# Layer on layer. 'Computational archaeology' in 15th-century Middle Dutch historiography

Stapel, R.J.

***published in***
LLC: the journal of digital scholarship in the humanities
2013

***DOI (link to publisher)***
10.1093/llc/fqs046

***document version***
Publisher's PDF, also known as Version of record

# Layer on layer. 'Computational archaeology' in 15th-century Middle Dutch historiography

**Rombert J. Stapel**
Fryske Akademy, Royal Netherlands Academy of Arts and Sciences/ Leiden University, Netherlands

## Abstract

The aim of this article was to distinguish different authorial layers within a 15th-century chronicle, a rare medieval autograph or author's copy (*Croniken van der Duytscher Oirden* or *Jüngere Hochmeisterchronik*), and to test specifically the validity of claims concerning the original composition of the text. A better apprehension of the creative process involved in composing the *Croniken* is essential for the interpretation and understanding of the purpose and intended audience of the text. Furthermore, it gives an insight into the historiographical activities in the 'peripheral' bailiwicks of the Teutonic Order. Computational techniques, in this case John Burrows' tried-and-tested *Delta* method, play an invaluable role in both the solution of these issues as well as in pointing in the direction of new enquiries. Here, the *Delta* method was used to create a walking window, only 2,000 words in length, across the entire chronicle. Despite the small sample size, chosen because in the present case, a finer granularity and precision in detecting the shifts in authorial styles was as important as reliability, *Delta* was able to pick out distinct parts of the chronicle, some as short as just >500 words.

**Correspondence:**
Rombert J. Stapel,
Institute for History, Leiden University, Doelensteeg 16, 2311 VL Leiden, Netherlands.
**Email:** r.j.stapel@outlook.com/ r.j.stapel@hum.leidenuniv.nl

## 1 Introduction

Medieval manuscripts are characterized by a great number of, what we could call, 'interfering' features. Scribes manually copied texts again and again, sometimes altering content and often altering orthography. Most original works have been lost, just as much of the copied material. To add even more difficulties, what we nowadays will easily refer to as 'original work' is much less clear-cut in the Middle Ages. Our modern notion of 'copyright' is virtually unknown to medieval men and women, and for a long time, the concept of *auctoritas* (author) was primarily used in referral to classic writers such as Aristotle and Augustine. Many medieval texts are thus written anonymously.

The situation could be characterized as chaotic by scholars used to relatively straightforward text corpora. Before you can begin your quest for a medieval author, you first have to find out, if even possible, what content is related to scribes and what can be attributed to the author. And just when you think you are making some progress, you find out that your 'author' has been merely compiling source texts, which he (or she) is copying word for word. A scholar addressing these texts should therefore meticulously peel off the different layers of the text. When confronted with these time-consuming scenarios, it might not sound that surprising that the number of studies involving the use of computational techniques and medieval texts is not that great. In recent years though, some progress has

Literary and Linguistic Computing, Vol. 28, No. 2, 2013. 

doi:10.1093/llc/fqs046 Advance Access published on 3 December 2012

been made (concerning Middle Dutch texts, e.g.: van Dalen-Oskam and van Zundert, 2007; Kestemont and van Dalen-Oskam, 2009; van Dalen-Oskam et al., 2010). Most of all, these studies show that it is possible—using computational techniques such as Burrows' *Delta* and Machine Learning—to overcome some of the difficulties in distinguishing, for instance, scribal and authorial layers within a single text.

## 2 *Croniken* and its Composition

This begs the question to what extent these techniques can be used in the examination of medieval texts and manuscripts; stripping the layers of authorial and scribal intervention. An excellent case-study is a manuscript in the collection of the Teutonic Order in Vienna. The Teutonic Order was a military order founded in the Holy Land in 1190 during the Third Crusade, following the example of the Hospitallers and the well-known Knights Templar. Later, the Teutonic Order shifted their focus increasingly to campaigns against the then heathen tribes in the Baltic region. Eventually this brought them into conflict with the Polish-Lithuanian kings. The manuscript in question (Deutschordenszentralarchiv Wien, Hs. 392) contains the oldest version of the so-called *Croniken van der Duytscher Oirden* or *Jüngere Hochmeisterchronik*: a 15th-century history of the Teutonic Order written in Middle Dutch (Stapel and Vollmann-Profe, 2010). It covers the supposed origin of the military order in the Old and New Testament, the crusades in the Holy Land, up to the disastrous Thirteen Years' War (1453–66), instigated by the Prussian cities and their ally the Polish-Lithuanian king against the Order, which felt as a betrayal to the chronicler.

Ever since the *Croniken van der Duytscher Oirden* has been studied in a scholarly context, there have been doubts about the initial composition (E.g.: Hirsch 1874, p. 42). The original Middle Dutch version of the chronicle begins with a long prologue, according to the text itself written by a certain bishop of Paderborn present at the Order's foundation in the 12th century. This assertion is usually discarded as a fabrication to create authority, as no Paderborn bishop was present in the Holy Land at that time (following Töppen 1853, pp. 64–65; Hirsch 1874, p. 24). The prologue is followed by a part that contains the lives and deeds of the Grand Masters of the Order, alternated by numerous privileges granted by popes and emperors. The *Croniken* ends with a history of the regional bailiwick of Utrecht (one of the thirteen 'provinces' of the Order in the Holy Roman Empire) and its Land Commanders. Especially that last part, sometimes dubbed 'bailiwick chronicle', is often put aside as a later addition to the chronicle.

It is essentially in the middle part that a unique characteristic of the manuscript is unveiled. Codicological, palaeographical, and philological evidence suggests that the person who committed the chronicle to paper was either responsible for the text's conception itself or worked in the direct vicinity of the author. A combination of those two is equally viable. In other words, the manuscript is what one could call an author's copy: an umbrella term for autographs and 'original' manuscripts such as an apograph (the first fair copy, produced by a (professional) scribe under close scrutiny of the author himself: Houthuys 2009, pp. 65–67). Apart from the negative argument that evidence of copyist interference (e.g. omissions, accidental repetitions, or the typical eye-skip) is completely absent, some alterations to the text seem content-related, directly linked to the use of the source texts as one would expect from an author. Watermark evidence shows that the Vienna manuscript is the oldest manuscript that contains the *Croniken*. Furthermore, all extant manuscripts of the *Croniken*, both in Middle Dutch and in German translations, stem from the same Vienna manuscript. This strengthens the claim that the manuscript is indeed an autograph or a manuscript sanctioned by or at least written in close proximity to the author himself.[1]

Author's copies are especially rare in a medieval context. A recent survey of Middle Dutch manuscripts mentions barely more than a 100 examples (Houthuys 2009). As a result, we have the unique opportunity to study a medieval text not affected by hindering interference by scribes, so common in texts from this era. However, as becomes clear

from the cautious tone earlier, it is not apparent that one person has authorial responsibility over the entire text of the *Croniken*. Material evidence for authorial interference is not that widespread throughout the text, and it cannot be excluded that certain parts were copied, dictated, and/or translated, and subsequently gathered. The scribe of the Vienna manuscript could well have united the role of author with that of what St. Bonaventure in his often-used classification of writers would have called a *compilator*. Compiling and copying word-for-word pieces of other one's works to create a new text was a perfectly viable working method in the Middle Ages and was in no way considered a copyright infringement.

An analysis of the authorial styles in such a text will be both intriguing and complex. However, a better appreciation of the original composition of the *Croniken*, its genesis, and the potential existence of predating historiographical sources will be necessary for understanding the scope and function of the historiographical activities of the Teutonic Order in the Late Middle Ages, especially in their provinces in the Holy Roman Empire. As the use of the historiographical genre, often stressing continuity, is typically associated to periods of transformation, this has immediate implications for a broader historical perspective on the Teutonic Order.

## 3 Text Corpus and Preparation

The person who physically wrote the Vienna manuscript of the *Croniken* left more than just this single manuscript. His hand is also responsible for twenty-five Middle Dutch land charters written for the Teutonic Order in Utrecht between 1479 and 1491, as well as a few Latin land charters written for various clients between 1489 and 1509. A handwritten copy by him of a Middle Dutch *Sachsenspiegel* written around 1499–1500 for the Utrecht Land Commanders also survived. Chance has it that in one of the land charters, his name is recorded: Hendrik Gerardsz. van Vianen. Most likely, Hendrik acted as secretary to Land Commander Johan van Drongelen (1469–92), who we will return to later in the text. All in all, he left a Middle Dutch corpus of approximately 131,000 words (see Appendix A). Perhaps this is not as large as the corpora modern literary scholars work with, but it is still substantial in medieval terms. The corpus includes the work of a single person working as manuscript scribe; writer of land charters; and last but not least having an ambiguous role as author and/or scribe of a history of his order.

All texts by Hendrik van Vianen have been transcribed and encoded according to the P5 guidelines of the *Text Encoding Initiative*. Ultimately they will be made available in a scholarly edition. However, for use in stylometric analyses and authorship attribution techniques, these XML transcripts were stripped of all the tags using XSLT. (Roman) numerals, sentences in other languages (Latin in effect), deletions, and unreadable passages were removed from the original files. All capital letters were transformed into lowercase. No lemmatization was provided, mainly because the available time was limited. Lemmatization helps in reducing potential negative effects of, for instance, inflection and spelling variation (e.g.: Kestemont *et al.*, 2010). It would be interesting to see, in what way this will affect the outcome. Nonetheless, in this case, even non-lemmatized texts were able to produce satisfactory results. Perhaps the *Croniken* benefits from a tendency in Middle Dutch to move towards a more standardized spelling from the 15th century onwards. The end product is therefore a set of basic plain text files. These files are accompanied by two non-related medieval texts (from a comparable period and genre) that will be used as test samples. Both are different adaptations of the so-called *Gouds Kroniekje*. They were prepared in a similar fashion.[2]

## 4 Method

A tried-and-tested method was used to tell apart the different possible authorial layers within the *Croniken*, John Burrows' *Delta*: a 'measure of stylistic difference and guide to likely authorship' (Burrows 2002). It has produced a steady track record in authorship attribution, has been applied successfully to Middle Dutch texts in the past, and is easily accessible. Furthermore, early experiments on

the *Croniken* text using *Delta* produced some promising results that legitimized further research. An important factor was the availability of freely available and easy-to-use tools. The *Intelligent Archive* was used to create word frequency lists of all the text samples (Craig 2010). For the calculation of the *Delta* scores, the *Delta Spreadsheets for Microsoft Excel*, provided by David Hoover on his Website, proved to be helpful (Hoover 2009). Since then, however, the *Stylometry with R* script was released by Maciej Eder and Jan Rybicki. Preliminary tests show a substantial gain of speed over Hoover's *Delta Spreadsheets* (Eder and Rybicki, 2011).

One of the problems with authorship attribution in the *Croniken*, or better, authorship distinction, is the lack of comparative text material. For authorship attribution problems concerning modern texts, each possible candidate would be assigned at least one primary sample taken from a text not considered in the discussed authorship issue. However, in this instance, this is not a viable option. Of course, the *Sachsenspiegel* and land charters by Hendrik van Vianen are available, but—apart from the obvious genre difference—the *Sachsenspiegel* was merely copied by Hendrik and presumably contains a completely different authorial style. In turn, the land charters were written in a highly formulaic language. Perhaps for those reasons, both texts have not yet proved to be effective as markers of authorial styles in most parts of the *Croniken*.

Therefore, a workaround to this issue is proposed. Instead of using external primary samples only, some primary samples will be collected from certain well-chosen parts of the *Croniken* itself. The text of the *Croniken* was divided in 181 equally sized parts of 2,000 words, each with an overlap of 500 words (see Appendix A). Choosing a perhaps risky sample size of just 2,000 words (compare Eder 2010) means we are exploring the limits of the *Delta* method. However, for the purpose of this study, precision and fine granularity is needed as much as reliability. Larger sample sizes smoothen out significant changes in style, as indeed an additional experiment with samples of 4,000 words and 500-word overlap confirmed (see later in the text). Word frequency lists of these 181 parts provided the data for our *Delta* analysis and will be addressed as the 'secondary samples'. By computing these secondary samples against primary samples from the same text, one is in effect visualizing possible discontinuities in authorial styles in the *Croniken*. In a hypothetical situation where we have three different styles in one continuous text, this might produce the following results (see Fig. 1). The bars represent the location from where the primary samples were taken, whereas the lines represent the *z*-score of *Delta*. These scores are computed for every continuous section of the text. The lower the *z*-score of one of the three primary samples, the higher the probability that a section was correctly identified as belonging to that style sample. However, how low a *z*-score should be to be significant cannot always be easily defined. Typically, this will be established on an ad hoc basis. In this particular hypothetical situation, the graph clearly shows the exact locations where the changes in style occur. In real-life experiments, these transitions might be less abrupt, especially of course if overlapping secondary samples are used. It is also important to note that the second-lowest *z*-score should be significantly higher than the lowest *z*-score.

## 4.1 Testing: privileges

To test the validity of the proposed *Delta* method and workaround, it is necessary to run some experiments in a controlled environment. The *Croniken* contains numerous (summaries of) privileges of the Teutonic Order, issued by both popes and emperors of the Holy Roman Empire. They range in size from anywhere between merely a couple of 100 words to well more than a thousand and are interlaced in the text, each at their correct place in the chronology, and often grouped together. There is reason to assume that both the papal and imperial privileges represent aberrant authorial styles to the rest of the *Croniken*. Twice for instance, the *Croniken* mentions the absence of many more privileges since 'they' have not 'witnessed the bulls or authenticated transcripts': suggesting the use of privilege collections prepared by others. This indeed turns out to be the case. For the imperial privileges, a complete cartulary of the Teutonic

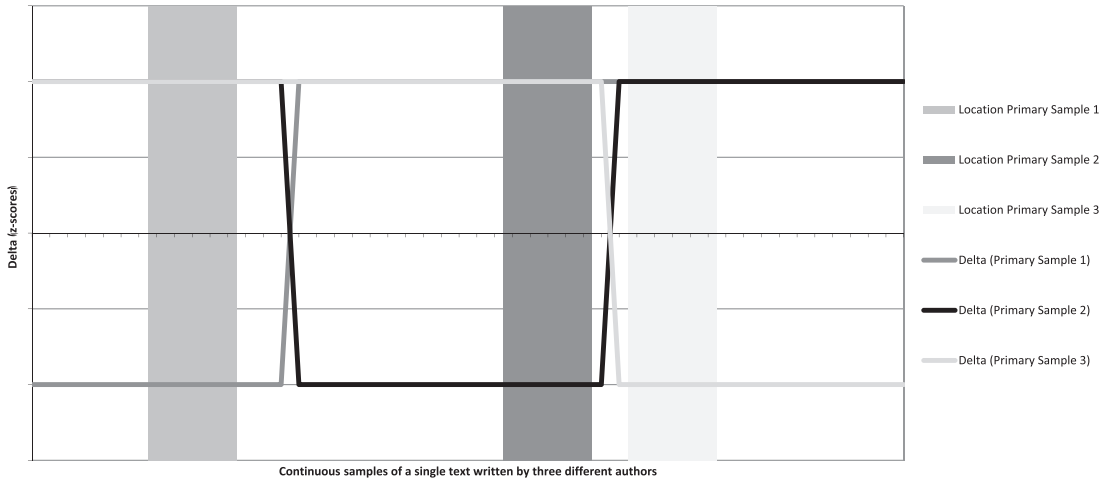


**Fig. 1** Hypothetical distribution of Delta scores and three authorial styles

Order issued at Heidelberg in 1428 has been integrated.[3] The original privileges were written in both Latin and High German and were translated (perhaps on-the-spot) into Middle Dutch, hardly altering the sentence structure of the privileges.[4]

Most likely, the papal privileges were also part of one collection. This collection was loosely based on a Latin cartulary that can be identified as one issued at Basel in 1434.[5] Not all privileges were included, some were added and many were severely shortened. Seldom were parts of the original Latin sentence structure left intact. Further, the summaries of papal privileges in the *Croniken* show a uniform visual appearance that reminds us of a similar—but not identical—existing Middle Dutch collection in the Utrecht bailiwick archive of the Teutonic Order.[6] It is unlikely that the writer of the *Croniken* manuscript, Hendrik van Vianen, used two completely opposite *modi operandi* for including both sets of privileges: one meticulously translating word-for-word from beginning to end, the other changing the order and reducing the content as he saw fit, leaving out and inserting privileges on the way. Therefore, it seems to be a fair assumption that the papal privileges were integrated word-for-word in the *Croniken* just as the imperial privileges, but from another existing Middle Dutch collection, now lost.

Thus, the imperial privileges were translated by one person, perhaps Hendrik van Vianen, who stayed true to the numerous original privileges dating from 1214 to 1415, written by an equally numerous number of people. The original editor of the papal privileges had much more room to introduce his own language preferences. It is likely that Hendrik van Vianen merely copied this collection. In both cases, however, one expects that the authorial style that can be construed from the samples of privileges is distinct from both the regular text of the *Croniken* and each other.

With this in mind, parts of both sets of privileges have been used to form two primary samples, each of three different sizes: 3,000, 6,000, and (circa) 8,000 words (see Appendix A). It should be noted that one of the drawbacks of using larger primary samples was that almost all privileges from the *Croniken* had to be included. Normally this would not be recommendable, but for the sake of examining the appropriate settings, we will accept the handicap. A third sample contained the complete text of the *Croniken* (thus including all the privileges but also all of the regular narrative of the chronicle). It acts as a baseline value. The goal was to determine whether it was possible to (1) single out privileges in general throughout the *Croniken* without wrongly identifying other areas in the text, and (2) to identify these parts of the chronicle correctly as containing either papal or imperial privileges: both representatives of different styles.

First, the *Delta* procedure was run against the 181 secondary samples for all three primary sample sizes

to determine what settings produced the best results. Twenty-four different sets of most frequent words (MFW) were used, ranging from 20 to 1,200. For each secondary sample, the largest class—that is, papal privileges, imperial privileges, and the remaining *Croniken* text—in terms of word size was marked. This resulted in 32 positive cases for the papal privileges, 16 for the imperial privileges, and 133 for the rest of the *Croniken*, totalling 181. A classification for a secondary sample was made based on the primary sample that produced the lowest *Delta* score. From here, we established the *true-positive rate* (TPR) and *false-positive rate* (FPR) of each set of variables (MFW; primary sample size). TPR is calculated by dividing the number of true-positive classifications by the total number of positives. If the TPR equals 1, all positive cases (e.g. aircrafts on a radar) are correctly identified as such. An identification in this case is expressed by the primary sample that produced the lowest *Delta* score. FPR is defined as the number of false-negative classifications divided by the total number of negatives. If the FPR is 1, all negatives (e.g. flocks of birds on the same radar) have been incorrectly classified as aircrafts in this example. A perfect classification occurs when the FPR is 0 and TPR is 1 (none of the flocks of birds is classified as aircrafts and all aircrafts are correctly identified). Generally, a pair of FPR and TPR values closest to (0,1) represents the optimal trade-off between many true-positive classifications and few false-positive errors. The distance to (0,1) can be measured for every set of variables. The following confusion matrix (*Croniken* baseline; 3,000-word primary samples; twenty MFW) will clarify this (Table 1).

Table 2 contains the distances for every combination of primary sample, primary sample size, and most frequently used words included in the calculations. Also included, for comparison, are the data produced by using longer 4,000-word secondary samples (500-word overlap).[7] The shortest distances are highlighted in each column. With a few notable exceptions, the optimum number of MFWs are all located in the higher regions. By adding up the distances of the three primary samples for each of the primary sample sizes in the final columns, the most appropriate MFW for that particular sample size is revealed. For the 3,000-word primary samples, this happens to be 650 MFW. Note that the performance of this primary sample size quickly worsens when more MFWs are added to the equation. When the primary sample size is increased to 6,000 or 8,000 words, the performance stabilizes in comparison with the 3,000-word sample size. Note as well that by increasing the secondary sample size to 4,000 words, there is a significant decrease in performance. The best results (an optimum between many correct and few incorrect classifications) can be expected by increasing the sample size to (approximately) 8,000 words and including the 1,000 MFW (Fig. 2).

We can conclude that the *Delta* method works rather well for this purpose and textual corpus.

**Table 1** Confusion matrix for the *Croniken* baseline sample, using 3,000-word primary samples and the twenty most frequently used words

| | **Positives** (*Croniken* is largest class) | **Negatives** (Papal or imperial privileges are largest class) |
|---|---|---|
| **Positive classifications** (*Croniken* produces lowest *Delta z*-score) | 131 | 18 |
| **Negative classifications** (Papal or imperial privileges produce lowest *Delta z*-score) | 2 | 30 |
| **Sum** | 133 | 48 |
| **FPR** | 0.375 | =18/48 |
| **TPR** | 0.985 | =131/133 |
| **Distance to FPR, TPR (0,1)** | 0.375 | =((FPR^2) + ((1 − TPR)^2))^0.5 |

**Table 2** Distance to FPR, TPR (0,1) for all combinations of variables

| Primary sample size | 3,000-word | | | 6,000-word | | | (Circa) 8,000-word | | | (Circa) 8,000-word, (4,000-word secondary samples) | | | Sum (*Croniken*, Papal, Imperial) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MFW | (Minimal) distance to FPR = 0, TPR = 1 | | | (Minimal) distance to FPR = 0, TPR = 1 | | | (Minimal) distance to FPR = 0, TPR = 1 | | | (Minimal) distance to FPR = 0, TPR = 1 | | | 3,000-word | 6,000-word | 8,000-word | 8,000-word, 4,000-word secondary samples |
| | *Croniken* | Papal | Imperial | *Croniken* | Papal | Imperial | *Croniken* | Papal | Imperial | *Croniken* | Papal | Imperial | | | | |
| **20** | 0.38 | 0.69 | 0.13 | 0.19 | 0.47 | 0.08 | 0.19 | 0.28 | 0.19 | 0.33 | 0.53 | 0.25 | 1.20 | 0.73 | 0.66 | 1.11 |
| **40** | 0.33 | 0.53 | 0.13 | 0.27 | 0.47 | 0.13 | 0.29 | 0.47 | 0.13 | 0.29 | 0.50 | 0.20 | 0.99 | 0.87 | 0.89 | 0.99 |
| **60** | 0.25 | 0.44 | 0.13 | 0.17 | 0.38 | **0.07** | 0.19 | 0.41 | 0.13 | 0.31 | 0.50 | 0.25 | 0.81 | 0.61 | 0.73 | 1.06 |
| **80** | 0.19 | 0.41 | 0.19 | 0.23 | 0.38 | 0.13 | 0.15 | 0.38 | 0.20 | 0.27 | 0.47 | 0.20 | 0.79 | 0.73 | 0.72 | 0.94 |
| **100** | 0.19 | 0.47 | 0.20 | 0.21 | 0.41 | 0.13 | 0.15 | 0.38 | 0.19 | 0.38 | 0.50 | 0.45 | 0.85 | 0.74 | 0.71 | 1.34 |
| **120** | 0.21 | 0.41 | 0.19 | 0.15 | 0.38 | 0.13 | 0.17 | 0.34 | 0.19 | 0.38 | 0.50 | 0.45 | 0.81 | 0.66 | 0.70 | 1.34 |
| **140** | 0.21 | 0.47 | 0.19 | 0.17 | 0.47 | 0.14 | 0.10 | 0.31 | 0.13 | 0.38 | 0.50 | 0.45 | 0.87 | 0.78 | 0.55 | 1.34 |
| **160** | 0.21 | 0.50 | 0.20 | 0.17 | 0.41 | 0.13 | 0.15 | 0.38 | 0.14 | 0.35 | 0.50 | 0.35 | 0.90 | 0.70 | 0.66 | 1.20 |
| **180** | 0.21 | 0.44 | 0.19 | 0.17 | 0.44 | 0.13 | 0.13 | 0.38 | 0.14 | 0.31 | 0.50 | 0.25 | 0.84 | 0.74 | 0.64 | 1.06 |
| **200** | 0.19 | 0.34 | 0.19 | 0.17 | 0.41 | **0.07** | 0.13 | 0.38 | 0.14 | 0.33 | 0.50 | 0.30 | 0.72 | 0.65 | 0.64 | 1.13 |
| **250** | 0.17 | 0.44 | 0.20 | 0.13 | 0.47 | 0.09 | 0.15 | 0.38 | 0.13 | 0.31 | 0.50 | 0.25 | 0.80 | 0.69 | 0.66 | 1.06 |
| **300** | 0.17 | 0.47 | 0.20 | 0.13 | 0.47 | 0.09 | 0.15 | 0.38 | 0.14 | 0.29 | 0.50 | 0.20 | 0.83 | 0.69 | 0.66 | 0.99 |
| **350** | 0.21 | 0.44 | 0.19 | 0.13 | 0.47 | 0.09 | 0.13 | 0.38 | 0.14 | 0.27 | 0.50 | 0.15 | 0.84 | 0.69 | 0.64 | 0.92 |
| **400** | 0.19 | 0.44 | 0.19 | 0.15 | 0.44 | 0.09 | 0.11 | 0.38 | 0.10 | 0.25 | 0.50 | 0.10 | 0.82 | 0.68 | 0.59 | 0.85 |
| **450** | 0.17 | 0.38 | 0.19 | 0.15 | 0.47 | 0.10 | 0.11 | 0.34 | 0.10 | 0.25 | 0.50 | 0.10 | 0.73 | 0.71 | 0.56 | 0.85 |
| **500** | 0.15 | 0.38 | 0.13 | 0.13 | 0.47 | 0.10 | 0.11 | 0.34 | 0.10 | 0.23 | 0.50 | 0.06 | 0.66 | 0.70 | 0.56 | 0.79 |
| **550** | 0.09 | 0.38 | **0.07** | 0.11 | 0.44 | 0.11 | 0.12 | 0.28 | **0.10** | 0.21 | 0.47 | 0.06 | 0.53 | 0.66 | 0.49 | 0.74 |
| **600** | **0.07** | 0.38 | 0.08 | 0.12 | 0.47 | 0.13 | **0.10** | 0.25 | 0.12 | 0.21 | 0.47 | 0.06 | 0.53 | 0.71 | 0.46 | 0.74 |
| **650** | 0.09 | 0.34 | 0.09 | 0.11 | 0.38 | 0.11 | **0.10** | 0.22 | 0.11 | 0.21 | 0.47 | 0.06 | **0.52** | 0.59 | 0.43 | 0.74 |
| **700** | 0.13 | **0.31** | 0.14 | 0.12 | 0.41 | 0.12 | 0.11 | 0.25 | 0.13 | 0.21 | 0.47 | 0.06 | 0.58 | 0.64 | 0.49 | 0.74 |
| **750** | 0.16 | 0.31 | 0.17 | 0.12 | 0.38 | 0.12 | 0.12 | 0.28 | 0.13 | 0.21 | 0.47 | 0.06 | 0.64 | 0.61 | 0.53 | 0.74 |
| **800** | 0.23 | 0.31 | 0.21 | **0.09** | 0.31 | 0.12 | 0.12 | 0.25 | 0.14 | 0.21 | 0.47 | 0.06 | 0.76 | 0.53 | 0.51 | 0.74 |
| **1,000** | 0.59 | 0.32 | 0.34 | 0.10 | **0.28** | 0.12 | 0.12 | **0.16** | 0.12 | **0.17** | **0.44** | **0.04** | 1.25 | **0.51** | **0.39** | **0.65** |
| **1,200** | 0.94 | 0.53 | 0.35 | 0.12 | **0.28** | 0.14 | 0.20 | 0.16 | 0.17 | 0.17 | **0.44** | 0.04 | 1.82 | 0.54 | 0.52 | 0.66 |

The shortest distances are in bold underline in each column.

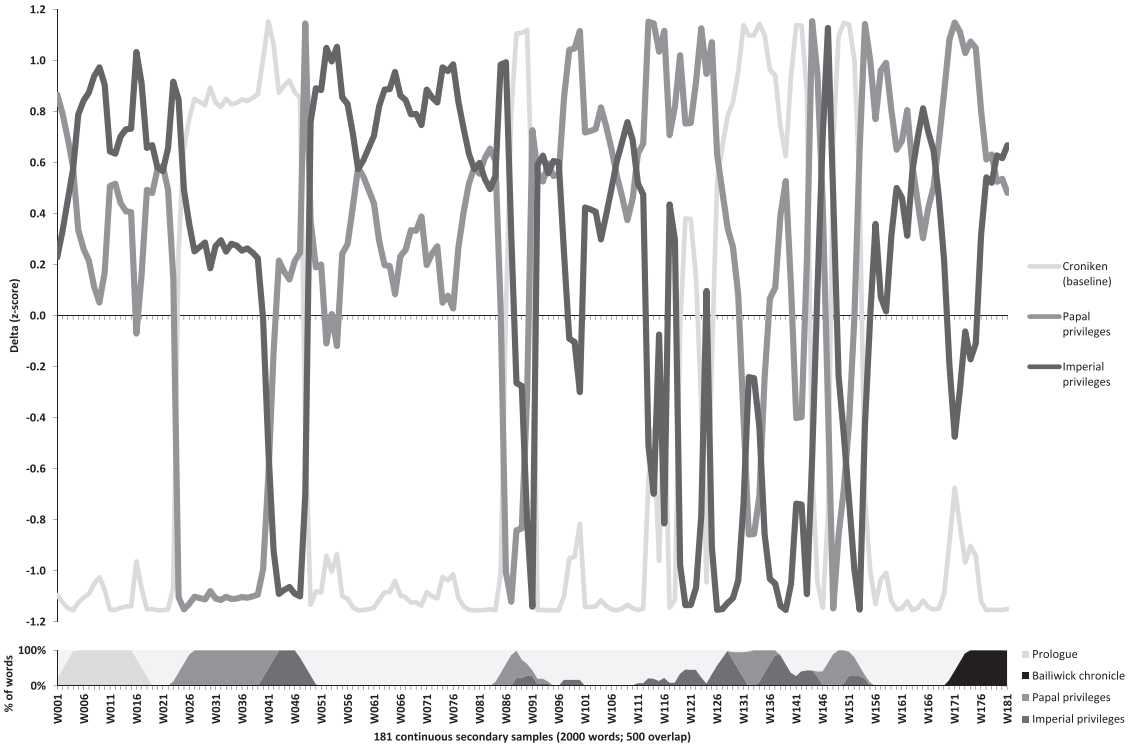


**Fig. 2** *Delta* analysis of the *Croniken* and its privileges: 1.000 MFW, 8.000 word primary samples. Immediately below is a stacked area chart containing the relative weight of the prologue, privileges and bailiwick chronicle in each of the 181 secondary samples

It is possible to pinpoint the location of the privileges in the text. Furthermore, it is possible to distinguish papal and imperial privileges from each other. The classification is even effective in cases where the privilege is small in size. The imperial privilege located at around part ninety-one, for instance, is only 566 words long, and still the *Delta z*-score of the imperial primary sample produces the lowest score. Note that this would be marked as an incorrect classification according to the criteria specified earlier. After all, the regular *Croniken* narrative is the largest class for this secondary sample with 1,118 words of 2,000. Perhaps this is related to specific properties of the baseline sample, incorporating the entire chronicle from beginning to end. The imperial privilege located around part 100 is the shortest privilege in the *Croniken*. But with only 340 words, the imperial primary sample struggles to produce the lowest *Delta* score. Thus, somewhere between 340 and 566 words seems to lie a tipping point for this particular primary sample. However, using the larger 4,000-word secondary samples, it becomes even difficult to detect privileges that stretch well >2,500 words in combined length.

As argued earlier, the privileges were part of existing collections that represent separate authorial styles, distinct from the rest of the *Croniken*. Hence, the fluctuations in Fig. 2 signify changes in authorship, instead of mere genre differences. If genre would be the most determining factor, both primary samples of privileges—belonging to the same genre—would be interchangeable in identifying all privileges in the *Croniken*; in fact, the papal privileges, for example, are seldom good indicators for privileges in general. This assumption that authorial style is measured is strengthened by further collected data that will be discussed later in the text. The areas in the text that contain

privileges constantly pop out as being distinct from the surrounding parts of the text, even when no primary samples containing privileges have been used.

## 5 Analysis *Croniken*

Now we know Burrows' *Delta* method and the proposed workaround is able to produce satisfactory results, it is time to investigate the questions of authorship that have been raised in the past (specifically regarding the prologue and bailiwick chronicle). Two primary samples were selected from areas in the chronicle that show evidence of the *Croniken*'s author actively creating his own narrative. Whether this author was Hendrik van Vianen himself or some other nearby person who made use of Hendrik's services is a difficult issue, but not necessarily at stake here. First, we want to look for shifts in authorial style in the *Croniken*. Only second comes the possibility of linking these styles to persons or source texts.

The presence of an author at work in these two primary samples becomes clear after carefully comparing the *Croniken* with the original source texts, often much earlier chronicles, and investigate how they are used to form a new text. For large proportions of the text, Hendrik (or his client) is not just a *compilator*, who picks and copies short pieces of text and assembles them in the right order, but he uses his sources more loosely to create a new narrative, placing it in his own words and order. There is even evidence that the author created his text while his original source texts were close by, thus excluding the possibility that he merely copied an earlier copy of the *Croniken*. Unfortunately, for this method, you need written source texts to compare them with the text of the *Croniken*. For some parts of the chronicle, this kind of source material is either nonexistent or lost. The bailiwick chronicle and perhaps the prologue are examples of this. For these parts, it could be interesting to use computational methods to figure out whether they have been created by the author who wrote the bulk of the chronicle text or copied from a now lost source, as can be expected for the privileges.

We have copied the approach that was chosen for the privileges (see Fig. 3). As the larger primary samples did produce the best results in the privileges experiment, their size was set at 10,000 words. We chose to include the 1,000 MFW for the same reason. Six primary samples have been selected (see Appendix A): the land charters drawn up by Hendrik van Vianen, his *Sachsenspiegel* copy, two samples of the *Gouds Kroniekje*, and the two aforementioned primary samples taken from the *Croniken* itself. The first of those two is just following the first lengthy set of privileges: corresponding roughly to parts 47–67. The second, roughly parts 100–120, is chosen because it is situated after folio 84vo. On or around this particular page, Hendrik van Vianen interrupted his project to wait around 10 years before picking up his quill pen once again.[8] It might be interesting to see whether this has influenced his writing style (for a closer examination of the theme of stylistic development over the years consider Stamou 2008).

One of the first striking features is that the two primary samples of the *Croniken* are poor markers of the privileges, as we have addressed earlier. The land charters and *Sachsenspiegel* perform much better for these specific areas in the text, indicating a different writing style compared with the regular parts of the chronicle. This confirms the experiment with the privileges. Note that often the land charters produce lower Delta *z*-scores in the areas with imperial privileges, whereas the *Sachsenspiegel* performs slightly better for the papal privileges.

The transition between the two *Croniken* samples is not a gradual one: it shifts abruptly, exactly at the point where folio 84vo can be found (around part 83): the spot where Hendrik van Vianen interrupted the writing process. It might suggest that little changes in the writing style of Hendrik van Vianen over the years—inevitable content-related variations here seem only minor in comparison with other areas in the *Croniken*—can be spotted using *Delta*. Interestingly enough, when we use 3,000-word primary samples (which are not discussed in this article), the two *Croniken* samples are almost interchangeable and the interruption around folio 84vo is not noticeable. By using 10,000-word primary samples, so it seems, more

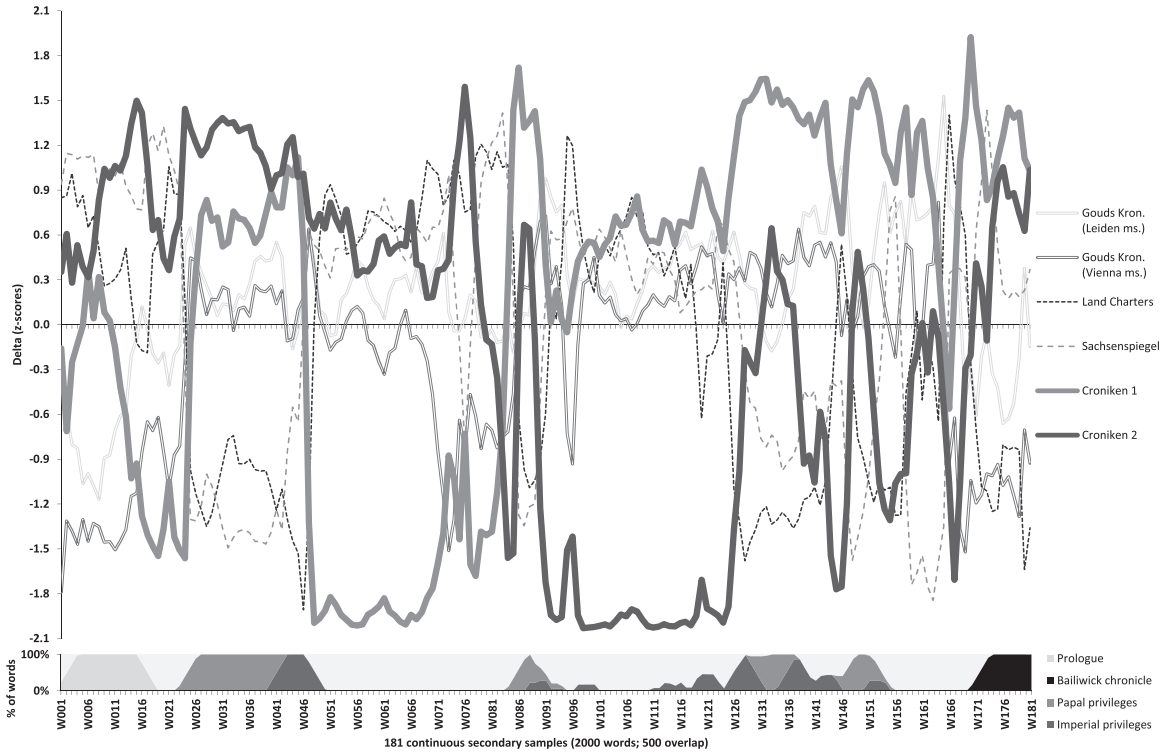


**Fig. 3** *Delta* analysis of *Croniken*: 1.000 MFW, 10.000 word primary samples

detail is preserved that looks to be averaged out in the 3,000-word analysis.

There are only a few areas in the text where the 'non-*Croniken*' samples more or less consistently outperform the samples taken from the *Croniken* itself: in the areas where the privileges are situated; in the prologue: roughly until the point where the author of the *Croniken* indicates that a certain bishop of Paderborn wrote the prologue; a number of indulgences around part 140–142; a short area around part 158–164 where copies of the pleadings of the Teutonic Order against the Prussian Federation before the court of the Holy Roman Emperor are written down (which could well have been copied word-for-word); and, finally, at the end of the *Croniken*, where the bailiwick chronicle is found. The chart becomes more clear when all the privileges and a list of Prussian commandries of the Teutonic Order are filtered from the secondary samples, retaining 123 parts of 2,000 words (Fig. 4). Now and then the *Sachsenspiegel* or land charters—both put down by Hendrik van Vianen—provide the lowest *Delta* *z*-scores, but even the non-related *Gouds Kroniekje* beats the *z*-scores of the chronicle's own samples ever so often. One could presume that the actual author of these areas is not included in any of the primary samples selected. The aforementioned would imply that the author of the *Croniken* was not responsible for creating these parts. He was responsible for gathering and copying them and by doing so enhanced his version of the history of the Teutonic Order.

Interesting in this perspective is a chart of the average number of alterations (additions and deletions to the text; a form of editing) made by Hendrik van Vianen (see Fig. 5). The highly altered list of commandries at the end of the *Croniken*, just before the bailiwick chronicle, is left out to prevent it from dominating the results. The bailiwick chronicle at the end and the first half of the prologue at the beginning of the *Croniken* clearly show less alterations than average. Caution should be in place,

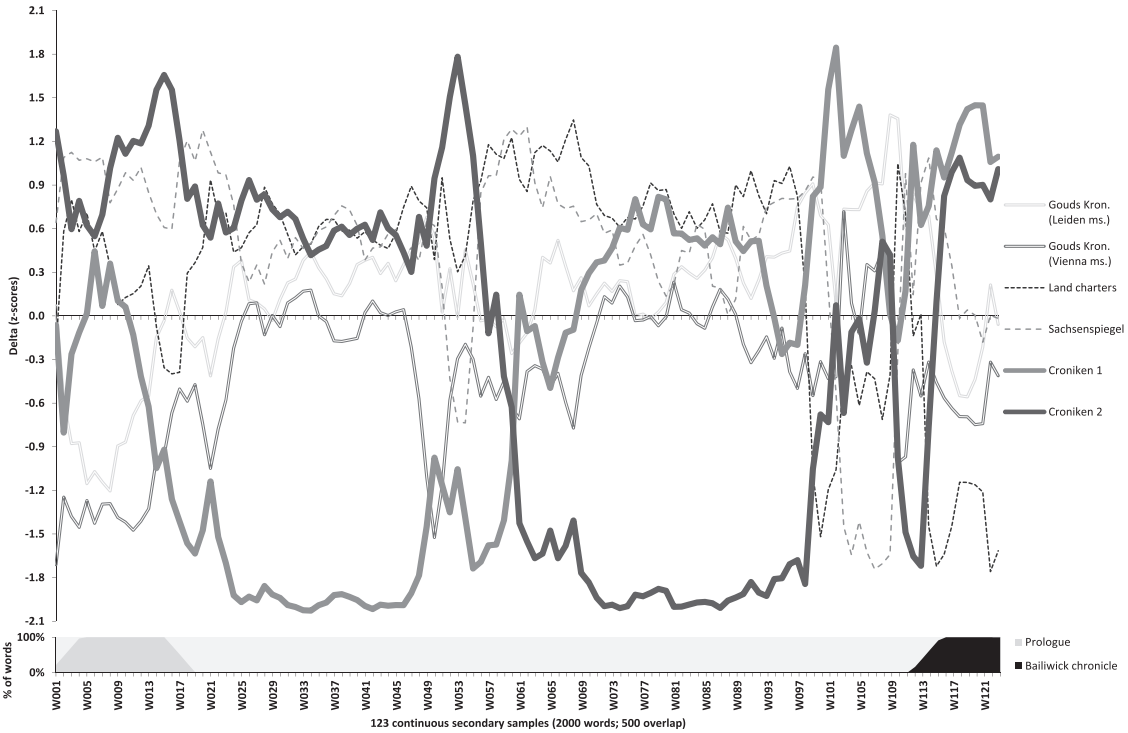


**Fig. 4** *Delta* analysis of *Croniken*, excluding all the privileges from the text: 1.000 MFW, 10.000 word primary samples

but assuming editing is an essential part of the writing process, could this affirm that Hendrik van Vianen was not writing but copying these parts? It has to be said that alterations, especially those non-related to the content, are brought into practice by copyists as well, so one should be careful.

## 6 New Insights to the *Croniken*

One of the possible consequences is that the bailiwick chronicle and part of the prologue could have existed before the *Croniken* itself. What then do we know of these historiographical ancestors of the *Croniken*? First, the prologue, supposedly written by a bishop of Paderborn. We cannot ignore the simple fact that investigations have shown that Bernard II von Ibbenbüren, the bishop of Paderborn at the time, could not have been present at the foundation of the Teutonic Order at Acre in 1190, in contrast to the statements made by the *Croniken*. But even when it is highly unlikely that this bishop of Paderborn wrote the prologue, based on the aforementioned *Delta* analysis, we cannot discard the possibility that there was indeed some other third author who wrote an earlier version of the prologue. This line of enquiry should be investigated further, just as the possibility that specific genre conventions for prologues could have influenced the *Delta* analysis. The existence of an earlier account of this prologue, with its far-reaching ideological program, would be an important find. It would indicate that ideas of these proportions circulated in the order's libraries yet even earlier than the late 15th century.

Then, the bailiwick chronicle. The content of the bailiwick chronicle unmistakably points for its author at the *vicinity* of Land Commander Johan van Drongelen, the direct superior of Hendrik van Vianen. Johan van Drongelen is known to have been active in historiographical circles in the Low Countries, so it could be possible that he himself authored the *Croniken* in corroboration with Hendrik van Vianen, while the latter was

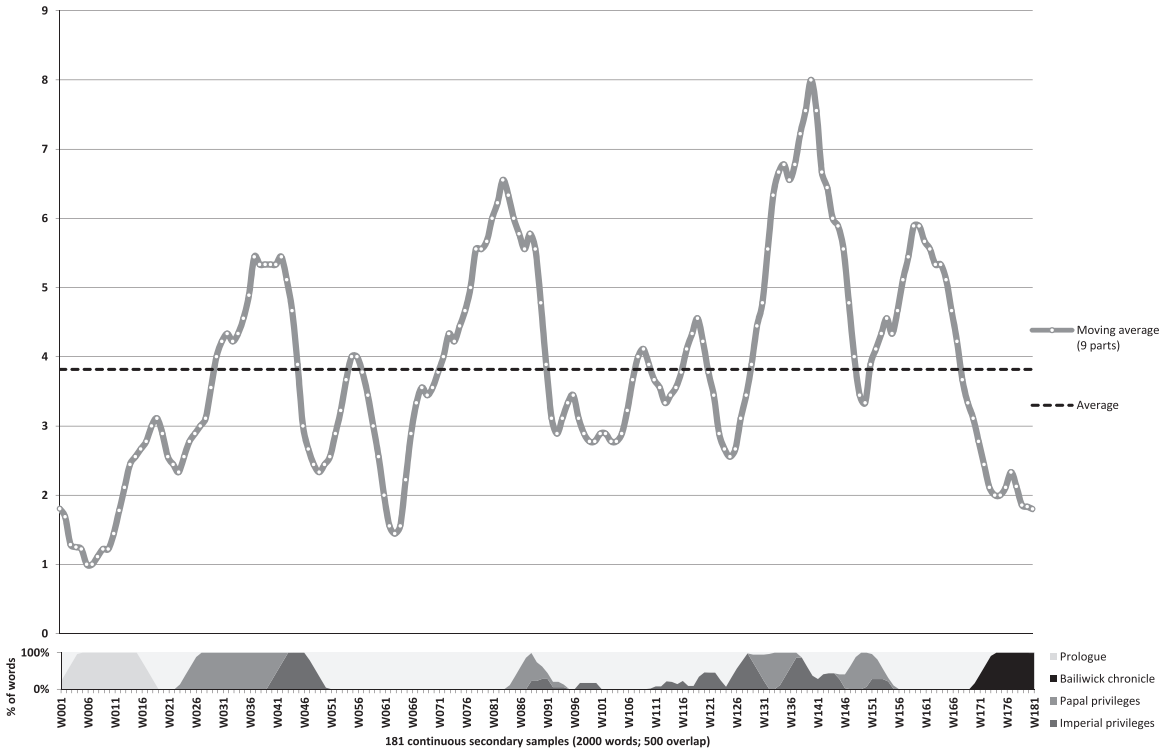


**Fig. 5** Moving average (nine parts) of number of alterations in the *Croniken*

responsible for the bailiwick chronicle. This however will be near impossible to verify—it could even be the other way around. Johan van Drongelen took up the office of Utrecht Land Commander in 1469, and assuming this part of the text was written after this appointment, an earlier version of the bailiwick chronicle must have been written in the following years—but before the rest of the *Croniken*. The conception of the *Croniken*, based on watermark evidence, started around 1480. The second half of the chronicle was continued around 1491. The bailiwick chronicle should therefore be dated between 1469 and around 1491 at the very latest. Palaeographical analysis shows that the final chapter of the bailiwick chronicle that covers Van Drongelen's own life was added in different stages by his secretary, Hendrik van Vianen; partly before and partly after his death in 1492. It was therefore not part of the original composition of the bailiwick chronicle.

Hence, one of new insights this research has provided is that the variety of historiographical works of the Teutonic Order in its bailiwicks could have been much broader than this one-off chronicle written in the Low Countries in the 15th century. Some preliminary examinations stimulated by this outcome have indeed revealed evidence of further historiographical activity in the bailiwicks that previously remained hidden or failed to be placed in the right context. It goes to show that the regional institutions within the Teutonic Order, the bailiwicks, were increasingly self-conscious about their place in history and the order. Just a few decades later, this would mature in tough negotiations by the north-western bailiwicks regarding the height of the contribution to the central authority of the Teutonic Order. Local interest prevailed above the common interest of the Teutonic Order as a whole: a military order that largely lost its legitimacy as defenders of the Christian faith against the heathens.

## 7 Conclusion

To conclude, this article has shown that it might be possible, by using fairly simple and freely available applications, to stretch the limits of Burrows' *Delta* to investigate the authorship and composition of complex medieval manuscripts. Even without the possibility to overcome some obvious problems such as spelling variation and other interfering features so common within these texts, promising results were produced. *Delta* was able to distinguish different authorial layers in the text, as short as just >500 words. Also, it has been shown to be fruitful to examine the authorial composition of a text when one lacks additional primary samples to compare the text with. This all could be of special interest to those scholars who unfortunately lack the time, energy, or skill to begin the laborious task of preparing, encoding, and tagging their texts. Furthermore, it can also help those researchers who work with manuscripts that show an elaborate authorial structure, such as compilations or composite manuscripts.

## Acknowledgments

I wish to thank Mike Kestemont and the two anonymous reviewers for their valuable remarks. All remaining shortcomings are purely my own.

## References

**Burrows, J. F.** (2002). "Delta": a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**: 267–87.

**Craig, H.** (2010). Intelligent archive. Budgerigar version. Newcastle, Australia: Centre for Literary and Linguistic Computing. http://www.newcastle.edu.au/school/hss/research/groups/cllc/intelligent-archive.html (Accessed 12 March 2011).

**Eder, M.** (2010). *Does Size Matter? Authorship Attribution, Small Samples, Big Problems. In Digital Humanities 2010 [Conference abstracts]*. London: Digital Humanities, pp. 132–5.

**Eder, M. and Rybicki, J.** (2011). Stylometry with R. https://sites.google.com/site/computationalstylistics/ (Accessed 12 October 2011].

**Hirsch, T.** (1874). Die Jüngere Hochmeisterchronik. In Hirsch, T., Strehlke, E., and Töppen, M. (eds), *Scriptores Rerum Prussicarum. Die Geschichtsquellen der Preussischen Vorzeit bis zum Untergange der Ordensherrschaft*. Leipzig: Hirzel, pp. 1–172.

**Hoover, D. L.** (2009). The delta spreadsheets. https://files.nyu.edu/dh3/public/TheDeltaSpreadsheets.html (Accessed 27 April 2011).

**Houthuys, A.** (2009). *Middeleeuws Kladwerk. De Autograaf Van De Brabantsche Yeesten, Boek VI (Vijftiende Eeuw)*. Hilversum: Verloren.

**Kestemont, M., Daelemans, W., and De Pauw, G.** (2010). Weigh your words – memory-based lemmatization for Middle Dutch. *Literary and Linguistic Computing*, **25**: 287–301.

**Kestemont, M. and van Dalen-Oskam, K. H.** (2009). Predicting the Past: Memory Based Copyist and Author Discrimination in Medieval Epics. In Calders, T., Tuyls, K., & Pechenizkiy, M. eds. *Proceedings of the 21st Benelux Conference on Artificial Intelligence. Benelux Conference on Artificial Intelligence (BNAIC) 2009*, Eindhoven, pp. 121–8. http://wwwis.win.tue.nl/bnaic2009/papers/bnaic2009_proceedings.pdf#page=134.

**Stamou, C.** (2008). Stylochronometry: stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, **23**: 181–99.

**Stapel, R. J. and Vollmann-Profe, G.** (2010). Cronike van der Duytscher Oirden. In Dunphy, R. G. (ed.), *Encyclopedia of the Medieval Chronicle*. Leiden: Brill, pp. 328–9.

**Töppen, M.** (1853). *Geschichte der Preussischen Historiographie von P. v. Dusburg bis auf K. Schütz. Oder Nachweisung und Kritik der gedruckten und ungedruckten Chroniken zur Geschichte Preussens unter der Herrschaft des deutschen Ordens*. Berlin: Hertz.

**van Dalen-Oskam, K. H., Thaisen, J., and Kestemont, M.** (2010). *Computational Approaches to Textual Variation in Medieval Literature. In Digital Humanities 2010 [Conference abstracts]*. London: Digital Humanities, pp. 37–44.

**van Dalen-Oskam, K. H. and van Zundert, J. J.** (2007). Delta for middle Dutch. Author and copyist distinction in Walewein. *Literary and Linguistic Computing*, **22**: 345–62.

## Notes

1 In my forthcoming dissertation on the *Croniken*, an extensive list of arguments concerning the role of the

scribe as author will be provided. This dissertation is expected to be made available early 2013.

2 My thanks go out to Dr. Antheun Janse (Leiden University), who was kind enough to allow me access to his digital transcripts of the chronicles. The two transcripts are produced from Leiden (ms. 'G') and Vienna (ms. 'F'). Leiden, Universiteitsbibliotheek, Ltk 1564; Vienna, Haus-, Hof- und Staatsarchiv, R 88.

3 Utrecht, Archief van de Ridderlijke Duitsche Orde, balije van Utrecht, inv.nr. 121.

4 In some privileges for instance, the German word for 'and' ('und'; also used in some of the eastern parts of the Netherlands) was left intact instead of translating it into the Dutch equivalent 'ende' that is used throughout the chronicle. Equally, the names in the witness lists have alternating Latin and High German origins.

5 Utrecht, Archief van de Ridderlijke Duitsche Orde, balije van Utrecht, inv.nr. 118.

6 A late 14th-century manuscript in the Utrecht bailiwick archive containing summaries of papal privileges bears a particularly strong resemblance: Utrecht, Archief van de Ridderlijke Duitsche Orde, balije van Utrecht, inv.nr. 120.

7 This constitutes 177 secondary samples in total, all of which have been classified by the same definitions used for the 2,000-word secondary samples. It resulted in 32 positive classifications for the papal privileges, 20 for the imperial privileges, and 125 for the regular *Croniken* narrative.

8 This is neither the time nor place to go into length on this subject. It is based on codicological evidence of the manuscript, palaeographical evidence of the hand of Hendrik van Vianen, as well as his changing personal taste for the use of abbreviations that can be plotted. The upcoming dissertation will address this evidence in more detail.

## Appendix A

**Table A1** Secondary samples. Diplomatic transcriptions by Rombert Stapel

| Text | Reference | Author(s) | Size (words) |
|---|---|---|---|
| *Croniken van der Duytscher Oirden* | Vienna, Deutschordenszentralarchiv (DOZA), Hs. 392 | Hendrik van Vianen (HvV); Anonymous | 91,984 |
| | "W001", "W002" [...] to "W181" | HvV; Anonymous | 2,000, 500 overlap |
| | "W001", "W002" [...] to "W177" | HvV; Anonymous | 4,000, 500 overlap |

**Table A2** Primary samples. Diplomatic transcriptions by Rombert Stapel, except *Gouds Kroniekje* by Dr. Antheun Janse (History Department, Leiden University)

| Text | Reference | Author(s) | Location relative to secondary samples (2,000 words) | Size (words) |
|---|---|---|---|---|
| *Croniken van der Duytscher Oirden* | Vienna, DOZA, Hs. 392 | HvV; Anonymous | | 91,984 |
| | "Baseline" | HvV; Anonymous | 1–181 | 91,984 |
| | Imperial privileges (3,000 words) | Anonymous (scribe/translations?: HvV) | ±42–47 | 3,000 |
| | Imperial privileges (6,000 words) | Anonymous (scribe/translations?: HvV) | ±42–47, 90–91, 98–99, 113, 116, 119–123 | 6,000 |
| | Imperial privileges (8,000 words) | Anonymous (scribe/translations?: HvV) | ±42–47, 90–91, 98–99, 113, 116, 119–123, 125–130 | 7,492 |
| | Papal privileges (3,000 words) | Anonymous (scribe: HvV) | ±25–30 | 3,000 |
| | Papal privileges (6,000 words) | Anonymous (scribe: HvV) | ±25–36 | 6,000 |
| | Papal privileges (8,000 words) | Anonymous (scribe: HvV) | ±25–41 | 8,267 |
| | *Croniken* 1 | HvV | ±47–67 | 10,000 |
| | *Croniken* 2 | HvV | ±100–120 | 10,000 |

(continued)

**Table A2** Continued

| Text | Reference | Author(s) | Location relative to secondary samples (2,000 words) | Size (words) |
|---|---|---|---|---|
| Sachsenspiegel | The Hague, Koninklijke Bibliotheek, 133 H 4 | Anonymous (scribe: HvV) | | 24,460 |
| | *Sachsenspiegel* | Anonymous (scribe: HvV) | First 10,000 words | 10,000 |
| Land charters | Utrecht, Archief van de Ridderlijke Duitsche Orde, balie van Utrecht, inv.nrs. 43.2; 107.2; 443.1a; 443.2g; 443.3; 490.1; 491.1; 503.1; 735.1; 735.5; 747.3; 772.1; 772.2; 772.3; 786.4; 791.2; 791.3; 806.2; 807.2; 808.2; 820.5; 825.2; 825.3; 836.1; 853.1; 853.2; 1777.1; 2248.2; 2287.1 | HvV? (scribe: HvV) | | 14,427 |
| | Land charters | HvV? (scribe: HvV) | First 10,000 words | 10,000 |
| Gouds Kroniekje | Leiden, Universiteitsbibliotheek, Ltk 1564 | Anonymous | | 40,774 |
| | *Gouds Kroniekje*, Leiden ms. | Anonymous | First 10,000 words | 10,000 |
| Gouds Kroniekje | Vienna, Haus-, Hof- und Staatsarchiv, R 88 | Anonymous | | 25,543 |
| | *Gouds Kroniekje*, Vienna ms. | Anonymous | First 10,000 words | 10.000 |