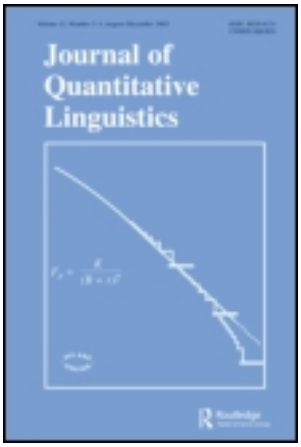

Journal of Quantitative Linguistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/njql20

Robust Rhymes? The Stability of Authorial Style in Medieval Narratives*

Mike Kestemont [a], Walter Daelemans [b] & Dominiek Sandra [b]

[a] University of Antwerp, Institute for the Study of Literature in the Low Countries andCLiPS Computational Linguistics Group

[b] University of Antwerp, CLiPSComputational Linguistics Group

Available online: 18 Jan 2012

To cite this article: Mike Kestemont, Walter Daelemans & Dominiek Sandra (2012): Robust Rhymes? The Stability of Authorial Style in Medieval Narratives*, Journal of Quantitative Linguistics, 19:1, 54-76

To link to this article: http://dx.doi.org/10.1080/09296174.2012.638796

Journal of Quantitative Linguistics
2012, Volume 19, Number 1, pp. 54–76
http://dx.doi.org/10.1080/09296174.2012.638796

Routledge
Taylor & Francis Group

Robust Rhymes? The Stability of Authorial Style in Medieval Narratives*

Mike Kestemont[1], Walter Daelemans[2] and Dominiek Sandra[2]
[1]University of Antwerp, Institute for the Study of Literature in the Low Countries and CLiPS Computational Linguistics Group; [2]University of Antwerp, CLiPS Computational Linguistics Group

ABSTRACT

We explore the application of stylometric methods developed for modern texts to rhymed medieval narratives (Jacob van Maerlant and Lodewijk van Velthem, ca. 1260–1330). Because of the peculiarities of medieval text transmission, we propose to use highly frequent rhyme words for authorship attribution. First, we shall demonstrate that these offer important benefits, being relatively content-independent and well-spread over texts. Subsequent experimentation shows that correspondence analyses can indeed detect authorial differences using highly frequent rhyme words. Finally, we demonstrate for Maerlant's oeuvre that this highly frequent rhyme words' stylistic stability should not be exaggerated since their distribution significantly correlates with the internal structure of that oeuvre.

STYLE AND AUTHORSHIP

Most statistically or computationally supported research into authorship attribution is nowadays style-based, convinced 'that by measuring some textual features we can distinguish between texts written by different authors' (Stamatatos, 2009, p. 538). The basic assumption of such stylometric research is consequently that each author has a unique set of linguistic characteristics or a 'stylome' (Van Halteren, Baayen, Tweedie, Haverkort & Neijt, 2005) that can be quantitatively distinguished from

*Address correspondence to: Mike Kestemont, University of Antwerp, City campus, Prinsstraat 13, Room D. 118, 2000 Antwerp, Belgium. Tel.: + 32 3/220 42 54.
Email: mike.kestemont@ua.ac.be

0929-6174/12/19010054

any other author's style. In this paper we shall focus on the possibilities of style-based authorship attribution for medieval narratives. Although authorship attribution has many relevant applications in the domain of historical studies, it is rarely applied to pre-modern data. In our case study we shall investigate a corpus of rhymed narratives (by Jacob van Maerlant and Lodewijk van Velthem) from the medieval Low Countries (ca. 1260–1330).

An innovative aspect of this research is that it is restricted to rhyme words, instead of plain words as is common in present-day authorship attribution, because plain words are typically vulnerable to corruption by medieval scribes. For present-day authors, highly frequent words have engendered a lot of scholarly interest, since they would contain reliable indications about authorship. In the first part of this paper, we shall therefore assess whether highly frequent rhyme words could be suited for medieval authorship attribution. Note that many contemporary authorship studies tacitly assume that an author's stylome remains relatively constant over time as well as across different texts, topics and text varieties. This supposition has however been challenged (Rudman, 1998). It has been doubted whether an author's style is necessarily constant (Holmes, 1998; Forsyth, 1999; Juola, 2007; Argamon, 2008; Stamatatos, 2009; Luyckx & Daelemans, 2011). In the final part of this paper, we will therefore attempt to determine to what extent the distribution of highly frequent rhyme words is affected by non-authorship related factors in the large oeuvre of a single medieval author.

AUTHORSHIP ATTRIBUTION AND MEDIEVAL LITERATURE

Although computational authorship attribution is surrounded by a lively discussion elsewhere, the discussion is nearly absent in medieval philology (500–1500), which is remarkable. One characteristic of medieval data is namely its problematic survival. For a variety of reasons (e.g. fires) a lot of important resources such as manuscripts have not survived or only in a severe state of damage. Therefore, scholars often lack meta-data on their texts: if a manuscript survives fragmentarily, it is often difficult to determine when or where it was produced. As far as authorship is concerned, we often possess the least information in medieval texts. Many texts are of unknown or disputed authorship and their attribution – to known authors or to the authors of other

anonymous texts – is therefore an important issue. Because of the particular transmission of medieval texts, authorship attribution for them is, however, anything but straightforward.

Before the advent of the printing press in Western Europe all copies of a particular work were manually produced by scribes (Salemans, 2000). Many medieval manuscripts that survive nowadays are in fact copies (of copies) of the original author's text; the original 'autographs' have rarely survived. Manual copying was an error-prone activity, so that scribes unwillingly introduced mistakes in a copy, 'corrupting' the authorial text (Roos & Heikkilä, 2009). No standard spelling or language existed, so that spelling was phonological, reflecting a scribe's personal dialect or regional spelling habits (Kestemont, Daelemans & De Pauw, 2010). Apparently, scribes saw no difficulties in adapting their exemplar's spelling and language and with each copy a text risked an increased deviation from the original (Spencer & Howe, 2001). Below is an example (Table 1) of how one line from the *Rijmbijbel*, one of the texts dealt with below, survives from a series of parallel manuscripts (Kestemont & Van Dalen-Oskam, 2009). Note how scribes have introduced subtle variations in the text or sometimes even changed the wording.

Recent research has shown that the influence of scribes might be even larger than previously assumed (Van Dalen-Oskam & Van Zundert, 2007). The alterations that scribes introduced tend to be systematic and often move beyond mere innocent spelling or dialectal adaptations. In a number of case studies, it has been shown that medieval scribes had a 'style' of their own (Kestemont & Van Dalen-Oskam, 2009). Apparently,

Table 1. An example of the variation in medieval text transmission: one line from the *Rijmbijbel* shows variant readings in a series of parallel manuscripts.

Manuscript	Variant reading for 1 line from *Rijmbijbel* ('On that moment and the same time')
D	Ter stont ende ter seluer vren
E	Tier stont ende ter seluer vren
F	TIere stont enter seluer vren
G	Tottien stonden en ter uren
H	TEn stonden ende ter seluer vren
I	Tjerst stont ende tier veren
J	Tyer stont ende tier seluer vren
N	TJer stont tier seluer vre

scribes enjoyed a large freedom in adapting texts, even to such an extent that their appropriation of texts can be modeled. This raises the issue to what extent the stylistic traits of an original author are preserved in subsequent copies. The strong impact of scribes suggests that the features that are traditionally used in authorship attribution might be of questionable relevance for medieval texts, since these are likely to contain markers for scribal, rather than authorial identity (Kestemont, 2010b).

One interesting and practical bypass for this problem has been suggested: rhyme words (Besamusca, 2003). Throughout the Middle Ages a good deal of the literature was rhymed, an acoustic quality of texts that was of course important for a semi-literate culture, in which literature was received through oral recitation rather than silent reading. For instance in the medieval Low Countries, the rhymed couplet was the preferred verse form for most of the narrative literature until well into the late medieval period (Lie, 1994). The rhymed couplet (aabbccdd . . .) was often used to structure medieval epics of a larger size. What is interesting is that rhyme words tend to be a stable element in medieval text transmission, very robust to scribal corruption (Kestemont, 2010b). Scribes generally refrained from manipulating the underlying rhyme words of a text (Besamusca, 2003), as is also clear from Table 1. It is of course cumbersome to try and change the rhyme words of a text, if one is not to *re*write a considerable piece of it. Even if scribes did change the spelling of rhyme words, the underlying lexemes were often left untouched.

It therefore makes sense to apply stylometric methods to the lexemes (lemmas) of words in rhyme position, since these are likely to contain non-contaminated indications about the original authorship of texts. The number of possible rhyme word combinations in a language is moreover limited: authors were bound to often recycle rhyme words. It is not inconceivable that they would display individual predilections for a subset of these words, and use them as 'stopgaps' or 'mnemonics', once they had proven useful. Frequently recurring rhyme words could therefore function as the 'fingerprint' of an author. The objective of this paper is therefore to study these rhyme words by means of a representative case study and determine whether it is feasible to apply stylometric methods to them. Note that we shall study the use of rhyme words in isolation from the combinations they appear in (e.g. pairs in the case of the couplet). Although rhyme combinations could contain

markers of authorial style too, they fall outside the scope of the present study.

CASE-STUDY: MAERLANT AND VELTHEM

Our case study is taken from the medieval Low Countries and focuses on the surviving works of two medieval Dutch authors (Jacob van Maerlant and Lodewijk van Velthem). Jacob van Maerlant (ca. 1240 - ca. 1300) was undoubtedly one of the most influential authors of the medieval Low Countries – one medieval poet called him the 'founding father of all poets who wrote in Dutch' (Van Oostrom, 1996). Maerlant has left us an extensive oeuvre of narrative texts. The following schema (Table 2) introduces the texts included in our corpus. Our corpus is described in detail at the end of this paper.

In this schema, the texts have been ranked according to their date of composition, suggested by the current state of the art in the research field (Van Oostrom, 1996). Some specific problems have to be taken into account. Both *M2* and *M3* survive in two unique manuscripts and for both of them there are indications that they might be heavily corrupted by the compilers of these manuscripts (Besamusca, Sleiderink & Warnar, 2009). It will therefore have to be determined whether these versions

Table 2. An overview of Maerlant's works included in the corpus.

Full Middle Dutch title	Abbreviation	Text variety
Alexanders Geesten ('The deeds of Alexander the Great')	*M1*	Chivalric
Historie van den Grale ('The history of the holy grail')	*M2*	Chivalric
Roman van Torec ('The romance of the knight Torec')	*M3*	Chivalric
Historie van Troyen ('The history of Troy')	*M4*	Chivalric
Heimelijkheid der Heimelijkheden ('The secret of secrets')	*M5*	Ethic-didactic
Der naturen bloeme ('The best of nature')	*M6*	Ethic-didactic
Rijmbijbel ('The rhyming Bible')	*M7*	Historiographical
Sinte Franciscus leven ('The life of Saint Francis')	*M8*	Historiographical
Spiegel historiael (Derde Partie) ('The mirror of history', third part)	*M9*	Historiographical

sufficiently preserve the original style of the author. In *M4* Maerlant included a text known to be written by another author, Segher Diengotgaf (Kestemont, 2010a). Because of the unclear status of this interpolation, we have excluded this part of *M4* from our corpus. Regarding *M5*, it has been doubted whether Maerlant has actually written it – his name only appears in two of the three surviving manuscripts – and moreover its date of composition is sometimes doubted (Van Oostrom, 1996). Although recent studies do not seem to doubt this anymore, this issue needs special attention. The technique of authorship attribution especially lends itself for addressing it.

It should be noted that the size of Maerlant's oeuvre is fairly large for a medieval author. Selecting a contemporary oeuvre that parallels Maerlant's is not without problems. Lodewijk van Velthem (beginning of the fourteenth century) seems the safest option. Velthem was a great admirer of Maerlant, even to such extent that he has been characterized the executor of Maerlant's literary testament (Van Oostrom, 1996). Only two works survive that can be attributed without any doubt to Velthem (Table 3).

Both works are continuations of two of Maerlant's works: *V1* is, for instance, the sequel to *M9* (see Table 2). *V2* firmly builds upon *M2* and is in fact extant from the same unique manuscript. The lack of an independent tradition for *M2* has sometimes raised the question to what extent Velthem altered Maerlant's original text (Besamusca, Sleiderink & Warnar, 2009). Note that the same is true for *M3*: it survives in a single manuscript, probably compiled by Velthem. Especially in this case, it is often claimed that Velthem might have manipulated Maerlant's *M3* to a large extent.

Because of the literary proximity between these authors, this corpus serves as a good test case for authorship attribution. Our hypothesis is that, if rhyme words can serve as reliable indicators of authorship, it should be possible to distinguish the texts containing the typical rhyme word 'fingerprint' that each author has left on them. The rhyme words in

Table 3. An overview of Velthem's works included in the corpus.

Full title	Abbreviation	Text variety
Spiegel historiael (Vierde en Vijfde Partie)	*V1*	Historiography
Merlijn-continuatie	*V2*	Chivalric

this corpus have been tokenized and lemmatized (Kestemont et al., 2010). All the experiments below are restricted to these lemma tags (i.e. the underlying lexemes), in order to abstract away from superficial spelling variation introduced by scribes. The schema in Table 4 presents some general facts about the lemmatized versions of the texts and rhyme words in this corpus.

AUTHORSHIP AND HIGH-FREQUENCY ITEMS

Although additional linguistic characteristics (such as syntax or character *n*-grams) are widely studied, lexical features remain popular in authorship attribution studies (Stamatatos, 2009). In these studies, text samples are represented as vectors, consisting of a fixed number of parameters indicating the normalized frequencies of a set of words. The main advantages of lexical features are that (a) their performance is generally acceptable; (b) their extraction generally requires little linguistic preprocessing (except tokenization); and (c) their stylistic relevance is often easy to interpret. Note that authorship attribution based on lexical features often does benefit from the inclusion of additional (e.g. syntax-based) feature types but that these other features types in isolation rarely outperform lexical features (Van Halteren et al., 2005; Luyckx & Daelemans, 2011). One feature type that generally does outperform

Table 4. General information about the rhyme words in the texts in the corpus.

Text	Number of lemma tokens	Number of distinct lemma types (per text)	Number of hapaxes (types)
M1	14.237	1790	128
M2	8.601	1224	76
M3	3.854	832	30
M4	38.391	2652	383
M5	2.156	793	40
M6	16.672	2298	409
M7	34.708	2454	327
M8	10.494	1461	105
M9	31.080	2463	293
V1	14.237	1910	209
V2	14.237	1179	84
Total	**188.667**	**5657** (in total)	**2085**

lexical features are character *n*-grams (Luyckx & Daelemans, 2011). However, because of scribal interference, these features are difficult to use for medieval authors (Kestemont & Van Dalen-Oskam, 2009).

Not all words are traditionally included in analyses, and feature selection is used to decrease the dimensionality of a problem by representing only the most relevant features. One obvious feature selection method is the restriction to the *n* most frequent items (Stamatatos, 2009). Only the *n* items that are most frequent in the entire data set are used in the representation of text samples, an approach that is again simple and language independent (at least, for most western languages). This feature selection method often yields excellent results and sometimes even outperforms more refined selection methods. Although many values for *n* have been proposed in the literature, ranging from 30 to 1000 (Stamatatos, 2009), researchers often limit themselves to a small number of frequent items, often less than 100 (Holmes, 1998). The reason for this is that the first hundred or so highly frequent words from any natural language corpus tend to consist of function or stop words, belonging to closed linguistic classes such as articles or prepositions. These highly frequent items are of distinct relevance in authorship studies because they are, among other things, typically (Stamatatos, 2009):

– used by all authors in a corpus (reliable base of comparison);
– to a lesser extent related to content than less frequent words (such as topic-specific nouns);
– not under an author's conscious control (robust to imitation);
– statistically reliable because of their high frequency;

According to the literature, representing text samples by their highly frequent lexical items and analyzing how these items are distributed in different texts seems to be a rather simple but reliable and well-anchored way of attributing texts to authors.

Interestingly, these ideas have a close art-historical parallel in the views of Giovanni Morelli (1816–1891). Many historical paintings have survived anonymously, hence the large-scale research into the authorship attribution of pictorial works. Morelli was among the first to suggest that the attribution of, for instance, a *Quattrocento* painting to some Italian master, could not happen based on 'content'. How Christ was depicted in a crucifixion scene or what kind of coat Mary Magdalene was wearing,

seemed all too much dictated by contemporary fashion or stylistic influences. He proposed to back off to inconspicuous details such as ears, hands and feet: such 'functional' elements were highly frequent in nearly all paintings because they were largely content-independent. It is a beautiful illustration of the stylometric findings discussed, that according to Morelli, authorship attribution in art history should be based on such frequent though inconspicuous elements – a painting's function words.

THE DISPERSION OF 'STOPGAPS'

As argued by the well-known linguist George K. Zipf, the words in any natural language corpus tend to be characterized by the following regularity: if the words from such a corpus are sorted in a list by decreasing frequencies, the word frequencies will be inversely proportional to their rank in the list (Zipf, 1949). Zipf formulated a series of famous laws to capture these regularities. A typical expression of this law can be visualized by ordering the words in a corpus according to their overall frequency and plot these words as points in a two-dimensional space, with the logarithm of their rank on the vertical axis and the logarithm of their frequency on the horizontal axis (Manin, 2009). The result (at least, the fitting regression line) will resemble a straight line with a slope that approximates -1. Zipf's laws should be handled with care: they are not a universal constant and they are highly dependent on the nature and size of text samples (Baayen, 2001; Köhler, 2005; Tuzzi et al., 2009). Nevertheless, the inverse relationship between rank order and frequency is interesting for authorship studies with their strong emphasis on the small 'crest' of highly frequent and well-spread items. Applied to rhyme words, we would like to find out whether rhyme word vocabulary has a similar crest of highly frequent, well-spread items that could be used for authorship attribution. In terms of poetics, we would expect such a crest to consist of 'stopgaps' or a nifty set of relatively vague and content-independent rhyme words that authors often use because of their generic employability. The notion of stopgaps has invited a lot of consideration in traditional philology, but so far has rarely been explored from a quantitative perspective.

If we plot the Zipf-curve (log rank as a function of log frequency) for the ranked list of rhyme words in our corpus (Figure 1), we see that these rhyme words seem to follow a Zipfian distribution. Note that the slope of

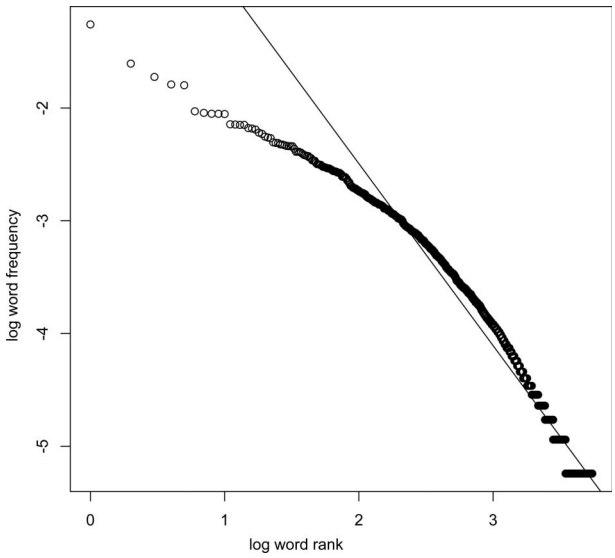

Fig. 1. Zipf-curve (log word frequency versus log word rank) for the 5657 distinct rhyme words in the corpus and the fitted regression line (formula: logfrequency = 0.74 − 1.61* logrank).

the fitted regression line in Figure 1 is lower (≈ − 1.61) than the value of − 1 in the 'classic' Zipf curve. However, fitting a linear model in which a word's logarithmic frequency is predicted by its logarithmic rank yields a highly significant outcome ($F(1, 5478) = 113700$, $p < .0001$, adjusted R^2: 0.95), demonstrating a linear relationship between these two variables in the current data set. It is important that rhyme words show a 'Zipf-*like*' distribution, a first indication that rhyme words as well have a crest of highly frequent, functional stopgaps that can be used in authorship attribution.

High-frequency words are considered interesting in authorship attribution, not because they are highly frequent in individual texts or oeuvres, but also because they tend to be frequent in all texts and oeuvres. All authors need to use e.g. definite articles and prepositions. These words thus 'scale' well to different texts, topics and oeuvres: they tend to be relatively independent from e.g. text variety, frequent and well spread over text samples. These three aspects are of course related: words can only be highly frequent and well spread over a large corpus, if they are indeed independent of text variety and vice versa. For the time being,

we cannot know for sure whether rhyme words display the same characteristics, although the Zipf-like distribution in Figure 1 suggests this to be the case.

Suppose, however, that a particular rhyme word is extremely frequent in one specific large text in the corpus (note that some texts are considerably larger than others, see Table 2). In that case, the rhyme word might pop up in the list of most frequent words in the corpus as a whole, although it is actually only frequent in one text and will not 'scale' very well to other texts. If we are to use highly frequent rhyme words for medieval authorship attribution, we first should determine whether these frequent rhyme words share the important characteristics of content-independence and good spread with their non-rhyme counterparts in modern texts. In the remainder of this section we shall use two procedures to determine this: firstly, the *inverse document frequency*, a weighting measure from information retrieval that depends on the mere occurrence of words (i.e. a Boolean value) in a collection of text samples, and secondly, the *coefficient of variation* that considers the actual frequency (i.e. real number) of words in samples.

These two measures we shall propose hereafter are strongly related to the concept of *polytextuality* (or *polytexty*) in quantitative linguistics. This is a modern systems-theoretical notion that has been coined in the framework of Synergetic Linguistics (Köhler, 1986; 2005). It considers the *dispersion* of linguistic (e.g. lexical) elements over text. This framework is especially relevant with respect to our specific research question. Note that the mere frequency of a lexical item depends on the communicative relevance of its meaning(s)/function(s) (Köhler, 2005) – a function word is frequent because it fulfills a communicatively important function. Note that *frequency* is, theoretically speaking, necessarily a function of *dispersion* but that this does not hold the other way around. From a theoretical point of view, high-frequency words need not have a high degree of *dispersion* (although this will often be true in practice). Essentially, our research question as such essentially relates to one of Köhler's synergetic laws: if a rhyme word is frequent in the whole corpus, is it then also frequent in parts of that corpus (Köhler, 1986)?

The first measure we can use in this respect is the *inverse document frequency* (IDF). This weighting measure is used in information retrieval to find out how *specific* a term is to a given document in a certain collection of documents (Manning, Raghavan & Schütze, 2009, pp. 117–118). The traditional IDF (Formula 1) for a term t is taken to be the

logarithm of the ratio of the total number of documents in the corpus (N) to the number of documents t appears in (df_t):

$$idf_t = \log \frac{N}{df_t} \quad (1)$$

Highly frequent words tend to be not document-specific and thus generally have a low IDF. Words that are rather document-specific, such as for instance hapaxes, tend to have high IDF's. Therefore, there is generally a strong correlation between the rank of a word in a frequency list and its IDF. In terms of quantitative linguistics IDF is a simple transformed proportion that relates to a term's polytextuality: the lower a term's IDF, the more 'polytextual' (communicatively relevant) it is (cf. Köhler, 2005). Note that neither the logarithm nor its base is in fact extremely important in the calculation, although this particular implementation remains standard in information retrieval.

To investigate whether this correlation would also hold true for rhyme words and to avoid the pitfall that a high-frequency rhyme occurred only in one text or even part of that text, we divided all texts in our corpus in equally-sized samples of 1000 rhymes, which yielded a total number of 184 samples for the corpus. Subsequently we listed the rhyme words occurring in these samples, ranked by their overall frequency in the corpus and collected for each rhyme word its relative frequency in each individual sample. Subsequently, we calculated the IDF for each word in the ranking in each of these 184 samples. By working with equal sample sizes we disentangled frequency and spread: high-frequency rhyme words that are well-spread across the set of 1000 samples should have a lower IDF than equally high-frequency rhymes that are restricted to fewer samples. Below, we have plotted the rank of these rhyme words and the corresponding IDF for these words in the collection (Figure 2).

As is clear from this figure there is indeed a significant positive correlation between the frequency rank of a word in the corpus as a whole and its IDF over samples of that corpus ($F(1, 5599) = 36550$, $p < .0001$, adjusted $R^2 = 0.87$), indicating that rhyme words that are relatively more frequent in the corpus are generally less 'sample-specific' than words with a relatively lower frequency in the corpus as a whole. Hence, the higher the overall frequency of a rhyme word, the better its spread.

Nevertheless, this way of working arguably overestimates the spread of words. First of all, since not all works are of equal length, longer works

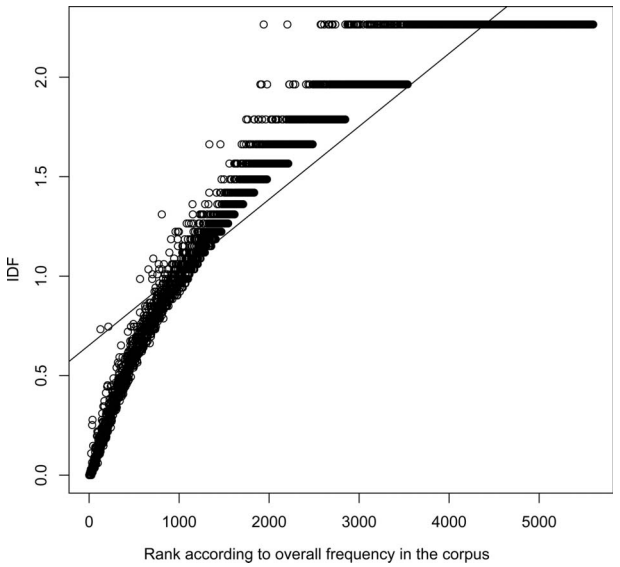

Fig. 2. *Inverse document frequency* versus word rank for 5601 rhyme words in 181 samples of 1000 rhyme words and the corresponding linear model (formula: IDF = 0.65 + 0.00037*rank).

are able to contribute more samples than shorter works. Secondly, the amount of Maerlant data severely outbalances the amount of data from Velthem. We therefore performed two subsequent analyses on subsets of the corpus. In the first one, we dropped the shorter texts (*M2*, *M3*, *M5*) and divided the first 10.000 rhyme words of the texts into samples of length 1000 (80 in total). We then performed the same IDF analysis on these samples as above. In a second analysis, we performed the IDF procedure on a selection that was balanced between Maerlant and Velthem, using 1000 word samples (56 in total) of the first 14.000 lines of 4 texts (*M1*, *M9*, *V1* and *V2*), granting an equal share to both authors. The results of these procedures were nearly identical to the initial one and are therefore not discussed further for the sake of brevity.

Note that IDF has one drawback as a tool for polytextuality-related measurements: it is based on the mere occurrence of a word in a given document (Boolean value) and does not consider the actual frequency of the word in that document (real number), so that it doesn't matter whether the words occurs 20, 200 or 50.000 times in a given document. Another, arguably more fine-grained means of measuring a word's

dispersion is the *coefficient of variation* (CV). This relatively simple statistic measure calculates the 'relative variability' in a series of observations and is defined as the ratio of the standard deviation over the mean for a series of observations (Lewontin, 1966). Consider a dummy corpus that only uses 5 words, occurring in five samples of equal size (Table 5).

Clearly, the CV of a word captures its dispersion in this dummy corpus: the lower a word's CV, the better its spread; the higher a CV, the more unstable the word's occurrence. A CV of 0 would mean that the word is equally frequent in all samples, i.e., that the standard deviation in the nominator of the ratio is 0. In terms of quantitative linguistics, a word's CV seems a fairly faithful way to capture its dispersion or 'polytextuality' (in fact, it seems more faithful than IDF which only considers Boolean occurrences of a word). We have applied the CV-measure to the same subcorpora as used in our IDF analysis. For instance for the first analysis, we first divided the texts in the corpus in samples of equal size, again using 184 samples of 1000 rhymes. Then, we created a list of all words occurring in those samples, ranked from frequent to less frequent. Next, we walked through the frequency list, calculating the CV for all words in the samples. The results are presented in Figure 3. We performed the same analysis for the subcorpora that we also used with the IDF measure. Again this yielded no significant differences with respect to this trend. These results too (Figure 3) reveal a strong positive correlation between a word's overall frequency in the corpus and its CV over samples of the corpus ($F(1, 5599) = 69170$, $p < .0001$, adjusted $R^2 = 0.93$).

The outcome of this section shows that the distribution of high frequency rhyme words (i.e. their degree of polytextuality) seems to offer the same benefits as high-frequency non-rhyme words in modern texts: they are frequent throughout the corpus and display a good dispersion over individual samples and oeuvres, suggesting relative content-independence and good scalability from one text to another. Hence, they are reliable indices for authorship attribution of medieval texts. We will now turn to this issue.

AUTHORSHIP AND DIMENSIONALITY REDUCTION

Nowadays it is quite common to perform an authorship attribution experiment with a Principal Components Analysis (PCA) or Correspondence Analysis (CA), based on the *n* most frequent lexical items in the

Table 5. Overview of fictional word frequencies in a dummy corpus of 5 texts (each 100 words), illustrating how the *coefficient of variation* captures the spread of words over the samples.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Total in corpus	Standard deviation	Mean	*Coefficient of variation*
Word 1	70	60	65	75	69	339	5.630275	67.8	0.08304
Word 2	20	10	15	25	20	90	5.700877	18	0.31671
Word 3	10	20	15	0	10	55	7.416198	11	0.67419
Word 4	0	10	5	0	0	15	4.472136	3	1.49071
Word 5	0	0	0	0	1	1	0.447213	0.2	2.23606
Total in sample	**100**	**100**	**100**	**100**	**100**	**500**			

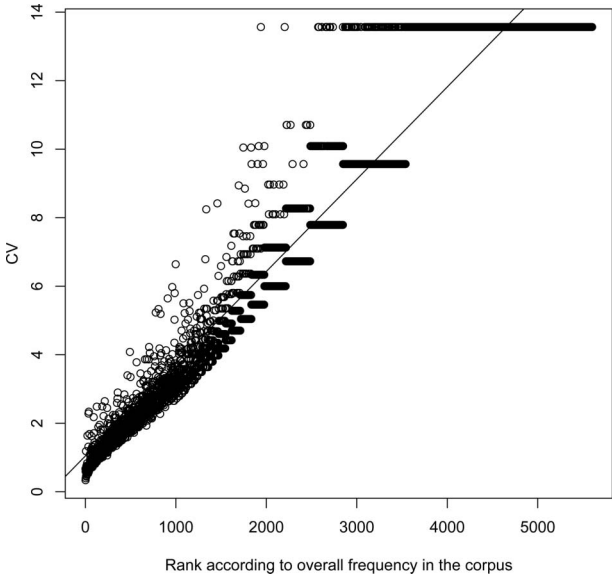

Fig. 3. *Coefficient of variation* versus word rank for 5601 rhyme words in 181 samples of 1000 rhyme words and the corresponding linear model (formula: CV = 1.04 + 0.0027*rank).

corpus (Holmes, 1998, pp. 113–114). These are unsupervised statistical procedures that try to transform a number of possibly correlated variables into a smaller number of uncorrelated variables or dimensions, often for the purpose of visualization. Their main advantage is that they can be easily applied to any corpus, since they are language-independent and require little preprocessing. Nevertheless, when plotting texts in the lower-dimensional space of the first dimensions (called 'components' or 'factors') resulting from such an analysis, texts of identical authorship tend to cluster. We performed a CA (Nenadić & Greenacre, 2007) of our corpus, selecting 81 samples of 2156 lines, so that the shortest text (*M5*) would be fully included. We used the 50 most frequent rhyme words in the corpus. When plotting the first two dimensions resulting from the CA we get the results shown in Figure 4. Analyses with different numbers (between 50 and 150) of rhyme words yielded similar results. Note that we have only plotted the texts in these plots (the 'rows' in the CA's) and left out the features (the 'columns') in order not to overload the plots with information.

Note that two clusters appear in Figure 4: the first dimension is horizontally separating Velthem's works (*V1*, *V2*) from the rest of the works in the corpus. Interestingly, *M3* is far closer to Velthem's works than Maerlant's: as we mentioned in the introduction, there were serious doubts as to whether Maerlant's style was still present in *M3*, since the compiler of the manuscript (Velthem) might have corrupted the text to a large extent. This CA strongly suggests that this is indeed the case. *M2*, on the other hand, does not seem to have suffered an equally large stylistic contamination and so far seems to reflect Maerlant's style rather faithfully. Note that *M5* – a text whose provenance has occasionally been doubted in the past – does not jump out in the analysis and seems to blend in neatly with the rest of Maerlant's texts. Although this analysis by itself (with only one control author) cannot be used as proof for Maerlant's authorship of the text, it certainly does not offer additional reasons to doubt the traditional attribution of *M5* to Maerlant.

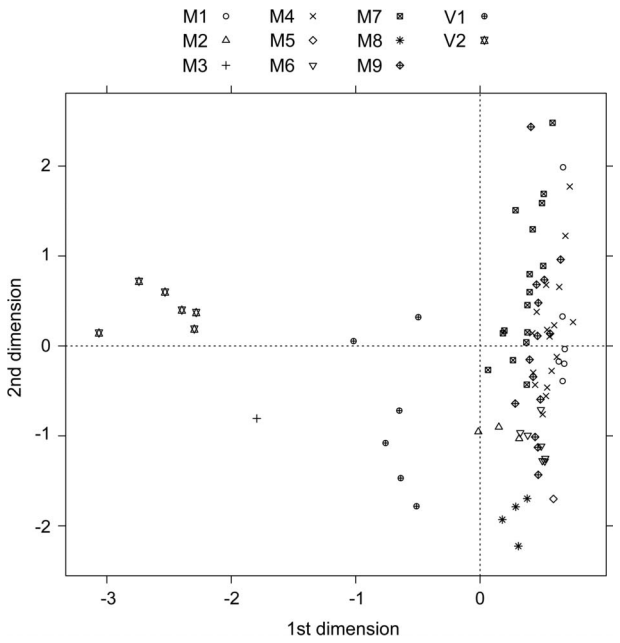

Fig. 4. Result of an unsupervised correspondence analysis of the 50 most frequent rhymes of 81 samples (of 2156 rhyme words) of both authors' texts in the corpus.

INTRA-OEUVRE STABILITY

The fact that such a large share of Maerlant's oeuvre is included in the data set enables some interesting intra-oeuvre analyses. It would be informative to inspect the stability of authorial traits in his texts. The result of CA of Maerlant's texts in the corpus, again for the 50 most frequent rhyme words (with 68 samples of 2156 rhymes) is presented Figure 5. Analyses for other numbers of highly frequent rhyme words (between 50 and 150) revealed strongly similar outcomes.

Here, we see three clusters emerging from the corpus: *M1*, *M2* and *M4* versus *M5* and *M6* versus *M7*, *M8* and *M9*. In order to determine whether this clustering effect is in fact significant, we have computed the Tukey's Honest Significant Differences between the positions of the samples from these three supposed clusters (respectively CHIV, HIST and DIDAC) in both dimensions of the CA (Figure 6). Importantly, this method considers all pair wise comparisons between the three groups we

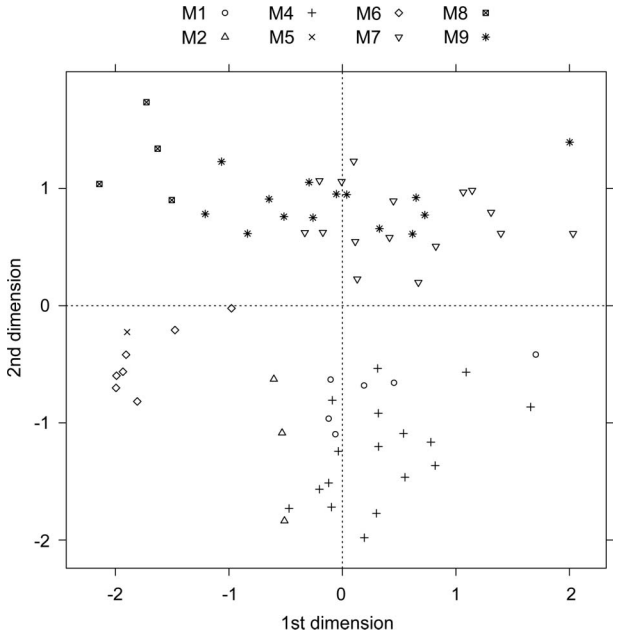

Fig. 5. Result of an unsupervised correspondence analysis of the 50 most frequent rhymes of 68 samples (of 2156 rhyme words) of Maerlant's texts in the corpus.

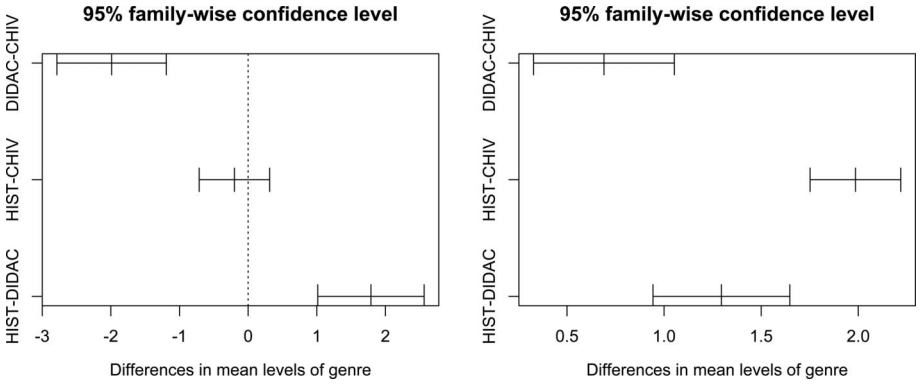

Fig. 6. Visual representation of the Tukey's Honest Significant Differences for the position of texts belonging to the three clusters (named after their genre) in the first dimension (left) and second dimension (right) of the CA in Figure 5.

have defined, while still reporting conservative *p*-values (Baayen, 2008). This way of working has been inspired by the work of Juola (2007), using bivariate statistics to demonstrate that the positions of works from the oeuvre of one author showed a pattern when plotted in a lower-dimensional space (obtained with Multi-Dimensional Scaling).

Figure 6 demonstrates that there is generally a strong correlation between the samples' coordinates and the three clusters we have visually distinguished: with these statistical tests, all differences whose range does not intersect zero are significant, which is true for all of them, except the difference in the first dimension between on the one hand *M1*, *M2* and *M4* and on the other hand *M7*, *M8* and *M9*, i.e., the comparison between Maerlant's chivalric and historiographical works ($p = .62$, all others $p < .0001$). A non-parametric Kruskal-Wallis rank sum test – also for higher numbers of rhyme words in the analysis – further supported the effect that the positions of samples from the three clusters show significant differences in both dimensions of the CA (1^{st} dimension: $\chi^2(2) = 18.55$, $p < .0001$; 2^{nd} dimension: $\chi^2(2) = 53.14$, $p < .0001$).

The bottom line of these statistical results is that this analysis sets out from a feature set of highly frequent rhyme words that are suited for authorship attribution – as demonstrated above – but that this need not imply that this same feature set should display stability or uniformity *within* the oeuvre of one author. On the contrary: the result of a correspondence analysis of these highly frequent rhyme words clearly

shows that their distribution over Maerlant's texts reflects the internal structure of his oeuvre. *M1*, *M2* and *M4* were Maerlant's first three works and they all belong to the text variety of chivalric epics. *M5* and *M6*, Maerlant both wrote in the middle of his career but both these texts have a very ethical, didactic character. *M7*, *M8* and *M9* were the last works Maerlant wrote but all three of them seem to belong to the genre of historiography (or hagiography). Note that the mutual differences between chivalric and historiographic texts are less outspoken, especially in the first dimension, than the differences between these two groups and the didactic group. From the point of view of poetics, this makes sense since texts from the historiographic and chivalric genres tend to be characterized by a similar kind of narrativity – in the end, they both 'tell stories' – and can thus be expected to employ similar highly frequent words to express a sequence of actions.

CONCLUSIONS AND FUTURE RESEARCH

In this paper we have explored the application of stylometric methods developed for modern texts to medieval narratives. Because of the peculiarity of medieval text transmission, highly frequent words are an unreliable base for authorship attribution. We have therefore proposed to use the highly frequent rhyme words in these narratives, since these are likely to contain markers for authorial identity. In the first section we have demonstrated that highly frequent rhyme words offer the same benefits for authorship attribution as normal highly frequent words in modern texts: two analyses (based on the *inverse document frequency* and *coefficient of variation*) show that they are relatively content-independent and well-spread over our corpus.

Further experimentation suggested that rhyme words are indeed suited for medieval authorship attribution. An unsupervised correspondence analysis of frequent rhyme words in the corpus was able to detect the authorial structure in our data, furthermore reflecting (and supporting) the state of the art in the traditional secondary literature about these texts.

Importantly, however, while highly frequent rhyme words thus seem to offer the same benefits as function words, their stability within one author's works should not be exaggerated. The results of a correspondence analysis of the highest frequency rhyme words in Maerlant's oeuvre revealed a significant correlation with its internal meta-structure.

Even if these 'stopgaps' seem suited for authorship attribution because of their content-independence, they seem indeed only *relatively* content-independent. Further research is needed to determine to which extent such *intra*-oeuvre differences interfere with *inter*-oeuvre differences. For medieval as well as present-day authors, it is clear that the challenging task of cross-text variety authorship attribution should be among stylometry's main priorities in the coming years. The current state of the art clearly cannot exclude the possibility that computational authorship attribution might only be reliable within the 'comfort zone' of a single text variety.

CORPUS DESCRIPTION

The selections from *M9* and *V1* are discussed in Kestemont (2010b) (but without the passages therein that are of doubtful provenance, i.e. the Heelu-interpolations and Book 4 of the fifth part). For the demarcation of the excluded Segher-part in *M4*, see Kestemont (2010a). The versions of *M6* and *M7* are the complete digital versions from the Corpus-Gysseling, maintained and annotated by the Institute for Dutch Lexicology (Leyden). The rest of these texts have been entirely harvested in their digital form from the standard editions on the *Cd-rom Middelnederlands* (1996). Note that for some texts (such as *V2*) only a representative sample is used. The *Maskeroen*-passage was not included in *M2* (Besamusca, Sleiderink & Warnar, 2009). The enriched version of the corpus will be made available for download in the public domain via the corresponding author's homepage.

ACKNOWLEDGEMENTS

Mike Kestemont is a researcher (aspirant) with the Research Foundation of Flanders (FWO) and gratefully acknowledges the Foundation's support. All authors would like to thank Frank Willaert and Guy de Pauw (as well as the anonymous reviewers of this journal) for their helpful and encouraging comments on several aspects of this paper.

REFERENCES

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, *23*, 131–147.

Baayen, H. R. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.
Baayen, H. R. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.
Besamusca, B. (2003). *The Book of Lancelot. The Middle Dutch Lancelot compilation and the medieval tradition of narrative cycles*. Cambridge: DS Brewer.
Besamusca, B., Sleiderink, R., & Warnar, G. (2009). Lodewijk van Velthem. Ter inleiding. In Authors (Eds.), *De boeken van Velthem: auteur, oeuvre en overlevering* (pp. 7–30). Hilversum: Verloren.
Cd-rom Middelnederlands (1996). Belgium & the Netherlands, The Hague & Antwerp: Sdu.
Forsyth, R. S. (1999). Stylochronometry with Substrings, or: a Poet Young and Old. *Literary and Linguistic Computing*, *14*, 1–26.
Holmes, D. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, *13*, 87–106.
Juola, P. (2007). Becoming Jack London. *Journal of Quantitative Linguistics*, *14*, 145–147.
Kestemont, M., & Van Dalen-Oskam, K. (2009). Predicting the Past: Memory-Based Copyist and Author Discrimination in Medieval Epics. In T. Calders, K. Tuyls & M. Pechenizkiy (Eds.), *Proceedings of the 19th Annual Belgian-Dutch Conference on Machine Learning* (pp. 121–128). Eindhoven: Eindhoven University Press.
Kestemont, M. (2010a). Seghers wapenfeiten. Oude en nieuwe hypotheses omtrent de *Trojeroman*, het huis van Gaasbeek en het handschrift-Van Hulthem. *Spiegel der Letteren*, *52*, 249–275.
Kestemont, M. (2010b). 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*. *Queeste: Journal of Medieval Literature in the Low Countries*, *17*, 1–34.
Kestemont, M., Daelemans, W., & De Pauw, G. (2010). Weigh your Words – Memory-Based Lemmatization for Middle Dutch. *Literary and Linguistic Computing*, *25*, 287–301.
Köhler (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.
Köhler (2005). Synergetic Linguistics. In R. Köhler, G. Altmann & R. G. Piotrowski (Eds.), *Quantitative Linguistik. Ein internationales Handbuch – Quantitative Linguistics. An international handbook* (pp. 760–775). Berlin & New York: Walter de Gruyter.
Lewontin, R. C. (1966). On the Measurement of Relative Variability. *Systematic Biology*, *15*, 141–142.
Lie, O. S. H. (1994). What is Truth? The Verse-Prose Debate in Medieval Dutch Literature. *Queeste: Journal of Medieval Literature in the Low Countries*, *1*, 34–65.
Luyckx, K., & Daelemans, W. (2011). The Effect of Author Set Size and Data Size in Authorship Attribution. *Literary and Linguistic Computing*, *26*, 35–55.
Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
Nenadić, O., & Greenacre, M. (2007). Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca-Package. *Journal of Statistical Software*, *20*, 1–13.
Roos, T., & Heikkilä, T. (2009). Evaluating Methods for Computer-Assisted Stemmatology Using Artificial Benchmark Data Sets. *Literary and Linguistic Computing*, *24*, 417–433.

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

Salemans, B. (2000). *Building Stemmas with the Computer in a Cladistic, Neo-Lachmannian Way: the Case of Fourteen Text versions of Lanseloet van Denemerken*. Nijmegen: Nijmegen University Press.

Spencer, M., & Howe, C.J. (2002). How Accurate Were Scribes? A Mathematical Model. *Literary and Linguistic Computing*, *17*, 311–322.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, *60*, 538–556.

Tuzzi, A., Popescu, I.-I., & Altmann, G. (2009). Zipf's Laws in Italian Texts. *Journal of Quantitative Linguistics*, *16*, 354–367.

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

Van Halteren, H., Baayen, H., Tweedie, F., Haverkort M., & Neijt, A. (2005). New Machine Learning Methods Demonstrate the Existence of a Human Stylome. *Journal of Quantitative Linguistics*, *12*, 65–77.

Van Oostrom, F. P. (1996). *Maerlants wereld*. Amsterdam: Prometheus.

Manin, D. Y. (2009). Mandelbrot's model model for Zipf's Law: can Mandelbrot's model explain Zipf's Law for Language? *Journal of Quantitative Linguistics*, *16*, 274–285.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley.