What Can Stylometry Learn From Its Application to Middle Dutch Literature?[1]

Mike Kestemont, University of Antwerp

Abstract: 'Stylometry' is a rapidly evolving subdiscipline of computational philology, focusing on the quantitative study of (literary) style. In recent years, stylometry has had some interesting applications in the field of Middle Dutch studies, especially in the domain of authorship attribution and scribe identification. In this paper I will highlight and discuss the main insights gained in these contributions. It will appear that these studies raise some challenging issues that deserve stylometry's careful consideration in coming years.

Keywords: Middle Dutch literature, Stylometry, Authorship attribution, Stylistics, Digital Humanities

> *When dealing with literary statistics it is best to err on the side of caution.*[2]
>
> *Saskia Murk Jansen*

Introduction

'Stylometry' refers to the quantitative study of (literary) style, nowadays often by means of computation.[3] Being a cross-disciplinary subfield within the so-called 'Digital humanities', stylometry currently attracts the interest of scholars from widely diverging disciplines such as linguistics, literary criticism, artificial intelligence and information science.[4] Stylometry tends to present itself as one of the more exact or empirical branches of modern philology, especially with its typical use of large, digital text corpora and statistics. Stylometric research is generally characterized by an emphasis on meta-data: the text itself is often of less interest to researchers than information *about* the text. Many studies, for instance, envisage the automatic inference of meta-data – such as authorship or date of composition – from a text's stylistic properties.

Although stylometry has been used in various problem settings, authorship attribution has undeniably been its most popular application in recent years.[5] The automatic classification of texts in terms of authorship is of course a fascinating issue, especially in forensic contexts where one would like to verify the [47] authorship of a suicide note – possibly faked by a murderer – or single out a bomb letter's author from a number of suspects. Automatic plagiarism detection is another widely studied issue, sometimes closely related to style-based authorship verification.[6] Although such high-profile applications are of course easier to 'sell' to the larger audience, most of the seminal work in stylometry has actually been done in the field of (historical) literary criticism. A good deal of our literary heritage has survived only fragmentarily (e.g. damaged manuscripts) or has only been published anonymously or pseudonymously. For these texts, an objective methodology to automatically verify their authorship or attribute them to known authors would be most welcome. Nevertheless,

stylometric studies remain somewhat rare and disparate in medieval philology and do not seem to have entered the mainstream literature yet.[7]

The field of Middle Dutch literary studies (ca. 1200-1550) is quite rich in cases of unresolved authorship. Famous examples include the identity of the compiler of the so-called *Lancelot compilation* – perhaps Lodewijk van Velthem, see Figure 1 – or the delineation of the œuvre by the 'Bible Translator of 1360' (possibly Petrus Naghel).[8] Therefore, it does not come as a surprise that stylometry has recently known a number of interesting case studies in the Low Countries' literary history. In this paper I intend to offer a concise overview of these publications, while, furthermore, using this opportunity to introduce the broader audience of this journal to the vibrant and rapidly evolving field of stylometry.[9] This paper's emphasis, however, will lie on a number of challenging issues and insights that resulted from the studies discussed. It will be demonstrated that these are quite relevant to the international stylometric community as well and deserve the field's careful attention in coming years.

Seminal Work

A milestone in twentieth century authorship attribution research has been the pioneering study in 1964 by the statisticians Mosteller and Wallace into the *Federalist Papers*, a collection of pamphlets advocating the ratification of the American Constitution, published under the pseudonym of *Publius* (1787-1788).[10] Although authorship attribution had of course been widely practiced before, their study introduced a groundbreaking novelty to the field. Previously, authorship attribution had often remained a largely impressionistic affair, in which attributions were only vaguely argued, based on the 'manual' and subjective reading experience of scholarly readers.[11] Scholars often turned to hand-selected 'checklists' of stylistic features to characterize an author's style and used these lists to accept or reject someone's authorship for a given text. These lists usually contained a limited set of stylistic peculiarities that had struck the expert reader's eye, such as the use of conspicuous nouns or unusual syntactic constructions.
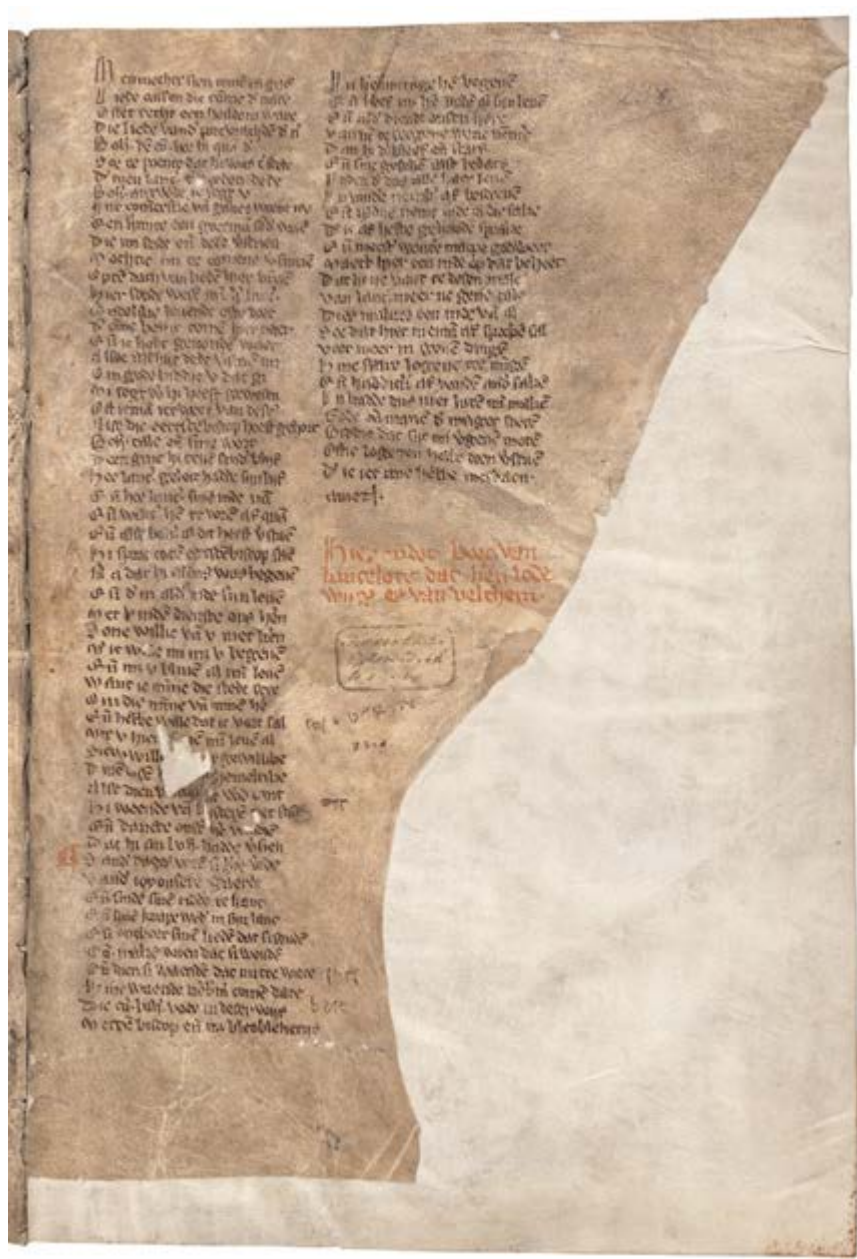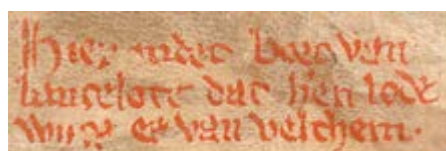
[48]

Figure 1: The last leaf of the Lancelot Compilation (The Hague, Koninklijke Bibliotheek, MS 129 A 10, fol. 238r). The mysterious colophon (in red) reads: 'Hier indet boec van lanselote dat heren lodewijcs es van velthem' ('Here ends the Book of Lancelot, which belongs to Sir Lodewijc van Velthem'). However, 'van' could also be understood as 'written by' or 'compiled by'. [49]

The use of such 'lists' often turned out to be problematic.[12] These lists tended, for instance, to be quite short, while it usually remained unclear why only this specific subset of features was included in the list and others were not. The main problem, however, was the low frequency of many of the items on the lists. Because usually stylistic peculiarities are already rare in an

œuvre itself and often even linked to a specific topic or genre, these features often did not scale well to other (possibly shorter) texts. Moreover, precisely because of their infrequent appearance, these features catch the human eye, which makes them extremely vulnerable to stylistic imitation and forgery. Mosteller and Wallace suggested to radically move away from these conspicuous, low frequency elements and focus on the exact opposite: a text's most frequent words, its function words.

Function words such as articles (*the*), prepositions (*under*) or pronouns (*she*) form a small and closed class of (typically short) common words, conveying only a very generic meaning.[13] Because the semantics of these function words is so pale, the same set of function words is extremely frequent in all texts. This makes these words very attractive for use in authorship attribution, not in the least because they are used by all authors writing in the same language and period and, thus, provide a statistically reliable base of comparison. Even more interesting is that these words are generally not related to a text's content and can therefore be used for attribution across different topics and genres. No matter what an author writes about (from music to politics to science), he or she will always need to use definite articles. A final advantage is that they are not usually under an author's conscious control and thus pretty robust to imitation.[14]

The earliest application of this revolutionary insight in Middle Dutch studies is to be found in an exceptionally early, but poorly recognized paper in *Dutch Crossing* (1988) by Saskia Murk Jansen.[15] In this study the author described an experiment concerning the authorship of the so-called mystic *Mengeldichten* ('Mixed poems'). Since not all these poems appear in the surviving manuscripts of Hadewijch's work (thirteenth century), a number of them are usually not attributed to the famous Brabantian beguine. In order to assess whether stylometry could add support to this thesis, Murk Jansen performed an experiment using *Principal components analysis* (PCA), a technique from multivariate statistics, nowadays commonly used for text clustering in authorship attribution. Explicitly referring to the work on the *Federalist Papers* by Mosteller and Wallace, Murk Jansen restricted her analysis to a small set of thirteen highly frequent Middle Dutch function words. Although her results offered neither a clear-cut confirmation nor a firm rejection [50] of existing hypotheses, the author seemed right in concluding 'that it would be interesting to pursue this line of analysis'.[16]

The Early Swallows

One swallow does not make a summer and Murk Jansen's experiment did not immediately inspire followers. In the 1980s and 1990s, however, a clear trend emerged in Middle Dutch studies, with several scholars increasingly exploring the possibilities of quantification in style-related research. Initially, only limited use was made of statistics in quantification and the computational aids that were (sometimes) involved in the end remained relatively modest. In 1983, for instance, Evert van den Berg completed his doctoral research about the development of versification in Middle Dutch epic literature.[17] In this study (and many subsequent articles), Van den Berg pursued a line of quantitative stylistic research, with an emphasis on the geographic and diachronic variation of features such as epithets and other formulaic language use. Van den Berg's work – however influential – did not rely on any advanced computational quantification but did occasionally include tests (e.g. Mann-Whitney) regarding statistical significance.

The same goes for Willem Kuiper's dissertation, published in 1989, about *Ferguut*, the Middle Dutch adaptation (ca. 1240) of the Old French Arthurian novel Fergus.[18] Kuiper argued that Ferguut had been written by two distinct authors, basing this on the difference between two parts of the text regarding the frequencies of various linguistic characteristics. He made extensive use of the well-known Chi-Square test to determine the significance of the observed differences. New in Kuiper's approach was that he partially relied on a computational processing of the text, for instance to automatically generate alphabetic word lists and concordances. A similar strategy was adopted in Bart Besamusca's investigations in 1991 of the so-called *Lancelot compilation*, a unique early fourteenth-century codex in which an anonymous 'compiler' left us an extensive Arthurian narrative, welded together from a collection of pre-existing Middle Dutch romances.[19] Besamusca, too, employed computer-generated word lists as an aid in his stylistic research into the identity of the mysterious compiler of this codex.

Computational Middle Dutch studies, however, only received their most important impetus in 1998, with the publication of the *Cd-rom Middelnederlands* ('CD-ROM Middle Dutch'), containing – apart from an entire digital dictionary of Middle Dutch – a near-exhaustive collection of the field's most prominent texts.[20] It is hard to overestimate the impact of this acclaimed *Cd-rom*, since it has been the most widely used resource in the field's recent history. The main advantage this text collection offered was that texts could not only be harvested from it in a machine-readable format, but also that all texts were fully searchable. The Cd-rom's advanced search options have been heavily exploited in authorship studies in the years following its publication. [51]

In 2001 Westgeest attributed the so-called Leyden *Lapidarium* fragments to the thirteenth-century poet Jacob van Maerlant, undoubtedly the most influential author in Middle Dutch literary history.[21] Exploiting the *Cd-rom*'s search possibilities, Westgeest was the first to use the disc as a sort of control corpus. He argued that stylistic features from the *Lapidarium* fragments were relatively typical of Maerlant's texts but infrequent in other works. Another well-known contribution appeared in 2002.[22] In this article, Joris Reynaert investigated the style of texts within the so-called 'Antwerp School', a coherent collection of ethical and historiographical texts that were all written in the first half of the fourteenth century by (one or more) poets from the city of Antwerp. Only one of the authors in this 'School' is known by name: Jan van Boendale, a learned shipping clerk. Reynaert argued that in fact many more, if not all works from this school should be attributed to Boendale. This 'School' of poets and texts might, in his eyes, well be a scholarly fiction, since it would in reality coincide with a single œuvre, namely Boendale's. In his research, Reynaert had turned to the *Cd-rom*, much in the same way as Westgeest had done. He argued that the Antwerp texts had a lot of stylistic features in common that they did not share with the other texts on the *Cd-rom*. Recently, in 2009, Marjolein Hogenbirk complemented Becamusca's study of the Lancelot compiler, adopting a methodology that was strongly reminiscent of Reynaert's way of working in 2002.[23]

Joining the Digital Humanities

It seems characteristic for the earliest period that researchers were enthusiastically exploring the possibilities of computational research but at the same time were still uncertain as to the methodologies that could be adopted. This situation changed in the first decade of the following century, when researchers from Middle Dutch studies were increasingly inspired by

international 'Digital humanities' research. The import of insights from abroad offered a remedy for one particular shortcoming of the earlier studies, namely the misinformation about recent advances in international stylometric research.[24]

Perhaps the earliest example of this development can be found in a brief but interesting book contribution from 2005 by Godfried Croenen, nowadays an important protagonist in computer-assisted medieval studies at the University of Liverpool.[25] In this paper he revisited the possibility of a dual authorship in the *Grimbergsche oorlog* ('The war of Grimbergen'), a fictionalized rhymed account of an early medieval feud between the lords of Grimbergen and the duke of Brabant. Croenen divided the text into samples of 500 verse lines. Subsequently, he turned to a statistic cluster analysis of these samples. Although his methodology is discussed only briefly, his results seemed to support the hypothesis that a second author took over from the poet who had initiated the text. Moreover, the cluster [52] analysis showed that the prologue could well be a later addition to the text by the second author.

One of the most important publications so far has been one in *Literary and linguistic computing* in 2007 by Karina van Dalen-Oskam and Joris van Zundert.[26] In this paper they investigated the *Roman van Walewein* ('Romance of Gawain'), an indigenous Middle Dutch romance about King Arthur's nephew, Walewein. With its positive and subtle depiction of the epic hero, *Walewein* occupies a unique position in the medieval European tradition and was recently fully translated into modern English.[27] The narrator of the text mentions that a certain Penninc had started the text but did not finish it: the *Walewein* (ca. 11,000 verse lines in total) would have been completed by a second author, *Pieter Vostaert*, adding 'about' 3,300 lines to Penninc's share, according to the narrator. Various scholars have been speculating as to the exact location in the text where Vostaert took over from Penninc, but at the time, conclusive evidence for one of the many existing hypotheses was still lacking. Van Dalen-Oskam and Van Zundert wanted to use quantitative authorship attribution to determine the exact spot of the authorial take-over in Walewein.

To this end, they applied 'Burrows's Delta' to the *Walewein*, a well-known stylometric measure for estimating the distance in authorial style between two samples of text. The influential metric was originally proposed by John Burrows, one of the founding fathers of modern stylometry.[28] Simply put, Delta compares two samples of texts in terms of the relative occurrences of a limited set of (e.g. fifty) highly frequent words. It outputs a dissimilarity score (hence, 'Delta') by statistically comparing the differences in these frequencies to the ones found in a large control corpus. A text of disputed signature can subsequently be attributed through calculating the Delta of this text to those written by a number of candidate authors. The disputed text will be attributed to the author whose work is most similar (viz. lowest Delta) to it. Empirical tests have shown that Delta, although conceptually simple and intuitive, often yields reliable results.

For the case of the *Walewein*, Van Dalen-Oskam and Van Zundert proposed a windowing procedure: they isolated the 'suspected' area where, roughly speaking, the takeover could have taken place – around line 8,000 – and divided it in consecutive samples ('windows') of equal size. Then, they extracted a writing sample from Penninc's beginning and one from Vostaert's completion (respectively well before and after the suspected area). Subsequently, they looped through the windows in the suspected area and calculated the Delta between on the one hand, each suspicious window, and on the other hand, the samples written by Penninc and Vostaert. They were hoping that the resulting list of values would show that at a specific point in the text, the windows would start to show a rapidly decreasing similarity to Penninc's share, while

displaying a fast increase in similarity to Vostaert's style. The location of this 'crossfade' would in that case be likely to coincide with the authorial takeover. [53]
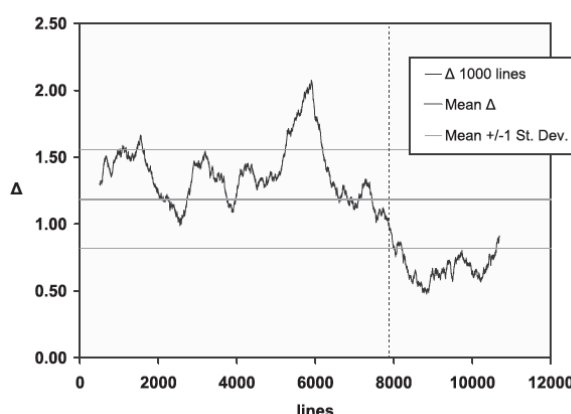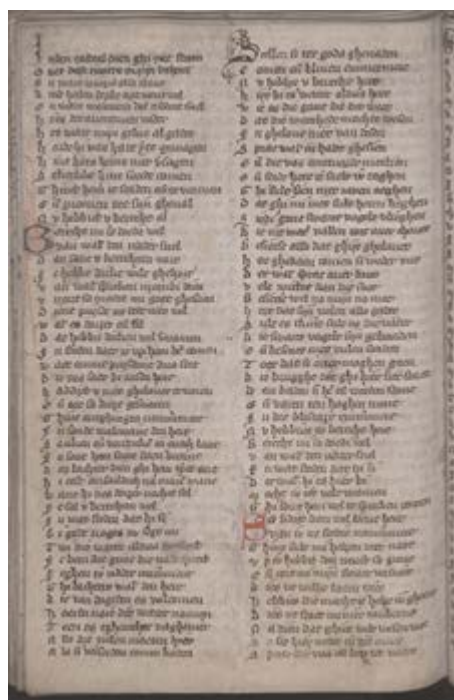
Figures 2a & 2b: 2a shows the upper-left part of fol. 152v in the Walewein-manuscript (Leiden, Koninklijke Bibliotheek, Ltk. 195) with the change between the two scribes: the two first lines on the verso side of this leaf are by the first scribe; the following by the second. 2b shows the result of an experiment by K. van Dalen-Oskam and J. Van Zundert ('Delta for Middle Dutch', p. 538, Fig. 19; courtesy of the authors), in which they were able to track the scribal takeover using common word frequencies (cf. the peak around line 5784).

The crossfade between Penninc and Vostaert, around line 8,000 was indeed clear from a number of their experiments. They made a strong case that Vostaert took over around line 7,880 of the text. However, their results were a school-book example of serendipity. Several of their graphs resulting from the windowing procedure showed a remarkably extreme 'peak' (or change in style) around line 6,000 of the text. Strangely enough, the text itself did not contain a narrative event or break in that region that could help explain this peak – which was in some experiments even more prominent than the change in authors. The two researchers quickly discerned that this peak coincided with the change in scribes that took place around that line. For their experiments, Van Dalen-Oskam and Van Zundert relied on an edition of the single complete manuscript of the *Walewein* that survives (Leiden, University Library, Ltk. 195). Paleographical research has shown that two scribes have had 'a hand' in this textual witness, the first one being responsible for the text up to line 5,783. The second scribe completed the copy [54] after that point and states at the end of the text that he finished his work in 1350. Van Dalen-Oskam and Van Zundert noted that Delta when applied to medieval texts was in fact not only sensitive to authorship but also, and perhaps even more, to the 'style' of the scribe who has copied the text. Scribal features therefore seem to interfere with authorial features in the computational analysis of medieval texts.

Hunting the Scribe

The *Walewein* case study uncovered crucial theoretical insights that would cause a turning point in computational scholarship. Apparently, modern attribution techniques are suited for the analysis of medieval authorship but have to take into account one major obstacle: *the scribe*. Before the invention of the printing press, texts were individually hand-copied by scribes. The 'autographs' of authors are moreover rarely extant. The absence of a standard language or spelling in the Middle Ages often caused a considerable variation between the parallel witnesses of texts, since scribes could freely adapt their exemplar to their own liking.[29] Scholars within the paradigm of 'Material Philology' have indeed been eager to stress that with each medieval copy of an old text, the opportunity arose to create an entirely new text.[30] For authorship studies, this evidently meant bad news, since with each subsequent copy, an original author's style risked fading even further.

For manuscript studies, on the other hand, the *Walewein* case immediately opened up exciting new research perspectives. If scribal features hinder authorship analyses, one could wonder just how stable these scribal features, e.g. spelling and dialect, really are. Would these, for instance, be stable enough to be able to perform scribe identification on medieval texts, much in the same way as people perform author identification? This would imply the intriguing possibility to recognize a scribe's hand, without ever having seen his or her handwriting. This could furthermore provoke a significant methodological advance in paleographic attribution studies. There too, the recognition of a scribe's hand can sometimes be a rather controversial affair, since the methodologies adopted are often 'expert-based' and, if not 'impressionistic', at least difficult to objectify.[31] That is especially the case when scholars attempt to recognize hands across different script types (e.g. *littera textualis* vs. *littera cursiva*). A computational methodology to assist in such difficult tasks would, of course, be a valuable aid.

This idea was first explored in a conference contribution from 2009 by Karina van Dalen-Oskam and myself.[32] The idea in itself was straightforward: given a large set of parallel copies of the very same authorial text, is it possible to computationally discriminate between the 'scribal style' of combinations of two copyists? In a series of experiments – the details of which cannot be discussed here – we tested this hypothesis on a diplomatic transcription of a series of parallel passages [55] from Jacob van Maerlant's *Rijmbijbel* ('Rhyming Bible') from fifteen parallel manuscripts. From our experiments it became clear that although not all combinations of scribes yielded equal results, a classification technique was generally fairly apt at distinguishing between scribes, using only linguistic features, such as spelling. The scribes in our corpus had apparently 'appropriated' the author's text to such an extent that we could tell them apart, with an accuracy that was significantly above chance.

More interestingly, we compared the performance of certain 'feature types' in this scribe identification task to their performance in a similar author identification task (Penninc vs. Vostaert). It turned out that the more 'superficial' textual features were (e.g. frequencies of single characters), the better they worked in scribe discrimination and the worse they worked in author discrimination. On the other hand, the 'deeper' the features used (the frequency of certain lemmas or part-of-speech tags), the worse they performed for scribe recognition and the better they worked for authorship attribution. The outcome of these experiments was intriguing on two levels: on the one hand our paper offered a proof-of-concept that – at least in this case study – it seemed possible to recognize Middle Dutch scribes. On the other hand, we demonstrated that this need not imply that this situation necessarily disqualifies authorship-

related research, since authorial characteristics – again, at least in this case study – seemed intact on another, perhaps 'deeper' level of texts. This dual outcome naturally fostered a second line of research in which Middle Dutch scholars started exploring under which circumstances, using which features, it would be possible to perform authorship attribution in Middle Dutch literature.

Coming to Terms With the Author

The inauguration of Herman Brinkman as a professor at the University of Amsterdam in 2009 has been significant for the development of stylometric research. This appointment had been made possible through the generous support of Brinkman's host institution, the Huygens Institute (The Hague). The institute is not mentioned as an aside, for the Huygens Institute has been positioning itself in recent years as an important protagonist in the Low Countries' 'Digital humanities' initiatives. (In that respect, it does not come as a surprise that both Van Dalen-Oskam and Van Zundert are also active at the Huygens.) Brinkman, who until then had not yet really ventured into the adventure of 'Digital humanities', surprised the audience with an excellent presentation entitled: *Als de nachtuil. Auteurschap en overlevering van middeleeuwse teksten* ('Like the night owl. Authorship and survival of medieval texts'). In his talk, Brinkman showed that state-of-the-art techniques from stylometry (e.g. hierarchical clustering) were in fact able to support traditional research in a large number of cases of unresolved authorship. [56]

Nevertheless, Brinkman too stressed that authorship attribution for medieval texts is not without problems. Research in medieval authorship attribution first of all shares a number of practical limitations with its present-day counterpart. If one is, for instance, to reliably attribute a text of disputed origin, both this text and the example data for the candidate authors need to be large enough. In addition, Brinkman stressed, medieval scholarship needs to account for a large number of other problems that are unique to medieval philology – the aforementioned scribal problems are the best example of this. In Brinkman's view, authorship attribution in medieval texts currently has to meet certain restrictions, if it is to be methodologically sound. Brinkman proposed 'ten commandments' for authorship attribution in Middle Dutch studies (paraphrased in the box below). For the authorial comparison of texts, commandment 5, 6 & 7, for instance, advocated a unity of genre and date and place of composition. According to Brinkman, this unity had to eliminate the disturbing influence of factors such as discourse, diachronic language change and dialect. Only if one would adhere to these principles, according to Brinkman, an investigation into medieval authorship attribution would be fully methodologically sound.

Brinkman's Ten Commandments[33]

1: Comparisons of authorial style based on word frequency analysis should be restricted to texts surviving in a copy by the same scribe (unity of scribe);

2: Authorial style analyses should avoid words that are dependent or indicative of specific content, genre, discourse or authorial perspective;

3: Writing samples need to be long enough to represent authorial usage (at least 2000 words);

4: One should only compare writing samples of a uniform length;

5, 6, 7: One should keep the factors 'genre', 'date of composition' and 'region of composition' stable across writing samples (unity of genre, date and region);

8: One needs to ensure a minimal distance between the original author's exemplar and the manuscript copy used;

9: In order to assess the differences between two texts, it is necessary to include a *tertium comparationis*;

10: A text that is to be analyzed in comparison to other texts is best divided into parts of equal length in order to check the stability of the outcome of the test.

Brinkman in his first commandment expressed the need for a unity of scribe in an authorship attribution experiment: if we compare the authorial style in two Middle Dutch texts, we could in principle only reliably compare them, when they survive in a copy that has been produced by one and the same scribe. Only in that way, we are able to isolate the impact of scribes on these texts. Already during his presentation, however, Brinkman indicated that his commandments are fairly restrictive. Note that if we would always fully adhere to them, we are actually unable to perform an authorial analysis on the bulk of the surviving Middle Dutch literature, since only a handful of texts have survived in a copy by the same scribe. In that respect, we could for instance never establish whether two of our most famous (and recently translated) texts, *Karel ende Elegast* and *Van den vos Reynaerde* were in fact written by one and the same author, since no copies of the texts by the same scribe survive.[34] It is clear that – although Brinkman's commandments will form an important touchstone for years to come – research is needed into ways of easing the criteria imposed by them.

Ongoing Research

In my current research project, I investigate one way of nuancing Brinkman's methodological criteria: *rhyme words*. It is a remarkable feature of Middle Dutch literature that a substantial share of it is rhymed, until well into the fourteenth century. Although rhymed couplets (aabbcc...) have been a prominent feature in many European medieval literatures, this stylistic device has occupied an extremely stable position in the Low Countries. Middle Dutch studies have long acknowledged that these rhyme words are an exceptional research object, not in the least because they are a relatively stable element in medieval text transmission, pretty robust to scribal corruption. Because rhyme words were often used to structure large, epic texts on the level of stylistics, scribes generally refrained from changing a text's rhyme words. It is indeed cumbersome to try and change a text's rhyme words, without having to rewrite a large portion of that text. Scribes would sometimes alter the spelling of rhyme words, but generally speaking, they seem to have left the underlying rhyme word (c.q. its lemma) intact. Interestingly,

restricting stylometric analyses to rhyme words might therefore eliminate scribal interference in medieval rhymed texts, thus relaxing the criterion of Brinkman's first commandment.

In my research, I investigate whether these rhyme words can be used to attribute for Middle Dutch epic texts. In particular, I explore the idea whether highly frequent rhyme words can serve as a surrogate for highly frequent normal words [58] (i.e. function words). Since a language's rhyme word vocabulary is limited and epic texts could be quite large, it has often been noted that poets repeatedly made use of the same 'stop-gaps': a nifty set of formulaic rhyme words that could be used over and over again in a variety of contexts. In a recent publication – joint work with Walter Daelemans and Dominiek Sandra – we have indeed shown that high-frequent rhyme words can safely be used for authorship attribution since they share a lot of advantages with the function words in modern texts: the same set of rhyme words is relatively context-independent and therefore frequent throughout most texts.[35] Moreover, it has been demonstrated that it is indeed possible to 'recognize' authors based on a stylistic profile that takes into account a poet's usage of this set of highly frequent rhyme words. For the authors of the *Spiegel historiael*, a thirteenth century vernacular adaptation of the Latin world chronicle *Speculum historiale*, it has indeed been shown that techniques from stylometry are reliably able to discriminate between the rhyming styles of Jacob van Maerlant and Filip Utenbroeke, two authors who contributed to the *Spiegel*.[36]

In another case study, I have investigated the rhyming style of Lodewijk van Velthem, a third author who added a substantial share to the *Spiegel historiael* in the early fourteenth century.[37] Although Velthem's style too could be easily modelled using computational techniques, it appeared that this was not true for all of his additions to the *Spiegel*. Some parts of his contribution were in fact so deviant in style from the rest of his œuvre, that they aroused suspicion. The largest fluctuation in style was observed in his famous account of the Battle of the Golden Spurs (Courtrai, 11 July 1302), a military encounter between the king of France and an untrained militia of Flemish townsmen, nowadays commemorated on the annual holiday of the Flemish community in Belgium. The deviance in style in Velthem's account turned out to be similar to other passages in his œuvre of which we know that the author did not write them himself but 'silently borrowed' them from other sources. Follow-up research showed that there is indeed ample reason to believe that Velthem – known to be a literary copycat – did not write this account himself but superficially reworked a pre-existing source text. To our knowledge, Velthem's detailed account held the earliest and most extensive Middle Dutch source of information on this well-known event. Stylometric research, however, showed that his account could in fact be even older.

Highly frequent rhyme words as such prove to be an interesting stylistic category for applied authorship attribution in historic literature. Apparently, a medieval author's use of such rhyme words contains a 'stylistic fingerprint' that is both constant and individual, so that we are able to model an author's 'stylome'. Nevertheless, it has been noted that the stability of one author's use of highly frequent rhyme words in his œuvre cannot be exaggerated. In one particular experiment we have looked into a single author's entire œuvre, namely Jacob van Maerlant's.[38] When applying a clustering technique (Correspondence Analysis) to the fifty most frequent rhyme words in Maerlant's entire œuvre, the results [59] showed that this text collection displayed a remarkable three-fold division. Although our analysis was restricted to a set of content-neutral stop-gaps, the Correspondence Analysis revealed three outspoken clusters: Maerlant's chivalric epics formed a first cluster, his historiographic works formed a second group, while his texts in more didactic genres clustered in a third group. Naturally, this

raises the question of how strongly genre-related features in texts in fact interfere with authorial style markers. Apparently, this question has not only been raised in Middle Dutch studies, but also elsewhere. There is a remarkable lack of research into an author's stability of style across different genres.[39] The little research that has been done on this topic (e.g. cross-topic attribution) clearly indicates that the international community currently underestimates the complexity of this issue. It should be stressed that for now there is in fact no evidence that Brinkman's fifth commandment – the unity of genre – would not be true for all stylistic attribution studies.

Conclusion: What Can Stylometry Learn From Its Application to Middle Dutch Literature?

Stylometry's potential for literary studies is enormous: the idea that one day we might have objective methodologies to automatically infer, for example, a text's authorship, is fascinating. At the same time, one should be careful not to forget that much work still needs to be done. Benchmark corpora to test existing methods generally remain under-exploited, so that it still remains unclear, for instance, how long an anonymous writing sample should in fact be in order to be able to attribute it to a known author.[40] Some interesting directions for future work have been proposed in Middle Dutch studies.[41] The issue of scribal interventions is perhaps the most remarkable one, since its effect could well be related to editorial interventions we find in other periods. To give but one example: since much empirical research into writing style is nowadays carried out on newspaper articles, 'there is a general worry with newspapers, that the texts of the authors are often changed by editor(s)'.[42] Techniques for recognizing scribal distortions thus might well be interesting for research into such present-day editorial interventions too. From a more literary perspective, the genre issue which was discussed at the end of this paper seems relevant for the international community too. The severe lack of cross-genre authorship attribution research makes it difficult to assess the real accuracy of contemporary attribution methods: even if techniques have been shown to perform well within a single genre, this certainly does not prove that they would work well outside the comfort zone of a single text variety. This raises interesting questions about an author's autonomy in writing. [60]

Acknowledgements

Bibliography

Allen, John R., 'On the Authenticity of the Baligant Episode in the Chanson de Roland', in Computers in the Humanities, ed. by John L. Mitchel (Edinburgh: Edinburgh University Press, 1974), pp. 65-72.

Argamon, Shlomo, 'Interpreting Burrows's Delta: Geometric and Probabilistic Foundations', in Literary and Linguistic Computing 23:2 (2008), pp. 131-47.

van den Berg, Evert, Middelnederlandse versbouw en syntaxis. Ontwikkelingen in de versifikatie van verhalende poëzie ca. 1200 – ca. 1400 (Utrecht: H&S, 1983).

Besamusca, Bart (ed.), Lanceloet. De Middelnederlandse vertaling van de Lancelot en prose overgeleverd in de Lancelotcompilatie. Pars 2 (vs. 5531-10740). Met een inleidende studie over de vertaaltechniek. (Assen: Van Gorcum, 1991).

Besamusca, Bart, The Book of Lancelot. The Middle Dutch Lancelot Compilation and the Medieval Tradition of Narrative Cycles (Cambridge: D.S. Brewer, 2003).

Binongo, José Nilo G., 'Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution', in Chance 16:2 (2003), pp. 9-17.

Bouwman, André and Besamusca, Bart (eds) Of Reynaert the Fox (Amsterdam: Amsterdam University Press, 2009).

Burrows, John, ''Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship', in Literary and linguistic computing 17:3 (2002), pp. 267-87.

Burgers, Jan W.J., De rijmkroniek van Holland en zijn auteurs. Historiografie in Holland door de Anonymus (1280-1282) en de grafelijke klerk Melis Stoke (begin veertiende eeuw) (Hilversum: Historische Vereniging Holland, 1999).

Cd-rom Middelnederlands. Woordenboek en teksten (The Hague etc.: Sdu, 1998).

Claassens, Geert H.M., 'Bible Translator of 1360 (fl. mid-14th c.)', in Medieval Germany: An Encyclopaedia, ed. by John M. Jeep (New York: Garland, 2001), pp. 405-07.

Croenen, Godfried, 'Het dubbele auteurschap van de Grimbergsche oorlog', in Maar er is meer: avontuurlijk lezen in de epiek van de Lage Landen. Studies voor Jozef D. Janssens, ed. by Remco Sleiderink, Veerle Uyttersprot and Bart Besamusca (Leuven: Davidsfonds, 2005), pp. 131-51.

van Dalen-Oskam, Karina and Van Zundert, Joris, 'Delta for Middle Dutch – Author and Copyist Distinction in Walewein', in Literary and Linguistic Computing 22:3 (2007), pp. 345-62.

van Dalen-Oskam, Karina, 'Digital Resources and Computational Methods', in Zeitschrift für deutsche Philologie, 130: Sonderheft (2011), pp. 375-390. [61]

Dimpel, Friedrich M., 'Der Verlust der 'Eneas'-Handschrift als Fiktion. Eine computergestützte, textstatistische Untersuchung', in Amsterdamer Beiträge zur älteren Germanistik 61:1 (2006), pp. 87-102.

van Driel, Joost, Prikkeling der zinnen. De stilistische diversiteit van de Middelnederlandse epische poëzie (Zutphen: Walburg Pers, 2009).

García, Antonio M. and Martín, Javier C., 'Function Words in Authorship Attribution Studies', in Literary and Linguistic Computing 22:1 (2007), pp. 49-66.

Hinskens, Frans and Van Dalen-Oskam, Karina, 'Kwantitatieve benaderingen in het taalen letterkundig onderzoek. Een ruwe schets', in Tijdschrift voor Nederlandse Taalen Letterkunde 123:1 (2007), pp. 1-21.

Hogenbirk, Marjolein, 'Is hij het? Lodewijk van Velthem en de compilator', in De boeken van Velthem. Auteur, oeuvre en overlevering, ed. by Bart Besamusca, Remco Sleiderink and Geert Warnart (Hilversum: Verloren, 2009), pp. 47-72.

Holmes, David I., 'The Evolution of Stylometry in Humanities scholarship', in Literary and Linguistic Computing 13:3 (1998), pp. 111-17.

Johnson, David and Claassens, Geert (eds and trans.), Dutch Romances. 1: Roman van Walewein (Cambridge: Brewer, 2000).

Kestemont, Mike, 'De meesters van de Spiegel. Auteursonderscheiding op basis van het frequente rijmwoord in het aandeel van Utenbroeke en Maerlant in de Spiegel historiael', in Tijdschrift voor Nederlandse Taalen Letterkunde 127:2 (2011), pp. 97-120. Kestemont, Mike, Daelemans, Walter and De Pauw, Guy, 'Weigh Your Words – Memory-Based Lemmatization for Middle Dutch', in Literary and Linguistic Computing 25:3 (2010), pp. 287-301.

Kestemont, Mike and Van Dalen-Oskam, Karina, 'Predicting the Past: Memory-Based Copyist and Author Discrimination in Medieval Epics', in Proceedings of the Twenty-First Benelux Conference on Artificial Intelligence (BNAIC 2009), ed. by Toon Calders, Karl Tuyls and Mykola Pechenizkyi (Eindhoven: Benelux Association for Artificial Intelligence, 2009), pp. 121-28. Kestemont, Mike, 'Velthem et al. A Stylometric Analysis of the Rhyme Words in the Account of the Battle of the Golden Spurs in the Fifth Part of the Spiegel historiael', in Queeste 17:1 (2010), pp. 1-34.

Kestemont, Mike, Daelemans, Walter and Sandra, Dominiek, 'Robust Rhymes? The Stability of Authorial Style in Medieval Narratives', in Journal of Quantitative Linguistics, 19:1 (2012), pp. 1-23.

Kuiper, Willem (ed.), Die riddere metten witten scilde. Oorsprong, overlevering en auteurschap van de Middelnederlandse Ferguut, gevolgd door een diplomatische editie en een diplomatisch glossarium (Amsterdam: Schipbouwer & Brinkman, 1989).

Kwakkel, Erik, 'Die dietsche boeke die ons toebehoeren': de kartuizers van Herne en de produktie van Middelnederlandse handschriften in de regio Brussel (1350-1400) (Leuven: Peeters, 2002).

Love, Harold, Attributing Authorship: An Introduction (Cambridge: Cambridge University Press, 2002).

Mosteller, Frederick and Wallace, David L., Inference and Disputed Authorship: The Federalist (Reading, MA: Addison-Wesley, 1964).

Murk Jansen, Saskia, 'Hadewijch's Mengeldichten: an experiment in statistical method', in Dutch Crossing 35 (1988), pp. 26-39. [62]

Nichols, Stephen, 'Why Material Philology? Some Thoughts', in Zeitschrift für deutsche Philologie 116 (1997), Beiheft [= special issue: Philologie als Textwissenschaft. Alte und neue Horizonte, ed. by Helmut Tervooren and Horst Wenzel], pp. 10-30.

Rem, Margit, 'Middelnederlands met de computer', in Madoc 19:1 (2005), pp. 24-32.

Reynaert, Joris, 'De 10e Brief van Hadewijch en het 41ste der Limburgse Sermoenen', in Leuvense bijdragen 63:2 (1974), pp. 137-49.

Reynaert, Joris, 'Boendale of 'Antwerpse school'? Over het auteurschap van 'Melibeus' en 'Dietsche doctrinale'', in Al t'Antwerpen in die stad. Jan van Boendale en de literaire cultuur van zijn tijd, ed. by Wim Van Anrooij and others (Amsterdam: Prometheus, 2002), pp. 127-57 and pp. 177-82.

Schreibman, Susan, Siemens, Ray and Unsworth, John (eds), A Companion to Digital Humanities (Oxford: Blackwell, 2004).

Stamatatos, Efstathios, Fakotakis, Nikos and Kokkinakis, George, 'Automatic text categorization in terms of genre and author', in Computational Linguistics 26:4 (2000), pp. 471-95.

Stamatatos, Efstathios, 'A Survey of Modern Authorship Attribution Methods', in Journal of the American Society for Information Science and Technology 60:3 (2009), pp. 538-56.

Stein, Benno, Lipka, Nedim and Prettenhoffer, Peter, 'Intrinsic plagiarism analysis', in Language Resources and Evaluation 45:1 (2011), pp. 63-82.

Westgeest, Hans, 'De Leidse lapidariumfragmenten: delen van Maerlants 'cortten lapydarys'?', in Queeste 8:1 (2001), pp. 1-26.

Notes

1. The author gratefully acknowledges the support of the Research Foundation of Flanders (FWO) at the University of Antwerp. He also would like to thank Frank Willaert and Walter Daelemans as well as the anonymous reviewers of this journal for their valuable comments on earlier versions of this paper.

2. Saskia Murk Jansen, 'Hadewijch's Mengeldichten: an experiment in statistical method', Dutch Crossing 35 (1988), pp. 26-39.

3. Usually, only writing style is considered in stylometry. For an excellent overview see David I. Holmes, 'The Evolution of Stylometry in Humanities scholarship', in Literary and Linguistic Computing 13:3 (1998), pp. 111-17.

4. For an accessible introduction see A Companion to Digital Humanities, ed. by Susan Schreibman, Ray Siemens and John Unsworth (Oxford: Blackwell, 2004).

5. A recent overview can be found in: Efstathios Stamatatos, 'A Survey of Modern Authorship Attribution Methods', in Journal of the American Society for Information Science and Technology 60:3 (2009), pp. 538-56.

6. See for instance: Benno Stein, Nedim Lipka and Peter Prettenhoffer, 'Intrinsic Plagiarism Analysis', in Language Resources and Evaluation 45:1 (2011), pp. 63-82.

7. Disparate yet interesting examples concerning medieval authorship include: Antonio M. García and Javier C. Martín, 'Function Words in Authorship Attribution Studies', in Literary and Linguistic Computing 22:1 (2007), pp. 49-66 for Old English; Friedrich M. Dimpel, 'Der Verlust der 'Eneas'-Handschrift als Fiktion. Eine computergestützte, textstatistische Untersuchung', in Amsterdamer Beiträge zur älteren Germanistik 61:1 [63] (2006), pp. 87-102 for German; John R. Allen, 'On the Authenticity of the Baligant Episode in the Chanson de Roland', in Computers in the Humanities, ed. by John L. Mitchel (Edinburgh: Edinburgh University Press, 1974), pp. 65-72 for French.

8. An English discussion of the Lancelot compilation in Bart Besamusca, The Book of Lancelot. The Middle Dutch Lancelot Compilation and the Medieval Tradition of Narrative Cycles (Cambridge: D.S. Brewer, 2003). For Petrus Naghel, consult Geert H.M. Claassens, 'Bible translator of 1360 (fl. mid-14th c.)', in Medieval Germany: An Encyclopaedia, ed. by John M. Jeep (New York: Garland, 2001).

9. A very interesting recent Dutch-language overview of the matter (partially overlapping with the present one) is presented in Frans Hinskens and Karina van Dalen-Oskam, 'Kwantitatieve benaderingen in het taalen letterkundig onderzoek. Een ruwe schets', in Tijdschrift voor Nederlandse Taalen Letterkunde 123:1 (2007), pp. 1-21. Another recent state-of-the-art article on quantitative studies is Karina van Dalen-Oskam, 'Digital Resources and Computational Methods', Zeitschrift für deutsche Philologie, 130: Sonderheft (2011), pp. 375-390.

10. Frederick Mosteller and David L. Wallace, Inference and Disputed Authorship: The Federalist (Reading, MA: Addison-Wesley, 1964).

11. A general discussion of (non-)traditional authorship attribution in Harold Love, Attributing Authorship: An Introduction (Cambridge: Cambridge University Press, 2002).

12. A more detailed discussion and critique of the traditional method can be found in in Mike Kestemont, 'De meesters van de Spiegel. Auteursonderscheiding op basis van het frequente rijmwoord in het aandeel van Utenbroeke en Maerlant in de "Spiegel historiael"', in Tijdschrift voor Nederlandse Taalen Letterkunde 127:2 (2011), pp. 97-120.

13. An introduction to the use of function words in authorship attribution in José Nilo G. Binongo, 'Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution', in Chance 16:2 (2003), pp. 26-39.

14. Note, however, that not all function words are necessarily highly frequent (cf. notwithstanding).

15. Murk Jansen, Mengeldichten. A similar early quantitative study is offered by Joris Reynaert, 'De 10e Brief van Hadewijch en het 41ste der Limburgse Sermoenen', in Leuvense bijdragen 63:2 (1974), pp. 137-49.

16. Murk Jansen, Mengeldichten, p. 32.

17. Evert van den Berg, Middelnederlandse versbouw en syntaxis. Ontwikkelingen in de versifikatie van verhalende poëzie ca. 1200 – ca. 1400 (Utrecht: H&S, 1983). Note that Van den Berg's method has been continued in the first chapters of Joost van Driel, Prikkeling der zinnen. De stilistische diversiteit van de Middelnederlandse epische poëzie (Zutphen: Walburg Pers, 2009).

18. Willem Kuiper (ed.), Die riddere metten witten scilde. Oorsprong, overlevering en auteurschap van de Middelnederlandse 'Ferguut', gevolgd door een diplomatische editie en een diplomatisch glossarium (Amsterdam: Schiphouwer & Brinkman, 1989).

19. Lanceloet. De Middelnederlandse vertaling van de Lancelot en prose overgeleverd in de Lancelotcompilatie. Pars 2 (vs. 5531-10740). Met een inleidende studie over de vertaaltechniek, ed. by Bart Besamusca (Assen: Van Gorcum, 1991).

20. Cd-rom Middelnederlands. Woordenboek en teksten (The Hague: Sdu, 1998). [64]

21. Hans Westgeest, 'De Leidse lapidariumfragmenten: delen van Maerlants 'cortten lapydarys'?', in Queeste 8:1 (2001), pp. 1-26.

22. Joris Reynaert, 'Boendale of 'Antwerpse school'? Over het auteurschap van 'Melibeus' en 'Dietsche doctrinale'', in Al t'Antwerpen in die stad. Jan van Boendale en de literaire cultuur van zijn tijd, ed. by Wim van Anrooij and others (Amsterdam: Prometheus, 2002).

23. Marjolein Hogenbirk, 'Is hij het? Lodewijk van Velthem en de compilator', in De boeken van Velthem. Auteur, oeuvre en overlevering, eds Bart Besamusca, Remco Sleiderink and Geert Warnar (Hilversum: Verloren, 2009).

24. Compare Jan W.J. Burgers, De rijmkroniek van Holland en zijn auteurs. Historiografie in Holland door de Anonymus (1280-1282) en de grafelijke klerk Melis Stoke (begin veertiende eeuw) (Hilversum: Historische Vereniging Holland, 1999), p. 82 (especially note 83).

25. Godfried Croenen, 'Het dubbele auteurschap van de Grimbergsche oorlog', in Maar er is meer: avontuurlijk lezen in de epiek van de Lage Landen. Studies voor Jozef D. Janssens, ed. by Remco Sleiderink, Veerle Uyttersprot and Bart Besamusca (Leuven: Davidsfonds, 2005).

26. Karina van Dalen-Oskam and Joris van Zundert, 'Delta for Middle Dutch – Author and Copyist Distinction in Walewein', in Literary and Linguistic Computing 22:3 (2007), pp. 345-62.

27. Dutch Romances. 1: Roman van Walewein, ed. and trans by David Johnson and Geert Claassens (Cambridge: Brewer, 2000).

28. John Burrows, "Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship', in Literary and Linguistic Computing 17:3 (2002), 267-87. Important theoretical interpretation in Shlomo Argamon, 'Interpreting Burrows's Delta: Geometric and Probabilistic Foundations', in Literary and Linguistic Computing 23:2 (2008), pp. 131-47.

29. Cf. Mike Kestemont, Walter Daelemans and Guy de Pauw, 'Weigh Your Words – Memory-Based Lemmatization for Middle Dutch', in Literary and Linguistic Computing 25:3 (2010), pp. 287-301.

30. Consult Stephen Nichols, 'Why Material Philology? Some Thoughts', in Zeitschrift für deutsche Philologie 116 (1997), pp. 10-30.

31. This issue is discussed in Erik Kwakkel, 'Die dietsche boeke die ons toebehoeren': de kartuizers van Herne en de produktie van Middelnederlandse handschriften in de regio Brussel (1350-1400) (Leuven: Peeters, 2002), p. 7.

32. Mike Kestemont and Karina van Dalen-Oskam, 'Predicting the past: Memory-based copyist and author discrimination in medieval epics', in Proceedings of the Twenty-First Benelux Conference on Artificial Intelligence (BNAIC 2009), ed. by Toon Calders, Karl Tuyls and Mykola Pechenizkyi (Eindhoven: Benelux Association for Artificial Intelligence, 2009).

33. I would like to thank Herman Brinkman for proofreading my translation.

34. Both of these important texts have recently been translated to English: Of Reynaert the Fox, ed. by André Bouwman and Bart Besamusca (Amsterdam: Amsterdam University Press, 2009). The translation of Karel ende Elegast is to appear in Olifant.

35. Mike Kestemont, Walter Daelemans and Dominiek Sandra, 'Robust Rhymes? The Stability of Authorial Style in Medieval Narratives', in Journal of Quantitative Linguistics, 19:1 (2012), pp. 1-23.

36. Kestemont, 'De meesters', passim. [65]

37. Mike Kestemont, 'Velthem et al. A Stylometric Analysis of the Rhyme Words in the Account of the Battle of the Golden Spurs in the Fifth Part of the Spiegel historiael', in Queeste 17:1 (2010), pp. 1-34.

38. Kestemont, Daelemans and Sandra, 'Robust rhymes?'.

39. Stamatatos, 'A Survey', pp. 553-54.

40. Stamatatos, 'A Survey', p. 553.

41. Hans van Halteren and Margit Rem (Radboud University Nijmegen) are currently investigating a number of fascinating case studies in scribe recognition and authorship attribution. A number of highly recommendable publications are forthcoming. Earlier work by Margit Rem has focused on computational linguistics, although the authorship of Ferguut is discussed in Margit Rem, 'Middelnederlands met de computer', in Madoc 19:1 (2005), pp. 24-32.

42. Efstathios Stamatatos, Nikos Fakotakis and George Kokkinakis, 'Automatic text categorization in terms of genre and author', in Computational linguistics 26:4 (2000), p. 485.