London	1370-1385	**I**	Brussels; OT	Around 1450
L	London	Dated 1393	**J**	Leiden	Around 1451
M	The Hague	Around 1330	**N**	The Hague	Dated 1453
N	The Hague	Dated 1453	**H**	Leiden; OT+NT	Around 1465
O	The Hague; OT	Around 1475	**O**	The Hague; OT	Around 1475

Table 1 The fifteen manuscripts of *Scolastica* that were used (fragments from other manuscripts were not included in the research). The places mentioned are the cities in which the library is located where the manuscript is currently kept. OT stands for Old Testament, NT for New Testament. Thirteen manuscripts contain OT, NT, and the adaptation of *De bello judaico*. The list is based on Postma (1991) with two small changes in the names of the manuscripts; her "bu" is our I, and her "hk" our O. The date of I is based on Deschamps and Mulder 2000: 1. The date of K is based on Van der Vlist and Rudy 2010: 39.

Manuscript	**Eva 'E'** 591-812= **222 lines**	**Debora 'D'** 7303-7520= **218 lines**	**Judith 'J'** 17444-659= **216 lines**	**NT 'M'** 22979-23219 **241 lines**	**Josephus 'T'** 33296-33511= **216 lines**
A	+	+	+	+	+
B	+	+	+	+	+
C	+	+	+	+	+
D	+	+	+	+	+
E	+	+	+	+	+
F	+	+	+	+	+
G	+	+	+	+	+
H	+	+	+	+	-
I	+	+	+	-	-
J	+	+	+	+	+
K	+	+	+	+	+
L	+	+	+	+	+
M	+	+	+	+	+
N	+	+	+	+	+
O	42 lines	139 lines	+	-	-

Table 2 Overview of the samples and of the availability of the samples in the manuscripts. The number of lines and the line numbering is that in the (edition of the) oldest of the manuscripts, manuscript C (Gysseling 1983). Other copies occasionally have less or more lines than C. In Manuscript O, several pages are missing.

Not many Middle Dutch texts are extant in a substantial number of copies. We chose a popular work by the most prolific Flemish author Jacob van Maerlant: the *Rijmbijbel* ('Rhyming Bible'), also called *Scolastica*, which is a translation/adaptation of the Medieval Latin *Historia scholastica* written by Peter Comestor, to which van Maerlant added an adaptation of Flavius Josephus's *De bello judaico (The Jewish war)*. Van Maerlant finished this work in 1271, and many fragments and fifteen manuscripts (though not all containing all parts of the text) survive, dating from ca. 1285 to the end of the fifteenth century. Most of the *Scolastica* manuscripts show no scribal collaboration: the complete manuscript was written by one scribe. Exceptions are the Eva sample in manuscript C, which is written by another scribe than the other four samples in C (Gysseling 1983: XII), and possibly manuscript I, which may have been written by several scribes (Deschamps and Mulder 2000: 1). Furthermore, it seems that no two manuscripts are written by the same scribe. Even when this had been the case, we would have considered them as different texts, since they may have been copied from different exemplars and/or with different intentions, e.g. due to different patrons or different intended users.

The oldest of these manuscripts, C, is available in a good edition (Gysseling 1983); the edition is also digitally available lemmatized and tagged for parts of speech at the Institute for Dutch Lexicology (www.inl.nl). One other manuscript was already edited in the nineteenth century, but the text needed to be digitized and the transcription needed to be carefully checked and adapted to modern standards. Transcriptions of the other manuscripts had to be especially made for this research.[1] Because of the length of the complete *Scolastica* (almost 35,000 lines), we had to work with samples. We chose 5 samples of 215-240 lines from different parts of the text, and transcribed the parallel texts (if available) from all 15 manuscripts. All samples contained episodes concerning biblical women: from the Old Testament about Eva and about Debora, and from the Apocryphal Books the one about Judith, from the New Testament about Martha and Mary and Mary Magdalen, and from *The Jewish War* about the unhappy Mary from Transjordania. This means that there is a certain level of content consistency in the episodes, although the stories about the women differ significantly.

This resulted in 70 short samples, together consisting of 87,617 tokens. The manuscripts are indicated by the sigla A, B, C, D, E, F, G, H, I, J, K, L, M, N and O (cf. Table 1 for more details about the manuscripts themselves and Table 2 for the availability of the different samples in the different manuscripts).

3. Predictions

If we want to answer our main question whether stylometric methods used in non-traditional authorship attribution can help to gain *more insight* in how medieval scribes dealt with the text they were copying, we need to go into the insights traditional methods such as close reading would give us. These then can be phrased as predictions for the results of the non-traditional approach and can later be used to set against the actual results, thus clearly highlighting the new information yielded by the stylometric methods. The differences between predicted and actual results could be seen as pointing to the 'secret life' of scribes – what the eye of the human scholar is not able to see.

While transcribing the text samples we carefully read them and we noticed that two manuscripts stood out compared to the others. The most notable is manuscript I, dating from around 1450 and only containing the Old Testament part and therefore only presenting the first three samples: Eva, Debora, and Judith. In the third of these, the Judith sample, manuscript I first runs parallel to the other manuscripts, but after a while it starts to show more and more differences in the text and even in the rhyme words. Several lines and even larger episodes are unique for this text. Therefore, we expect this sample to show up as a significant outlier compared to the other fourteen Judith samples in our multivariate analysis. The other manuscript standing out in the eyes of a human reader, but less than the Judith episode in manuscript I, is manuscript E, written around 1400 and containing all five samples. In several of the five episodes it is clear that the scribe (or the scribe of the exemplar of this text or the scribe of the examplar of that text (etc.)) adapted the content of the text on a micro-level. As it seems, the scribe had difficulties understanding certain sentences and wanted to make the text clear for his/her readers by 'correcting' them. But these corrections are in fact often mistakes - which we know, because E's adapations do not agree with what Peter Comestor wrote in his *Historia Scholastica*, Maerlant's direct source text, or with the Latin Bible, the most important source text of Comestor and also an important secondary source for van Maerlant. An intriguing example is found in the Judith sample. Judith is brought to Holofernes, who is sitting 'Onder eenen diere sporware' ('under an expensive canopy'). Manuscript E reads, however, that Holofernes is sitting 'Ond*er* .i. boom vp sijn hant .i. sporeware' ('under a tree, on his hand a sparrow hawk'). The scribe responsible for the reading in E obviously did not know the Middle Dutch word *sporeware* in the meaning 'canopy', but only in the meaning of 'sparrow hawk' and thought the text needed correction...[2] These kind of corrections occur in all five samples in manuscript E. In the last one, from the Josephus part, E shows even more differences from the other manuscripts, sometimes even differing rhyme words. Since the differences in the first four samples seem relatively minor, we would expect E not to differ too much from the other manuscripts there. A larger difference, however, is expected in the fifth sample, in a comparable (but much less significant) way as I differs from the other manuscripts in the Judith sample.

With these predictions based on our reading experience we will move on to the results of the measurements. We will present graphs for the Judith episode and will succinctly describe the results for the other episodes. A more elaborate description of all results can be found in the Appendix.

4. Measurements on the Middle Dutch text

Although we want to explore content differences between the manuscripts and would like to ignore spelling differences, the first measurements we present are done on the Middle Dutch transcriptions as they are, in their very diverse spelling. The measurements were done with the Minitab software (version 16), using The Intelligent Archive for the export of the sample data in the necessary format for further analysis in Minitab.

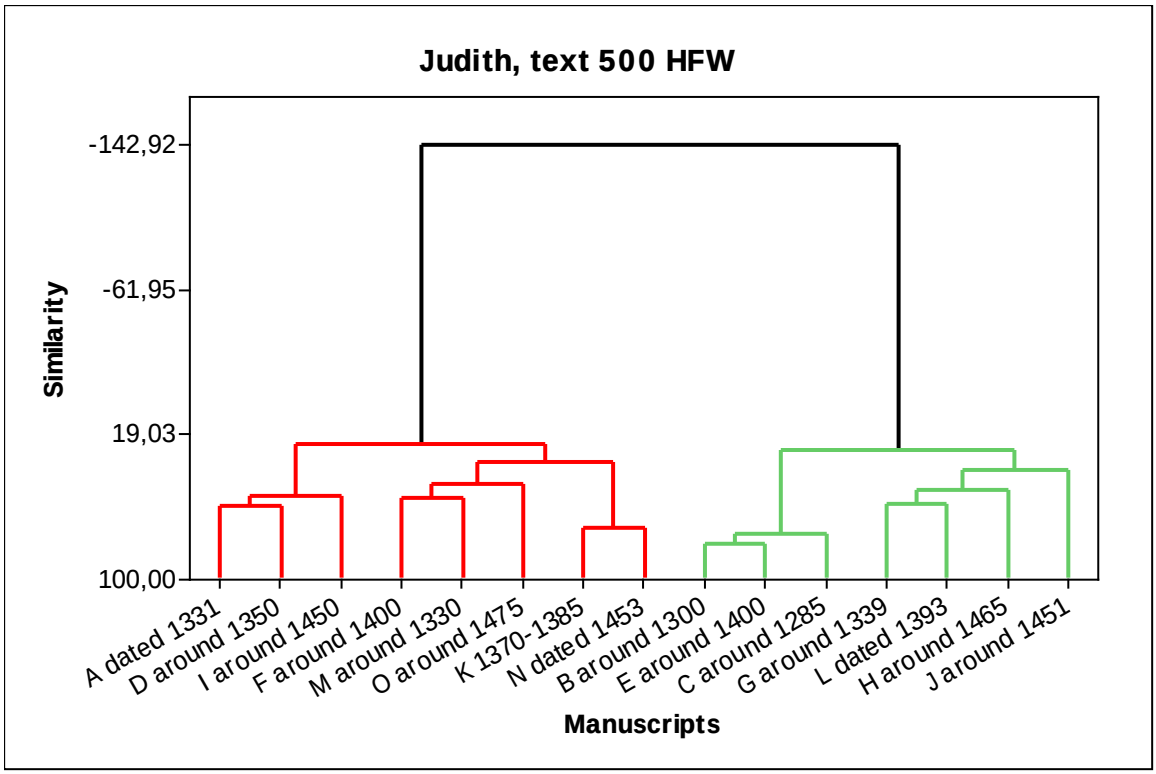

Fig. 1 Cluster analysis based on the Middle Dutch text of the Judith episodes, using the 500 most frequent words (types)

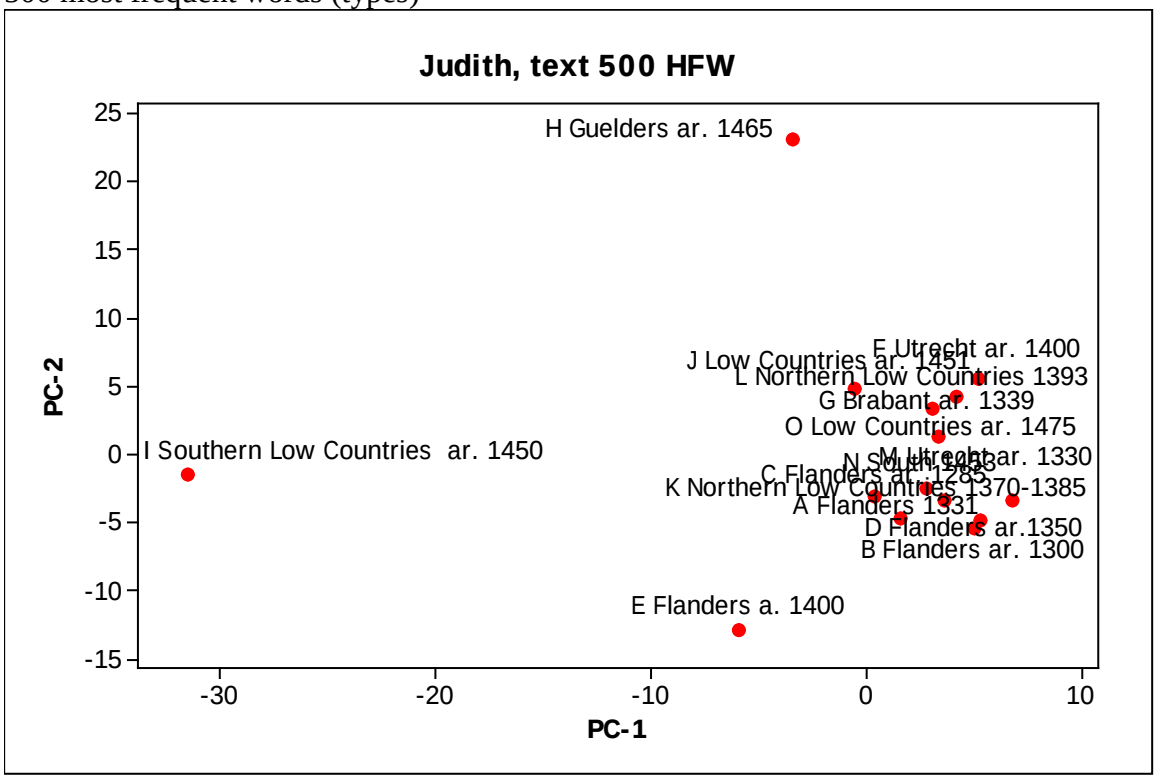

Fig. 2 Principal component analysis based on the Middle Dutch text of the Judith episodes, using the 500 most frequent words (types)

Since Minitab is limited to measuring 999 rows of data, this meant we could not take all word forms into account for the multivariate measurements on the Middle Dutch text - all Judith episodes together, for instance, have 1949 different words (types). We did many different measurements using different amounts of types; below we present the ones using the 500 most frequent types. The cluster analysis shows that the two main groups of manuscripts consist of eight and seven manuscripts. The principal component analysis presents a slightly different view, with most of the manuscripts clustering together and only three having some distance to this main group.

The biggest difference, shown on the first component (the horizontal axis), is calculated for manuscript I. This is the manuscript that has an increasingly idiosyncratic version of the Judith story, as described in section 3. Manuscript E differs slightly on the second component, which also agrees with our predictions in section 3. However, manuscript H differs more than E from the main group on this component, which we did not predict based on our close reading. Manuscript H is a late manuscript, but certainly not the only one. What may be unique for this manuscript, however, is that it is located around the Dutch - German border, the eastern part of the Middle Dutch language area. This could mean that Eastern spelling idiosynchrasies in some high-frequency words are responsible for the distance to the main group.

The cluster analyses for the other episodes (not shown here in a graph) also show no single manuscript standing out. In the principal component analysis, manuscript I is not only an outlier for Judith but also for Debora, which we did not predict. Manuscript E is a (predicted) relative outlier in the Jewish War episode.

When we look at the loadings for the Judith episode, we find that some of the highest frequency words are variant spellings of the same word:

hi – hy (pronoun *he*)
si – sy (pronoun *she*)
zi – zy (pronoun *they*)
soe – so (*thus*)
Holofernes – Olofernes

This seems to suggest that the measurements on the Middle Dutch text are indeed highly influenced by irrelevant spelling differences. If we want to abstract from spelling, we need to do the measurements on normalized text, or on lemmatized text. We opted for the last choice, expecting to get the best connection to the content of the manuscripts from the highest abstraction level, even reducing different verb forms to one and the same lemma.

5. Measurements on the Lemmas

The lemma, for the sake of convenience the Modern Dutch head word, groups spelling variants of the same noun under the same head word, or the many inflected forms of an adjective, and the forms of verbs in all their tenses and so forth. To enable the exploration of possible differences in the use of parts of speech, each token was also tagged with a code denoting the part of speech. Since no lemmatizer/tagger for Middle Dutch was

available yet when we started our research, we lemmatized the samples and tagged them for parts of speech manually, assisted by several Perl-scripts. The lemmas have the form of the Modern Dutch dictionary entry (or the form the Modern Dutch entry would have had, had the word survived into present-day Dutch). We differentiated between ten parts of speech: noun, proper name, adjective, main verb, copula / auxiliary verb, numeral, pronoun, adverb, preposition, and conjunction.

For the Judith episode, lemmatization and part of speech tagging resulted in reducing the amount of data from 1946 types to 642 lemmas - counting the same lemma with two (or more) different PoS tags as two (or more) lemmas. To add to the spelling examples given in section 4:

hi – hy both get the lemma HIJ
si – sy, zi - zy all get the lemma ZIJ
soe – so both get the lemma ZO
Holofernes – Olofernes both get the lemma HOLOFERNES

The cluster analysis of the manuscripts with the Judith episode (Fig. 3) now clearly separates mansucript I from the other manuscripts. In the principal component analysis (Fig. 4) the differences between the main group and manuscript I as well as E are much larger than in the graph based on the Middle Dutch text, but again I differs on the first component and E on the second. Manuscript H, an outlier on the second component in Fig. 2, does not stand out anymore. This seems to indicate that the outlying quality was based on spelling variation rather than content differences.

The cluster analyses for the other episodes show no single manuscript standing out for the episode about Eva. For Debora, however, manuscript I is clearly separated from the rest, which when only reading the texts we did not notice. Manuscript E in the Debora sample is the most extreme among all other manuscripts clustering together. In the New Testament sample about Martha, Mary and Mary Magdalen, E does not stand out and is part of the largest cluster. In the *Jewish war* sample, E forms its own cluster, next to the other manuscripts.

In the principal component analysis, the Eva sample shows a wide spread of all manuscripts with none really standing out, although manuscript E is on the fringes and manuscript I is not. In the Debora sample, manuscript I is an outlier on the first component, which we did not predict. Manuscript E stands apart from the group on the second component. In the New Testament sample E does not stand out and is part of the largest group (manuscript I does not have the New Testament and the Josephus part). Surprisingly, for the New Testament sample two other mansucripts, N and F, are outliers, in different components, which they are in none of the other samples. Reading the samples did not prepare us for this. In the *Jewish war* sample, manuscript E opposes the other manuscripts on the first component. Additionally, manuscript G proves an outlier, be it on the second component - which is also new information from the traditional reading perspective.

The principal component analysis thus shows more manuscripts to be outliers than the cluster observations does, adding to manuscripts E and I a special role for N and F in the New Testament sample, and G for the *Jewish war* sample. It should also be remarked, that there seems to be a trend in the course of the texts, toward ever more

differences in manuscripts I and E, from which the New Testament episode in E is an exception.

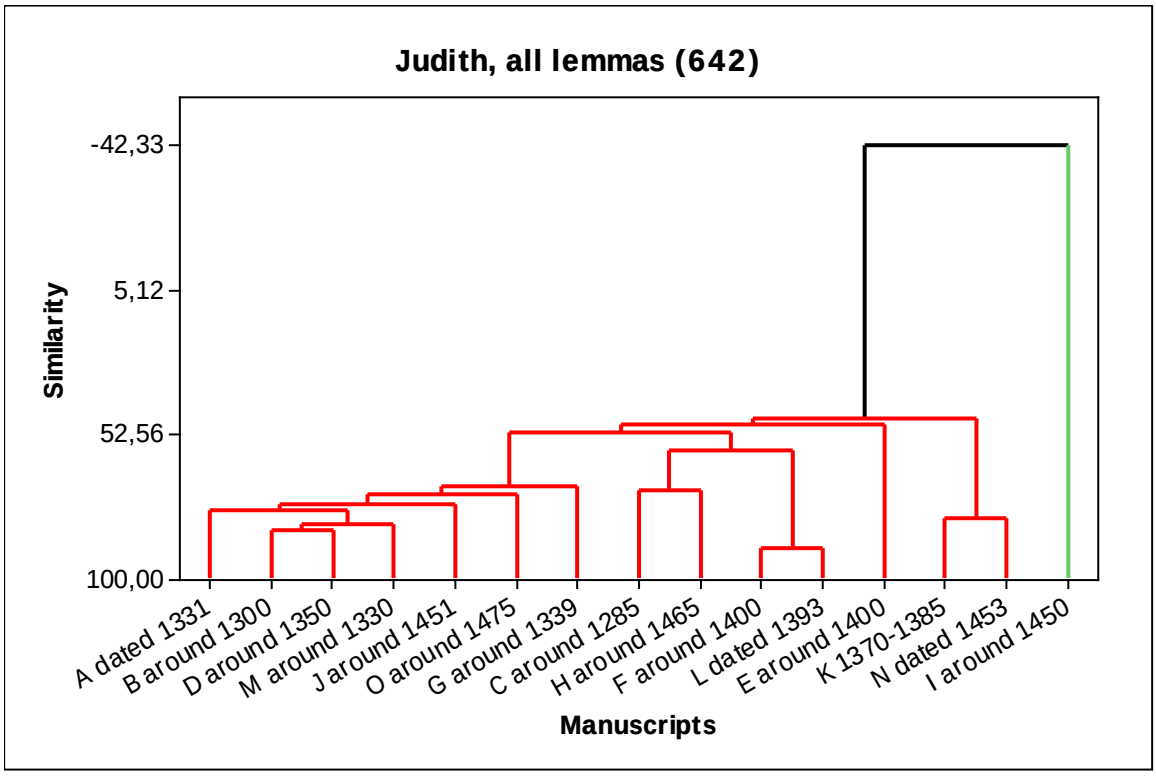

Fig. 3 Cluster analysis based on all lemmas of the Judith episodes

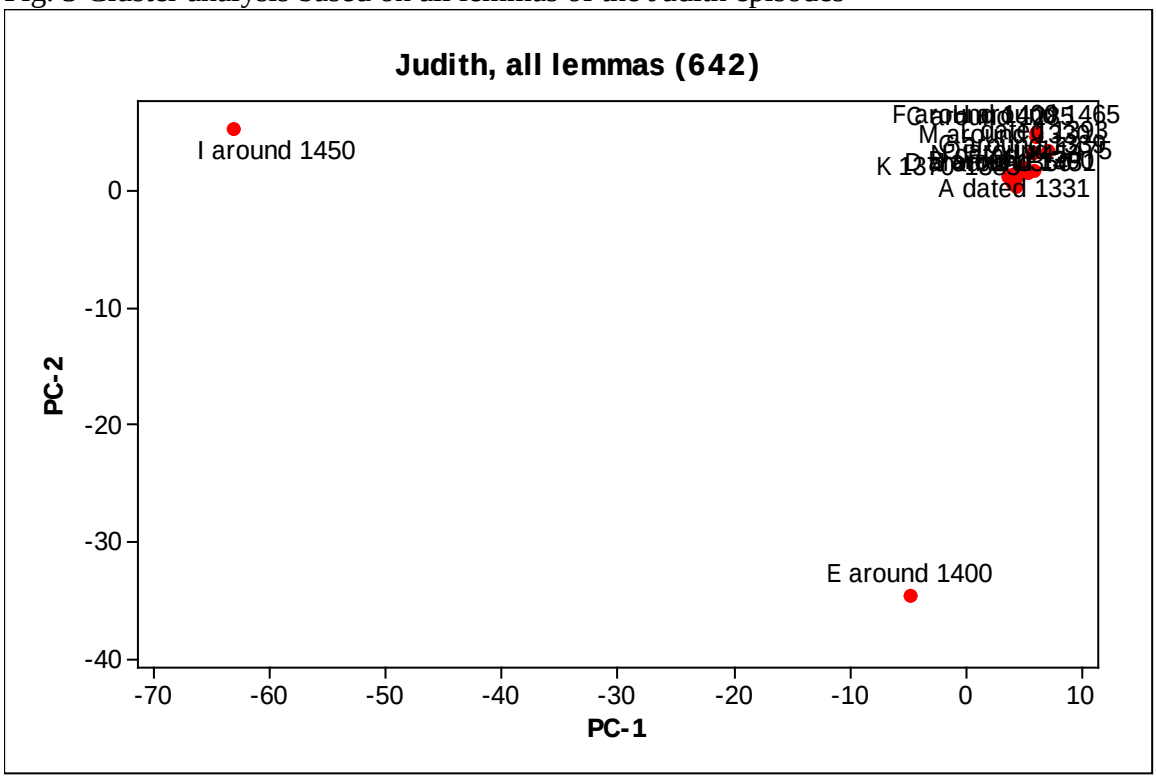

Fig. 4 Principal component analysis based on all lemmas of the Judith episodes

A plot revealing the loadings of the PCA in Fig. 4 (fig. 4a) shows so many lemmas in the extremities that the plot is not very helpful for further analysis. This may be inherent to the fact that we are dealing with copies of the same text, in which there are hardly any vocabulary differences but the differences are all in the details of the frequencies of the same set of lemmas.

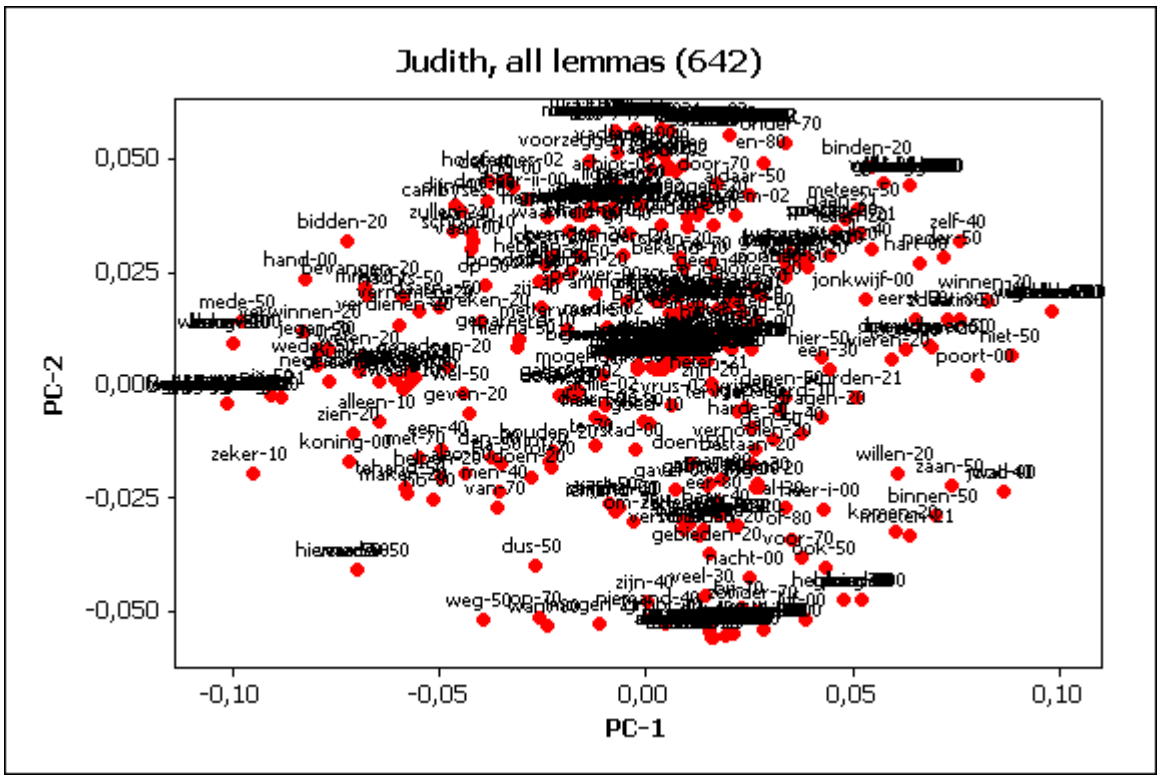

Fig. 5 Loadings of the principal component analysis based on all lemmas of the Judith episodes as shown in Fig. 4

This is confirmed when we look at an overview of the scores of all lemmas on PC-1 and PC-2. The ranges for both are rather small. The score range for PC-1 (where manuscript I was a big outlier) is between -0.056 and 0.057; Fig. 4b shows that more than 100 lemmas score in the two extremities of this range. The range for PC-2 (with manuscript E being an outlier) is between -0.102 and 0.098, where most lemmas score around 0.02 and 35 lemmas score around -0.01. A first look at the lemmas scoring at the extremities does not show any clear trends as to content or part of speech, but a closer look at them in follow-up research may still yield some useful information.

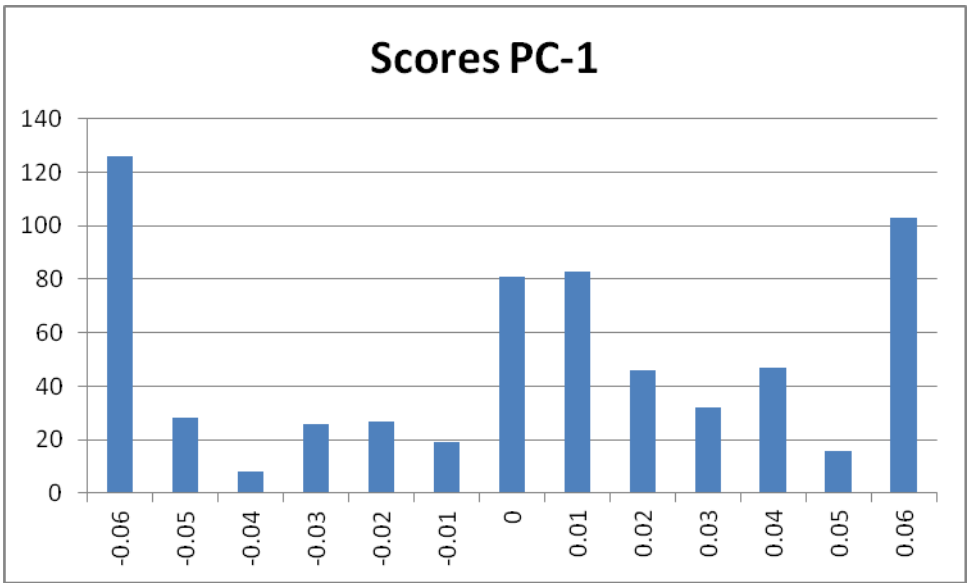

Fig. 6 Number of lemmas (vertical axis) per (rounded) score on PC-1 (horizontal axis)

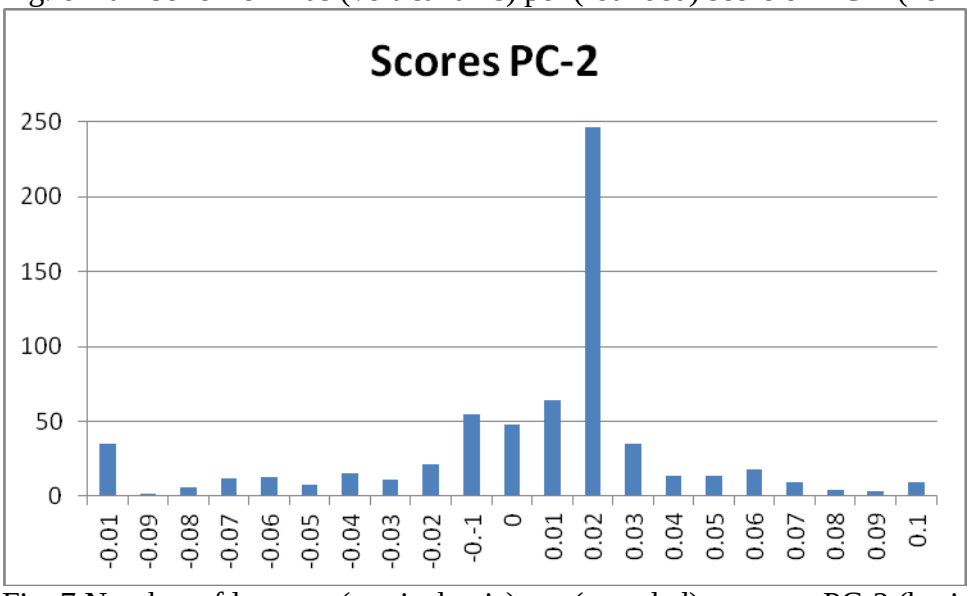

Fig. 7 Number of lemmas (vertical axis) per (rounded) score on PC-2 (horizontal axis)

An experiment with Burrows's Delta Procedure (Burrows 2002) taking into account the 150 highest frequency lemmas in all of the Judith samples led to the numbers in Table 3.

Manuscript	count	sum	mean(=delta)	stdev
D	125	18.292729	0.146342	0.152065
K	127	18.911179	0.148907	0.160308
B	129	20.119246	0.155963	0.188407
A	129	20.670763	0.160238	0.200395
J	131	21.044738	0.160647	0.224539
M	127	22.120806	0.174180	0.196136
N	128	23.026195	0.179892	0.243907
G	129	24.315943	0.188496	0.200230
O	128	26.937486	0.210449	0.299380

L	125	26.750524	0.214004	0.261642
C	126	27.696051	0.219810	0.283555
F	125	28.570724	0.228566	0.304883
H	126	33.859472	0.268726	0.334306
E	130	48.975728	0.376736	0.410536
I	129	77.099439	0.597670	0.594144

Table 3. Delta (column 4) from lowest to highest for each Judith episode compared to all Judith episodes together.

These Deltas confirm the PCA results in Fig. 4, showing Manuscript I with the largest difference, followed by Manuscript E. The Delta measurement even gives more insight in the cluster of the other thirteen manuscripts than the PCA plot (more about this in Section 8).

6. Parts of Speech Analysis

We also did a cluster analysis and PCA for groups of parts of speech, to find out in which part of their vocabulary the different manuscripts in the five samples diverged. This builds on the idea that scribal variation may be largely visible in function words, as opposed to variation in content words when a scribe adapts the content of the text and starts acting as a rewriter or even a new author (Van Dalen-Oskam and van Zundert 2007, 2008). We distinguished between three groups. In the group of *content words* we gathered all nouns (including proper names) and all verbal forms functioning as main verbs. The group of *function words* consisted of all conjunctions, all pronouns, all prepositions and all verbal forms functioning as modal or auxiliary verbs. The third group consisted of adverbs, numerals and adjectives, not wanting to lump them together as of yet with the content words to which they are usually seen to belong. We called this group the 'grey words'.

What stands out the most in the cluster analysis is that for each of the different PoS groups the manuscripts cluster in a different way, even in each of the samples. Content words as well as 'grey' words tend to be highly similar between most manuscripts, with the exception of the *Jewish War* sample, but they form different clusters. Most remarkable is manuscript A, which stands apart in the cluster observation of the Eva content words (also a bit for the Eva 'grey' words) and for the Debora 'grey' words. There is a remarkably high similarity of function words in the Judith sample for several manuscripts, amongst which manuscript I. That the content words in the *Jewish War* sample show no near 100 % similarity for some manuscripts is exceptional. In these observations, the manuscripts that were noted as 'different' while reading the texts, E and I, do certainly not stand out, with only this one exception, manuscript E in the content words in the Transjordania sample.

In the principal component analysis, manuscript E is a relative outlier in the content words in the Eva sample. In the Debora and the Judith sample E stands apart on the second component for all three PoS groups. In the New Testament sample, E does not stand out. In the *Jewish War* sample, however, E opposes the other manuscripts on the first component in all PoS groups. Manuscript I is separate on the first component from

the other manuscripts in the Debora and Judith sample (and stops after the Old Testament). Other special manuscripts: F is a relative outlier for content words in the Eva sample, and is opposite to the other manuscripts on the first component in the New Testament sample on all three PoS groups. In the *Jewish war* sample, F is slightly separate in the large cluster together with A and C, but not extremely. H opposes the large cluster in the Judith sample on the second component only for the function words. G stands out on the first component for content words in the *Jewish War* sample. And N is separate from the large cluster on the second component for all three PoS groups in the New Testament sample.

Principal component analysis thus again shows more manuscripts to be outliers than the cluster observations did, adding to manuscripts E and I a special role for N and F in the New Testament sample, and G for the *Jewish war* sample. It should again be remarked, that there seems to be a trend in the course of the texts, from which the New Testament sample is an exception.

7. What about Judith and the Scribe of Manuscript I?

The measurements clearly show that manuscript I stands out in several samples. In section 3, close reading of all samples showed that the uniqueness of manuscript I is most clearly visible to the human eye in the Judith episode, which was confirmed in the cluster analysis and the principal component analysis in section 5. This certainly leads to more research questions. This particular manuscript is known as manuscript 720-22 of the Royal Library of Belgium in Brussels. Deschamps and Mulder (2000) describe it as a convolute manuscript consisting of two parts. The first part contains a Dutch version of *The travels of John Mandeville* and a short other text and can be dated between 1425-1450. It is written by one scribe, who writes a *littera cursiva*. The second part contains Jacob van Maerlant's *Scolastica*. Deschamps and Mulder date this part around 1450. The text is written in *littera cursiva* and *littera hybrida* and may have been written, they state, by more than one scribe. Deschamps and Mulder do not provide a localisation of the manuscript, although some information about the provenance is given (the manuscript belonged to the Augustinian Canon Regulars at St-Maartensdal in Leuven). Our impression is that the dialect of the second part of manuscript is most likely to reflect a background in the Southern Low Countries, but we have not done systematic research to confirm this. For now, this (still rather broad) localisation is used. The Judith episode does not show a change in script and therefore is probably written by one and the same scribe.

So what was it about Judith that may have inspired this scribe (or the scribe of the examplar (etc.)) to present a version that is significantly different from all the other surviving manuscripts? We might even state that this unique version of the Judith sample, having unique lines and different wording and rhymes at several places compared to the other manuscripts, is not to be considered as a copy of Jacob van Maerlant's text but as a new text, and thus that the scribe responsible for this unique text can be better characterized as being an *author* in stead of a copyist. We will not be able to give a definitive answer to the question “What about Judith?”, but we will give some pointers for further research.

Judith is the well-known heroine of the apocryphal biblical Book of Judith who single-handedly killed Holofernes, the leader of the army that besieged her home town. Her story was popular in the Middle Ages and can be found in several separate texts from the early to the late Middle Ages in many languages (Lähnemann 2006). She also is one of the women described in *Neun guten Frauen* ('Nine good women'), a list of important women drawn up parallel to the better-known *Neuf preux* ('Nine Worthies') text. *Neun guten Frauen* presents three sets of three women, from respectively a heathen, Jewish, and Christian background. The chosen women were Lucretia, Veturia, and Virginia, Esther, Jael, and Judith, and Saint Helen, Saint Birgitta, and Saint Elisabeth (Van Anrooij 1997: 92-93). The text originates from Southern Germany near the end of the fifteenth century and was popular in the Low Countries from the fifteenth to the early seventeenth century (Van Anrooij 1997: 104). The probable date of the text is a couple of decades later than the probable date of manuscript I, so the *Neun guten Frauen* may not be the direct inspiration for the new version of the Judith story in this *Scolastica* manuscript. The existence and popularity of the text, however, does show that the story of Judith was very much alive in the fifteenth century - of which manuscript I is then also a witness. If the scribe/author of manuscript I was influenced by some sort of list of famous women could for example be checked by analyzing the episodes concerning the other Jewish *guten Frauen* Esther and Jael, which were not included in our samples. Also, historical and archival research may help us find out more about the cultural environments in which Judith played a special role. However, since the provenance of the manuscript is not completely clear, it will be difficult to find a good point of entry into this kind of research. Progress in this area seems to depend on a lucky find that may help us further.

8. Judith without Manuscript I

There is one more thing we would like to explore before we draw up an answer to our main question. Which visualizations do we get when we exclude manuscript I from our measurements on the Judith sample? Since manuscript I deviates so much, it may be that any differences between the other manuscripts are blotted out. For this reason we redid the cluster observations and the principal component analysis on the Judith samples excluding manuscript I. Removing manuscript I from the dataset brings the amount of lemmas down from 642, being significantly more than the amount in the other four samples, to 238, which is significantly lower than for the other four samples. The measurements result in Fig. 8 and Fig. 9.

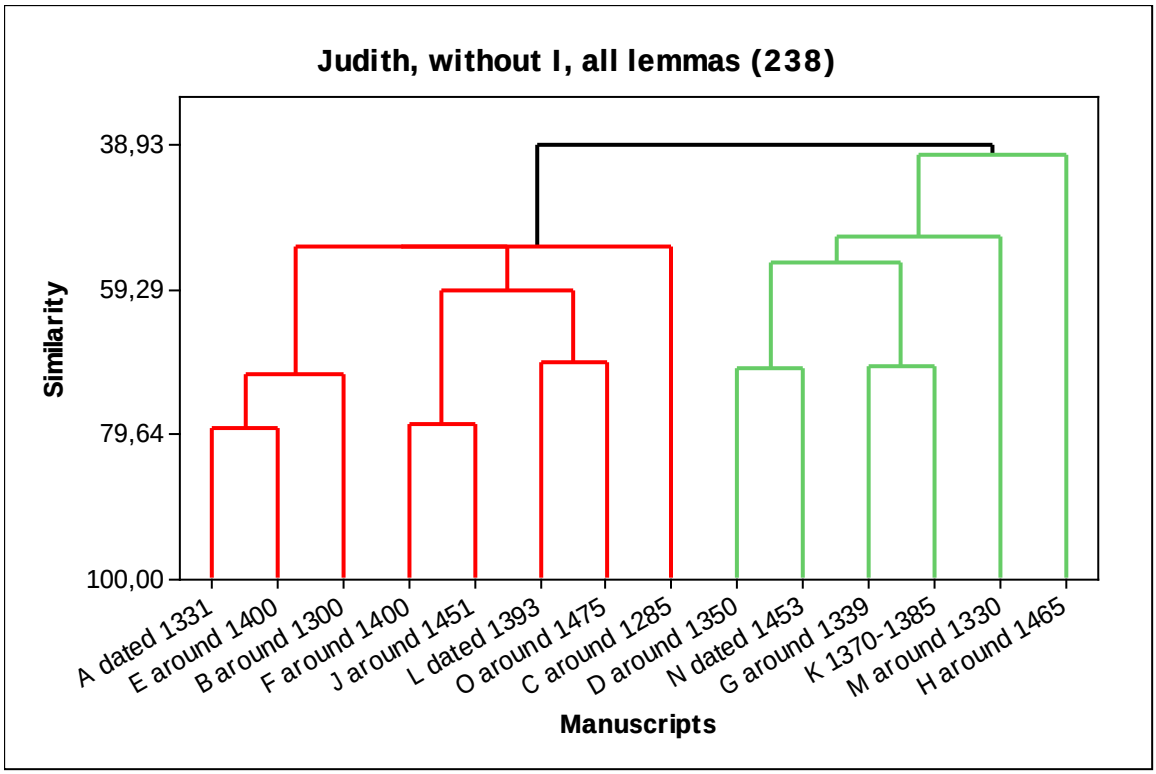

Fig. 8 Cluster analysis based on all lemmas of the Judith episodes, excluding manuscript I

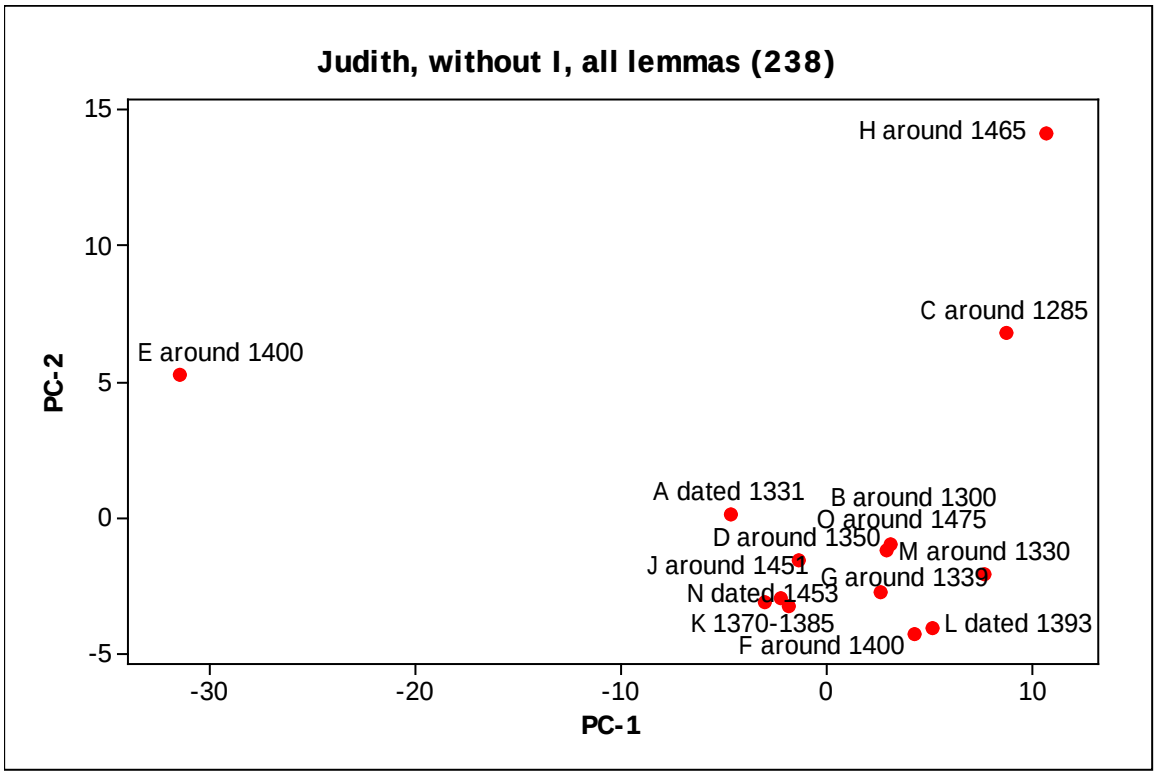

Fig. 9 Principal component analysis based on all lemmas of the Judith episodes, excluding manuscript I

The graph visualizing the cluster analysis (Fig. 8) now shows two large clusters of manuscripts, in stead of manuscript I forming a cluster on its own. The principal component analysis as visualized in Fig. 9 now shows manuscript E as an outlier on the first component, and, surprisingly, manuscript H is back (cf. section 4) as an outlier on the second component, with the oldest manuscript C placed between H and the group with the other mansucripts. Without further research it is not exactly clear what is happening here, but this seems to suggest that manuscript H may not only be significantly differing in spelling of some high-frequency words (cf. section 4 and 5), but also to some extent in content. This exploration thus also leads to new information compared to our predictions in section 3. It leads to the wish to have a closer look not only at manuscript E but also at manuscript H in comparison to the other manuscripts for the Judith sample. The small experiment with Burrows's Delta Procedure reported on in Section 6 also showed Manuscript H to have the third highest difference. There is much more to be said about the application of Delta on the lemmas in the Scolastica samples, but this may have to be the topic of a follow-up article. Analysing the Worksheets on which the Delta calculations are made, could be a very good way to zoom in on what is happening in these manuscript episodes on lemma level to find out more about the ways of the scribes.

9. Conclusions

Our main question was: What new insights in how scribes dealt with the text they were copying can we gain by applying stylometric methods used in non-traditional authorship attribution? The selected methods were cluster analysis and principal component analysis. We wanted to take the content level of the copies into account and for that reason focused our attention on the measurements done on the lemmatized text samples. Testing the measuments on the 'raw' Middle Dutch texts showed that this approach was indeed worth while.

The cluster analysis and principal component analysis on the lemmatized and PoS-tagged samples highlighted the fact that no copies of the same sample in the different manuscripts were exactly the same. Most manuscripts, however, clustered together and only occasional outliers were found. It was significant that the clustering was clearly different for all different episodes, which strongly suggested the necessity to analyse and explain the activity of scribes on an episodic level. Exploring the differences in usage of parts of speech led to the same conclusion, which suggests that each scribe may have had different approaches to different parts of speech as to the freedom they allowed themselves, although some trends were clearly visible, showing the least variation in the content words and most in the function words.

Several methodological next steps need to be considered. The first would be to establish the level of influence that language change and dialect differences may have had on the graphed differences between the manuscripts. They do not seem highly interfering, but even so when these can be filtered out, we may have the best view on the changes that actually relate to a different approach of the content. An explanation of these differences then needs traditional research methods into the historical and cultural context of each of the manuscripts and/or scribes to find out what the possible reasons for their approach were. The Delta Worksheets, in effect, could well have the most useful data set

for this further analysis, with information about standard deviation and z-scores on lemma level.

The application of stylometric methods has clearly resulted in a lot more information than the traditional approach by reading the texts. Cluster analysis and principal component analysis have resulted in clear pointers to those manuscripts and those episodes which are the most interesting for further research. The Judith episode in manuscript I was already significant after close reading, as was manuscript E in the *Jewish War* sample. But several other episodes and manuscripts turned up in the results of the non-traditional methods. It seemed that both manuscript E and I had an increasing level of difference compared to the other manuscripts throughout the manuscripts, not being noticable between the other manuscripts in the Eva episode, but becoming more of an outlier in every next episode. It was also becoming clear that the New Testament episode was an exception to this trend. This needs to be examined as well, since it could show that the most daring scribe (of manuscript E in this case) was especially cautious in the New Testament - and other scribes, not standing out in the other episodes, here took their opportunity. It will be very useful to find out in which ways the approach to the New Testament differed to that of the other episodes and what the backgrounds of the manuscripts standing out in that episode where.

The big advantage of this top-down way of zooming in on the copies of the same text is that the next steps are certain to yield useful information - we know in which manuscripts and episodes something seems to happen and we have some very clear ideas as to which new episodes would be worthwhile to analyze. This is a great extra stimulus to spend time on the laborious work of transcribing more episodes from the still unedited manuscripts, some of which will not have been read for several centuries.

Acknowledgements
Many thanks to the anonymous reviewers, who made me go back to the visualization of the PCA loadings and to an experiment with Delta which I had decided not to include in earlier versions of the article. The Delta calculations were done by Meindert Kroese (Huygens ING). Willem Kuiper (Huygens ING) helped with the transcription and collation of the manuscript samples. Others who have been very helpful with comments on earlier versions of this contribution are Peter Boot, Mike Kestemont, and Joris van Zundert.

Notes

1. The transcriptions were as close to what the manuscript reads as possible (a diplomatic transcription). Abbreviations were solved according to full forms, when available, or when not available to the most common form in Middle Dutch dialects.

2. The same mistake seems to be made by Michiel van der Borch, the illuminator of manuscript M, who depicted not a canopy above Holofernes's bed, but a sparrow hawk (Van Dalen-Oskam and Meuwese 2003).

References

Brefeld, J. (1994). *A guidebook for the Jerusalem pilgrimage in the late Middle Ages: a case for computer-aided textual criticism.* Hilversum: Verloren (Middeleeuwse Studies en Bronnen XL).

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17: 267-287.

Cerquiglini, B. (1989). *Éloge de la variante. Histoire critique de la philologie* Paris: Éditions du Seuil.

Cerquiglini, B. (1999). *In praise of the variant. A critical history of philology*. Baltimore: Johns Hopkins University Press.

Craig, H. a**nd Kinney, F.** (2009). *Shakespeare, computers, and the mystery of authorship.* Cambridge University Press.

Deschamps, J. and Mulder, H. (2000). *Inventaris van de Middelnederlandse handschriften van de Koninklijke Bibliotheek van België (voorlopige uitgave), Derde aflevering.* Brussel. http://opteron1.kbr.be/manus/BELGICA/deschamps_mulder/aflevering_3.pdf

Greco, G. L. and Schoemaker, P. (1993). Intertextuality and large corpora: a medievalist approach. *Computers and the humanities* 27: 349-355

Gysseling, M. (1983). *Rijmbijbel/tekst* In: *Corpus van Middelnederlandse teksten (tot en met het jaar 1300).* Uitgegeven door M. Gysseling m.m.v. en van woordindices voorzien door W. Pijnenburg. Reeks II: literaire handschriften, deel 3, Leiden.

Holmes, D. I. (1994). Authorship attribution. *Computers and the humanities* 28: 87-106.

Hoover, D. L. (2010). Authorial style. In Dan McIntyre & Beatrix Busse (Eds.). *Language and style. In honour of Mick Short.* New York etc.: Palgrave Macmillan, pp. 250-271.

Kestemont, M. and Van Dalen-Oskam, K. (2009). Predicting the Past: Memory-Based Copyist and Author Discrimination in Medieval Epics. *Proceedings of the twenty-first Benelux conference on artificial Intelligence (BNAIC 2009).* Eindhoven, pp. 121-128.

Lähnemann, H. (2006). *Hystoria Judith. Deutsche Judithdichtungen vom 12. bis zum 16. Jahrhundert*. Berlin, New York: Walter de Gruyter (Scrinium Friburgense, Bd 20)

Postma, A. (1991) Overzicht van Scolastica-handschriften. In J. van Moolenbroek en M. Mulder (ed.). *Scolastica willic ontbinden. Over de Rijmbijbel van Jacob van Maerlant*. Hilversum, p. 145. (Middeleeuwse Studies en Bronnen XXV)

Spencer, M. and Howe, C. J. (2001). Estimating distances between manuscripts based on copying errors. *Literary and Linguistic Computing* 16: 467-484.

Spencer, M. and Howe, C. J. (2002). How accurate were scribes? A mathematical model. *Literary and Linguistic Computing* 17: 311-322.

Spencer, M., Windram, H. F., Barbrook, A. C., Davidson, E. A., and Howe, C. J. (2006). Phylogenetic analysis of written traditions. In *Phylogenetic methods and the prehistory of languages*, ed. P. Forster & C. Renfrew, pp. 67-74. Cambridge.

Thaisen, J. (in press). A Probabilistic Analysis of a Middle English Text. In Brent Nelson and Melissa Terras (eds). *Digitizing Medieval and Early Modern Material Culture*. (New Technologies in Medieval and Renaissance Studies). Tempe: Arizona Center for Medieval and Renaissance Studies.

Van Anrooij, W. (1997). *Helden van weleer. De Negen Besten in de Nederlanden (1300-1700).* Amsterdam: Amsterdam University Press.

Van Dalen-Oskam, K. and M. Meuwese (2003). 'Een vreemde vogel in de Meermanno-Rijmbijbel' *Millennium* 17: 13-25.

Van Dalen-Oskam, K. and van Zundert, J. (2008). The Quest for Uniqueness: Author and Copyist Distinction in Middle Dutch Arthurian Romances based on Computer-assisted Lexicon Analysis. In Mooijaart, M., van der Wal, M. (eds.) *Yesterday's words: contemporary, current and future lexicography.* [Proceedings of the Third International Conference on Historical Lexicography and Lexicology (ICHLL), 21-23 June 2006, Leiden]. Cambridge: Cambridge Scholars Publishing, pp. 292-304.

Van Dalen-Oskam, K. and van Zundert, J. (2007). Delta for Middle Dutch – Author and Copyist Distinction in *Walewein*. *Literary and Linguistic Computing* 22: 345-362.

Van der Vlist, E. and Rudy, K. M. (2010). Het geschreven boek in Nederland tot omstreeks 1400. Continuïteit en emancipatie. In: Kopij en druk *revisited*. *Jaarboek voor Nederlandse boekgeschiedenis* 17: 15-52.

Windram, H. F., Shaw, P., Robinson, P., Howe, C. J. (2008). Dante's *Monarchia* as a test case for the use of phylogenetic methods in stemmatic analysis. *Literary and Linguistic Computing* 23: 443-63.

***Minitab*:** http://www.minitab.com

***The Intelligent Archive*:**
http://www.newcastle.edu.au/school/hss/research/groups/cllc/intelligent-archive.html

Appendix

Descriptions of the results of the measurements

1 The data

In the article, the main number of tokens and lemmas were given for the Judith episode. It may be useful to present them for all five samples. Some numbers will also give an idea of the differences between the measurement of lemmas and text. Table 4 therefore presents the total amounts of lemmas, types, and tokens in each of the episodes, for all of the copies of the episode together. Two types of tokens are given: 'tokens raw', based on the Middle Dutch text itself, in which enclitic forms such as *datsi* ('that they') are counted as one, and 'tokens tagged' in which the token count is based on the tagged text and in which *datsi* ('that they') is taken to consist of two tokens.

	Eva	Debora	Judith	Maria's	Transjordania
Lemmas	540	526	642	490	578
Types	1963	1775	1949	1673	1624
Tokens raw	16890	16189	17449	16293	14098
Tokens tagged	18174	17084	18510	17529	15323

Table 4. Numbers for each episode of lemmas (different head words / dictionary enrties), types (different word forms), and tokens (different occurrences of word forms). Tokens raw are the tokens in the untagged Middle Dutch samples. Tokens tagged are tokens derived from the amount of tags for lemma and part of speech. Manuscript O is not included in the calculations for Eva and Debora. I and O do not have the New Testament episode and I, O and H do not have the *Jewish war* episode.

2 Cluster Observations based on Middle Dutch text (types)

For all: Linkage Method: Ward, Distance Measure: Euclidian, showing 2 clusters. Applied with 990 highest frequency words, 500, 250, 125, 75, and 50 highest frequency words.

Eva. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. (1963 types.) Highest similarity of manuscripts B and E, then G and N, and I and M. Two groups: ABCEIJLM versus DFGHKN.

Debora. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. (1775 types.) Highest similarity of manuscripts C and K, then D and M, and E and N. Two groups: ABCGHK versus DEFIJLMN.

Judith. All extant fifteen manuscripts are represented. (1949 types.) Highest similarity of manuscripts B and E, then K and N. Two groups: ADFIKMNO versus BCEGHJL. (Fig. 1)

New Testament. New Testament is not included in manuscripts I and O. (1673 types.) Highest similarity of manuscripts G and M, and A and F. Two groups: ACDEFGJM versus BHKLN.

Jewish War. Josephus' *De bello judaico* is not included in manuscripts H, I, and O. (1624 types.) Highest similarity of manuscripts B and D. Two groups: ABDFGMN versus CEJKL.

3 Principal Component Analysis based on Middle Dutch text (types)

For all: Correlation, 2 components. Applied with all words, the 990 highest frequency words, and 500, 250, 125, 75, and 50 highest frequency words. We will again give a description of the situation per sample, for the sake of clarity repeating some of the factual information also given above.

Eva. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. (1963 types.) The manuscripts roughly cluster together, except for manuscript H which is an outlier on the second component for all frequency ranges. In the lower measurements (125 and less) H gets closer to the group, and J is the outlier on the second component.

Debora. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. (1775 types.) Manuscript H is an outlier on the second component in the higher frequency ranges and on the first component in the lower frequencies. Manuscript I is a relative outlier on the first component in frequencies 500 and 250, and on the second component in frequency 125. In lower frequency ranges (75 and 50) manuscript I is on the fringe on the first component.

Judith. All extant fifteen manuscripts are represented. (1949 types.) Manuscript I is an outlier on the first component. Manuscript E and H differ from the main cluster on the second component, in different directions. (Fig. 2)

New Testament. New Testament is not included in manuscripts I and O. (1673 types.) Manuscripts are widely scattered. In lower frequency measures (125 and less) H separates from the group on the first component. N separates from the rest for 75 and 50 HFW on the second component.

Jewish War. Josephus' *De bello judaico* is not included in manuscripts H, I, and O. (1624 types.) Manuscript E is an outlier on the first component. At 500 and 250 HFW, G is slightly separate from the big cluster. For 125 and less, C is an outlier on the second component.

4 Cluster Observations based on lemmas

For all: Linkage Method: Ward, Distance Measure: Euclidian, showing 2 clusters. Applied with all words, the 250 highest frequency words, and 125, 75, and 50 highest frequency words.

Eva. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. (540 different lemmas.) Highest similarity of manuscripts B and D, then F and L, and K and N. Two groups: ABDEJKMN versus CFGHIL Minor shift in placement of manuscript I in different frequency measurements.

Debora. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. (526 lemmas.) Highest similarity of manuscripts F and L, then B and D, and A and M. Two groups: ABCDEFGHJKLMN versus I. E is the most extreme in the large group. Relatively minor shift in placement of manuscripts H and J in different frequency measurements.

Judith. All extant fifteen manuscripts are represented. (642 lemmas.) Highest similarity of manuscripts F and L, then B and D. Two groups: ABCDEFGHJKLMN versus I, I being significantly more different than in the Debora sample. E is the most extreme in the large group. Minor shifts in placement of clusters FL, KN, and G and O in different frequency measurements. (Fig. 3)

New Testament. New Testament is not included in manuscripts I and O. (490 lemmas.) Highest similarity of manuscripts A and L, C and M, and D and J. Two groups: AFL versus BCDEGHJKMN. In the large group minor shifts in placement of J, K, and N in different frequency measurements.

Jewish War. Josephus' *De bello judaico* is not included in manuscripts H, I, and O. (578 lemmas.) Highest similarity of manuscripts D and J, or D and M. F and L, and K and N. Two groups: ABCDFGHKMN versus E, E being at a significant distance, comparable to I versus the other manuscripts in the Judith sample. A lot of shifts in placement in the large group in different frequency measurements.

5 Principal Component Analysis based on lemmas

For all: Correlation, 2 components. Applied with all words, the 250 highest frequency words, and 125, 75, and 50 highest frequency words. We will again give a description of the situation per sample, for the sake of clarity repeating some of the factual information also given above.

Eva. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. (540 lemmas.) The manuscripts occur without any very clear pattern in the scatterplot for all measured frequency sets. There is some clustering of manuscripts ABDJKN, in the lower measurements (75 and 50 HFW) joined by L and M. Relative outliers are EFGH, but there is a lot of shifting around (except for E).

Debora. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. (526 lemmas.) A graph for all words could not be made, because the texts in some areas were too much alike. Manuscript I and E are outliers, all the other manuscript form a dense cluster. I and E diverge on the first resp. the second component.

Judith. All extant fifteen manuscripts are represented. (642 lemmas.) Manuscript I and E are outliers, all the other manuscript form a dense cluster. I and E diverge on the first resp. the second component. The differences between I and the other manuscripts are more extreme than in the Debora sample. (Fig. 4)

New Testament. New Testament is not included in manuscripts I and O. (490 lemmas.) Manuscript N and F are outliers, in different components. All other manuscripts form a relatively spacious cluster. When applying PCA on the 75 and 50 highest frequency words, A and L are leaving the large group and approaching F.

Jewish War. Josephus' *De bello judaico* is not included in manuscripts H, I, and O. (578 lemmas.) When taking into account all words and the 250 highest frequency words, manuscripts E and G are outliers, on component 1 resp. 2. The other manuscripts show a relatively spacious cluster. For 125 HFW, G moves a bit closer to the group while manuscript C (the oldest of all manuscripts) starts moving away from the group. In 75 and 50 HWF, G is part of the group and C is a clear outlier (and on another component than E).

6 Cluster Observations based on parts of speech

For all: Linkage Method: Ward, Distance Measure: Euclidian, showing 2 clusters. Applied with all words, the 250 highest frequency words, and 125, 75, and 50 highest frequency words. No shifts at all in the clusters in the different frequency ranges, except for the content words in the Transjordania sample.

Eva. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. **Content words** (287 lemmas in total): near 100 % similarity for B and D, K and M and for I and L, J and N, and for C and G, which then closely relate to F. Two groups: A versus BCDEFGHIJKLMN. A is a significant outlier versus all the other manuscripts. **Function words** (83 lemmas): differences between the manuscripts are much larger than for the content words. Most similary between I and M, A and L, and B and C. Two groups: ADEGJKLN versus BCFHIM. **'Grey' words** (170): near 100 % similarity between F and L, E and H, and N and I. Other combinations than the content words, but the same level of nearness. Two groups: AFKLM versus BCDEGHIJN. In the first group, A is a significant outlier.

Debora. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. **Content words** (315): Two groups: ABCDGHIJKLM versus EFN. Most of the manuscripts in the first group have a near 100 % similarity. In the second group, E and F are almost similar, while N is a clear outlier. **Function words** (69): the manuscripts are clearly less similar than in the content words. Two groups: ABCDEFGJLMN versus HKI. **'Grey' words** (142): About half of the manuscript have a near 100% similarity: C and D, E, and J, F and G, H and M, L and N. Two groups: A versus BCDEFGHIJKLMN.

Judith. All extant fifteen manuscripts are represented. **Content words** (381): near 100 % similarity between E and O, F and Gm B and D, H, M and L, and I and J. Two groups, rather far apart: ABCDEFGHLMO versus IJKN. **Function words** (93): similarities never nearing 100 %. Closest are F and J, and A and E. Two groups: ABCEFJLO versus DGHIKMN. **'Grey' words** (168): near 100 % similarity between I, K and O, F, L and N, and D and M. Two groups: AIJKO + CFHLN (with a rather big difference between them) versus BDEGM.

New Testament. New Testament is not included in manuscripts I and O. **Content words** (273): very high similarity (nearing 100 %) for B and C, and H and M. Two groups: ABCEKL + DN (rather far apart from each other) versus FGHJM. **Function words** (78): no close similarities such as for the content words. Closest are K and L, and E and G. Two groups: ABCDFKL versus EGHJMN. **'Grey' words** (139): several almost 100 % similar manuscripts, F and L, B and G, J and M. Two groups: AFHL versus BCDEGJKMN.

Jewish War. Josephus' *De bello judaico* is not included in manuscripts H, I, and O. **Content words** (300): no near 100 % similar manuscripts. Two groups ABCDFGJKLMN versus E. When measuring different frequency ranges, manuscripts move around in the large group. **Function words** (92): no very near similarities. Two groups: ABDJMN versus CEFGKL. **'Grey' words** (186): many near 100 % similarities, such as A and K, B and D, C and G, and F and J. Two groups: ABCDEGK versus FJLMN.

7 Principal Component Analysis based on parts of speech

For all: Correlation, 2 components. Applied with all words, the 250 highest frequency words, and 125, 75, and 50 highest frequency words. We will again give a description of the situation per sample, for the sake of clarity repeating some of the factual information also given above.

Eva. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. **Content words** (287 lemmas in total): the picture is not very clear, but manuscript E, F, an I are relative outliers, though they are never very far from the rest of the manuscripts and I sometimes is a clear part of the cluster. A does not stand out here, while in the cluster observations it did. **Function words** (83 lemmas): a very scattered pattern for all manuscripts. **'Grey' words** (170): no clear pattern, no clear outliers.

Debora. Manuscript O lacks some pages, resulting in too short a sample text to include in the measurements. **Content words** (315): Most manuscripts cluster in a big group. I is a clear outlier on the first component and E on the second component. In the scatterplot for 75 HFW J moves away from the big cluster, and in the plot for 50 HFW J is back in the herd while N has moved out a bit. **Function words** (69): I is a clear outlier on the first component and E on the second component, the rest is in one big cluster. **'Grey' words** (142): I stands out on the first component and E on the second. J stands slightly aside from the cluster with the rest of the manuscripts.

Judith. All extant fifteen manuscripts are represented. **Content words** (381): I is a clear outlier on the first component and E on the second component, the rest is in one big cluster. **Function words** (93): I is a clear outlier on the first component and E on the second component. H is also separate from the rest on the second component. **'Grey' words** (168): I is a clear outlier on the first component and E on the second component, the rest is in one big cluster.

New Testament. New Testament is not included in manuscripts I and O. **Content words** (273): F is a clear outlier on the first component and N on the second component, the rest is in a relatively spacious cluster. **Function words** (78): same as content words. **'Grey' words** (139): same as content words and function words.

Jewish War. Josephus's *De bello judaico* is not included in manuscripts H, I, and O. **Content words** (300): E is a clear outlier on the first component and G on the second component, the rest is in a relatively spacious cluster in which ACF are slightly apart from the rest. **Function words** (92): E is a clear outlier on the first component and C on the second. The rest is in a relatively spacious cluster from which G is slightly apart in the plot for all 92 lemmas. **'Grey' words** (186): same as content words.