Predicting the Past: Memory Based Copyist and Author Discrimination in Medieval Epics

Mike Kestemont[a] Karina Van Dalen-Oskam[b]

[a] *CLiPS Computational Linguistics group, University of Antwerp, Belgium*
[b] *Huygens Institute KNAW, The Hague, The Netherlands*

Abstract

In this paper we will focus on the scribal variation in manually copied medieval texts. Using a lazy machine learning technique, we will argue that it is possible to discriminate between scribes, implying that they did adapt texts when copying them. Consequently, we will assess to what extent scribal interventions compromise our ability to detect the original authorship of medieval texts. It will be shown that, if the right features and weighting methods are used, the automated discrimination of both copyists and authors is possible for medieval texts. The case studies presented suggest that scribes only corrupted the original texts in a shallow and superficial way, leaving authorial features generally intact on deeper levels. This result will be of interest for research into e.g. contemporary newspaper articles when trying to detect editorial interventions.

1 Introduction

Applications in the field of Artificial Intelligence become practically relevant, whenever their performance on a given task equals or surpasses the performance of humans. That is to say: an application should be able to extract and interpret information from a certain data source in such a way that the result of this process could not have been obtained by a human. Literature, for that matter, is an interesting case. Humans have since long enjoyed reading texts and reflecting about them. People have aesthetic abilities which allow them to easily answer the question whether they consider texts e.g. 'boring', 'moving' or 'beautiful'. However, readers, be they scholars or not, will often have difficulties performing other tasks, such as attributing texts to authors. Research has shown Machine Learning algorithms to be equally or better fit for this attribution task. Especially for modern text material, it has been shown that (combinations of) features can be extracted from texts that seem to be linked to the identity or 'stylome' of specific authors [14, 6, 7]. The ease with which such problems can be solved, should not be exaggerated, as this seems highly dependent on, for instance, the number of potential authors and the amount of training material representing them in an experiment [9]. Nevertheless, given a realistic, confined authorship attribution task, good results can be obtained.

In this contribution we will not focus on modern material – like most studies in this area do. Instead, we will focus on two Middle Dutch texts, written in the medieval Low Countries. A complicating factor is that medieval literary texts have often only come down to us in *manu*scripts. Manual copying not only invited unconscious mistakes, it also gave scribes the opportunity to freely adapt spelling, wording, or even content according to their own wishes [13, 11]. Since we often only have copies of (copies of copies of ...) medieval texts, it is *de facto* uncertain what the original text was like. The inspiration for this present contribution lies with previous research into a particular Middle Dutch epic, the *Roman van Walewein* [2, 5]. This Arthurian romance was originally written in Flanders, probably around 1260 AD. Its 11,202 verse lines only survived in a single complete manuscript, copied by two scribes: the first scribe copied lines 1 to 5783; the second lines 5784 to 11,202. Near the end, the text states that the original *Walewein* had been written by two authors: Penninc started it, and Pieter Vostaert would have finished it, by adding a prologue before Penninc's share and 'appending about 3,300' lines to it. This dual authorship had long intrigued Middle Dutch studies, when in 2007 Van Dalen-Oskam and Van Zundert approached the matter from a computational perspective [5]. Using Yule's K and Burrows's Delta they 'windowed' through the text, trying to find the exact location where Vostaert took over from Penninc [1, 15]. Their results proved promising but not concluding: their outcome pointed to the expected area of the author change (somewhere in the range of lines 7,872 - 7,930) but remained unable to pinpoint an exact location. Copyist interference appeared to be a complicating factor, as their method seemed to have less difficulties discriminating copyists than discriminating authors. As such, it was unclear to what precise extent authorial character traits were blurred or even wiped out by scribal interventions.

This state of affairs resulted in our present research interest. In section 2, we will assess the possibilities of medieval copyist discrimination on the basis of a difficult test corpus. We will report on the feature extraction (2.1) during our experiments (2.2) and optimization techniques (2.3). Finally, we return to the *Walewein* (3): using our knowledge from previous experiments, we will experiment on both copyist discrimination (3.2) and author discrimination (3.2). Conclusions will be drawn in the final section (4).

2 Test corpus for scribal variation

Discriminating copyists is a topic on which there has been little quantitative research; exceptions are a.o. [13, 11]. It does not seem a trivial task, since, generally speaking, scribes seem to have altered texts only on a superficial level (e.g. spelling or dialect). Therefore, it is unclear to what extent copyists could have left their fingerprints on the texts they copied. To assess the possibilities of copyist discrimination we set out with experiments on a rather difficult test corpus. We chose a text written by the influential Flemish author Jacob van Maerlant in 1271: the so-called *Rhyming Bible* (*Rijmbijbel*), a Middle Dutch translation of the Medieval Latin *Historia Scholastica*.[1] The Middle Dutch text consists of almost 35,000 lines and is handed down to us in fifteen parallel text witnesses, each written by a different scribe, which we refer to as A to O. From each of these (partially damaged or incomplete) manuscripts we selected and compiled a corpus of 15 parts of ca. 230 parallel verse lines. Subsequently, this corpus was manually lemmatized and enriched with part-of-speech tags (a very basic tagset of 17 tags was used).

Scribe	Line
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

Table 1: Illustration; some scribal variants of the first line in the parallel corpus

2.1 Feature extraction

Deciding which kind of features to extract for copyist discrimination was not straightforward. A lot of seemingly 'obvious' features that could be included for this task, have been used in authorship related research as well [7]. It was unclear beforehand which features would be reliable indicators of scribal, rather than authorial identity. Our feature selection was related to three linguistic notions: lemmata, part-of-speech categories and clitics. Lemmatization refers to the assignment of plain *tokens* to a headword. In English, for instance, *am*, *are* and *been* are tokens that can all be related to the same headword or *lemma*, in this case 'TO BE'). Part-of-speech tags (PoS) are linguistic labels that express to which morpho-syntactic category a token belongs. In the sentence 'John eats chicken.' *John* would be labeled a 'proper name', *eats* would belong to the 'verb' category, whereas *chicken* would be a 'noun'. A clitic token is a single token that actually contains several concatenated tokens that in normal writing would be separated by white space – in colloquial writing one might use the atomic token *wassup* whereas official spelling would prefer *what's up*

[1] English references are sparse; refer to [10] for an introduction to the text complex in English.

(with white space). We did experiments with three broad categories of feature sets: (a) ratios, (b) character n grams and (c) word-level tags. The first category included 10 rather simple ratios, measuring different kinds of textual ‘richness’ (cf. table 2). The character n grams category deals with grapheme or character frequencies. From the texts, we extracted the relative frequencies of common (combinations of) graphemes, including whitespace [13]. We considered unigrams (e.g. ‘ ’,‘u’,‘n’, ‘i’, ‘g’,‘r’,‘a’,‘m’,‘s’,‘ ’) as well as the 50 most frequent bigrams (e.g. ‘ b’,‘bi’, ‘ig’, ‘gr’,‘ra’,‘am’,‘ms’,‘s ’) and trigrams (e.g. ‘ tr’,‘tri’,‘rig’, ‘igr’, ‘gra’,‘ram’,‘ams’, ‘ms ’). The final category concerned the relative frequency of tokens, as well as the lemmata and PoS-tags assigned to them. We measured the set of the 50 most frequent tokens (without any kind of spelling normalization), the set of the 50 most frequent lemmata and finally the set of the 15 most frequent part-of-speech tags. The effectiveness of these features would be determined by only one factor: the degree in which they could avoid ‘picking up’ textual characteristics that were not related to scribal variation. Many of these features would probably suffer from interference of content or even authorial style, hindering scribal classification.

Ratios
- characters per verse line (#characters / #lines)
- tokens per verse line (#tokens / #lines)
- characters per word (#characters / #tokens)
- unique tokens (#unique tokens / #tokens)
- token richness (#tokens / #lemmata)
- clitics (#clitics / #tokens)
- PoS-richness (#unique tags / #tags)
- lemma-richness (#unique lemmata / #lemmata)
- spelling consistency (#unique tokens/#unique combinations of lemma and PoS)

Table 2: Description of the features in the ratios category

2.2 Experimental set up

We approached copyist discrimination as a boolean classification task: given a set of both positive and negative training examples for a particular target scribe, a machine learning algorithm should be able to classify held out test instances as (not) belonging to that scribe’s hand. For our experiments, we used a so called *lazy* machine learner, the *Tilburg Memory Based Learner* (TiMBL) [3].[2] Because of the limited amount of training data available for each scribe, eager learners would probably have difficulties not to *over*generalize over the data. Especially for language data, a classifier should not only detect regularities in the data, but should also be sensitive to the many *sub*regularities that are so typical of linguistic data. A lazy learner would hopefully not abstract away from the specificity of rare (sub)cases in the sparse instance space and could be expected to display robustness during classification [4]. On the front of classification, no real new techniques will be presented. The novelty of our contribution, however, lies with the experimental application of AI-techniques onto medieval textual criticism, a field that up until now has been largely ignored in AI-studies.

Our experiments were set up as follows. Each scribe’s share would be split up into *n* (=10) samples of *x* verses. *N* and *x* were rather ‘tricky’ factors: an increasing/decreasing *n* value would lead to a smaller/larger amount of training instances, but given the little data for each copyist, this would also imply smaller/larger samples and thus less/more representative feature vectors. Each scribe’s parts (*target copyist*) in the corpus would be confronted with the share of every other one (*confusor*). On these twenty overall samples, we performed 10-fold cross validation, whereby the training material in each fold would consist of 9 positive examples of the target copyist and 9 negative examples belonging to the confusor (18 in total). The test set in each fold would consist of two held outs: one positive example and one negative example. For each of our 10 folds we calculated the average F-score for the target copyist class, representing the harmonic combination of both recall and precision for the target.

In many ways, the ranking of the results, presented in table 3, is not surprising. The feature set with the highest discriminating power is the token category. As we are dealing with tokens and not with their normalised lemmata, this clearly indicates that this feature set does not so much pick up e.g. content variation

[2] The software and reference guide are available from http://ilk.uvt.nl/timbl.

Ranking	Feature set	F-score
#1	*tokens*	69.55
#2	*unigrams*	62.76
#3	*bigrams*	59.05
#4	*PoS*	55.49
#5	*trigrams*	52.48
#6	*ratios*	51.62
#7	*lemma*	43.14

Table 3: Average F-scores per feature set

but rather the spelling and dialectal variation for these highly frequent words. Next come character unigrams: apparently, these are able to abstract away from content, modelling scribal differences in a rather reliable way. Next in the ranking come the bigrams. Given the fact that these larger grams are more likely to suffer from content interference, they do worse than the simple unigrams, but logically better than the trigrams. PoS are ranked fourth, which seems to indicate that copyists did have some influence on the syntax of a text. Subsequently, the ratios category with only 10 features does surprisingly well, especially when compared to the worst performing category: lemmata, which seems to be all too much related to deeper content.

However, an important issue should be raised here. In figure 1, we plotted the result for each of the top three scoring feature sets per copyist. The figure is hardly revealing but that is exactly the point: not only does every scribe differ in what *degree* he can be discriminated from his confusors, scribes also highly differ in which *aspects* they can be distinguished. This illustrates the fact that we are not dealing with colourless transmittors, but individuals who all fingerprinted the text in their very own way. An important challenge, therefore, lies with fine tuning our techniques, so that they would not only obtain better results, but can also cope with these individual variations.

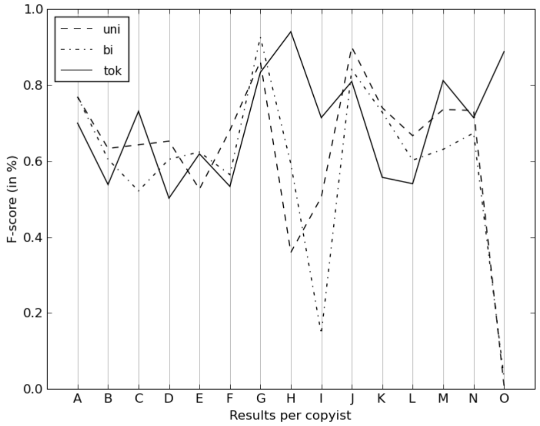

Figure 1: Individual results of the top-3 feature sets per copyist

2.3 Optimization

As is common in authorship attribution studies, combining the features discussed above in an intelligent way might boost our initial scores (e.g. [6]). However, the mere combination of the feature sets discussed above would result in too large feature vectors per instance, hardly fit for this task. To automatically compress our feature vectors to a more reasonable size, we used a Chi Square weighting method [6, 9]. From the bulky combination of our original feature sets, only the top-n values (displaying the largest divergence between the positive and negative items during training) were included in the actual learning phase. We plotted the effect of this feature selection in figure 2. We started off with the combination of the three best scoring feature sets, gradually adding the feature sets that performed worse during initial experimentation. Adding features does boost accuracy and subsequent combinations rather consistently outperform the initial model.

The chi square metric thus seems to do its work, selecting only those features that seem most revealing for the discrimination task under scrutiny. Compressing the feature sets also helps: going from 125 to 25 features, we observe a fairly constant decrease of some 50% in the overall error rate.

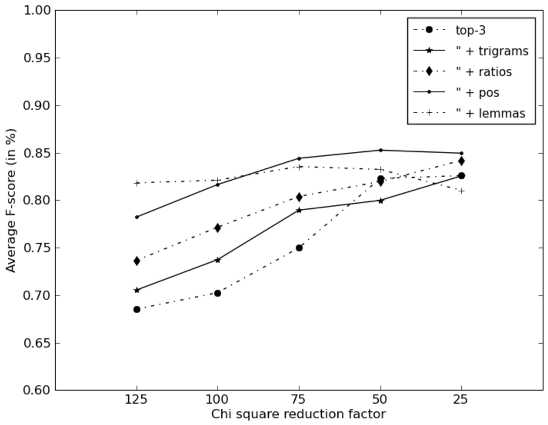

Figure 2: Effect for chi square selection of features

3 Back to Walewein

3.1 The scribes of the Walewein

Our initial experiments clearly show that copyist discrimination is possible, although hard. Even when trying to discriminate between two fairly similar copies of an identical text, intelligent feature extraction allowed us to discriminate between copyists to a fair extent (in the best case ca. 85% for the average F-scores). Given the promising results for this hard and limited data, it would be interesting to return to our initial case, where much more training data was available: the two authors and two copyists of the Walewein.

We experimented on copyist discrimination for the *Walewein*. We divided the text into two roughly equal parts: copyist A being responsible for the first 5783 lines and B for the rest of the manuscript. Subsequently, we divided both categories into samples of equal size. On this data set we performed leave one out cross-validation, whereby each time one example (either A or B) would be held out and the rest of the samples would serve as training material (see figure 3). The classification task remained similar: predicting whether the held out sample belonged to scribe A or B. Using the same optimized feature extraction techniques from our initial *Rhyming bible*-experiments, we were able to gain competitive scores, presented in figure 3. As much more training data was available for this particular task, much higher accuracy scores could be obtained. Sample size matters: starting from 50 verse lines per sample, leave one out testing displays reasonable scores but from 125 lines onwards the technique used seems even flawless.

3.2 The authors of the Walewein

One could wonder whether our binary classification techniques for copyist discrimination, would also be of any use in authorship discrimination. We therefore did some exploratory experimentation on a smaller, distinct data set. As mentioned above, the second author of the *Walewein* would have taken over from his predecessor somewhere between the verselines 7500 and 8000. As training data we therefore selected two sets. First, the 1717 lines before line 7500 (certainly written by author A, but written down by copyist B) and secondly, the 1717 lines after line 8000 (certainly written by author B but again written down by the same copyist B). We performed leave one out cross-validation on samples of 75 verse lines from these 'safe areas', located outside the edges of the area where we could expect the change in authors. Table 4 displays the effectivess of several feature sets for this task.

In this table we also present the ranking for each feature set in table 3, where we looked at the relevance of each feature set in discriminating scribes. One large trend emerges: features that did well in

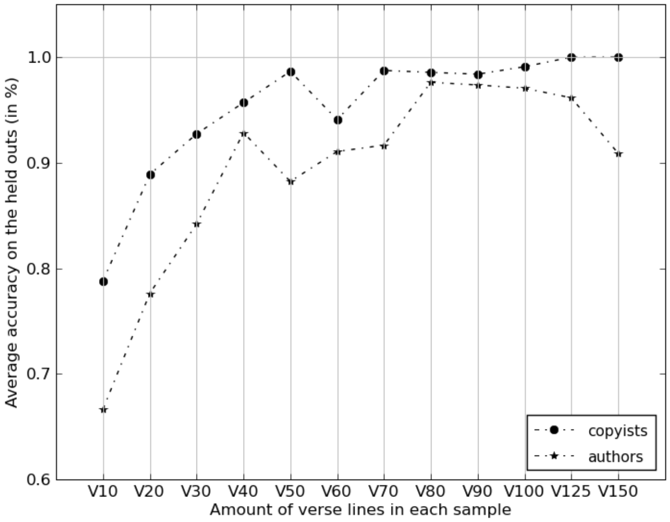

Figure 3: Effect of sample size on classification of the authors and copyists in the Walewein

Current ranking	Previous ranking	Feature set	Overall accuracy
#1	/	*lemma (51-101)*	86.36
#2	#6	*ratios*	81.82
#3	#5	*trigrams*	79.54
#4	#2	*unigrams*	72.73
#5	#3	*bigrams*	70.45
#6	#1	*tokens*	68.19
#7	#7	*lemma (1-50)*	65.91
		Chi weighted top five	**95.45**

Table 4: Average F-scores per feature set for author discrimination

discriminating scribes, perform poorly in telling authors apart. The best example are the trigrams versus the bigrams. Trigrams seem more likely to be influenced by content or authorial style than bigrams and thus performed worse than the bigrams in telling scribes part. However, for exactly the same reasons, they obviously outperform bigrams for authorship discrimination.

One other interesting matter can be observed. The top-50 occuring lemmata work equally bad for copyists and authors; they seemed to be a neutral indicator of language or even discourse. However, following the work of Van Dalen-Oskam and Van Zundert 2007 we experimentally added an extra feature set. As mentioned, they grounded their analysis on type-token ratios of frequent words and Burrows's Delta [5]. When operating in higher frequency strata, their windowing techniques did well in detecting the change in scribes, but not so well in authors. However, when backing off to slightly lower frequency strata, their results for detecting the change in authors significantly improved. This led them to formulating the hypothesis that it was in higher frequency strata that changes in copyists were to be detected, whereas changes in authorship were only to be witnessed in slightly lower strata. To test their hypothesis we added one frequency stratum to our own feature set: the range of most frequent lemmas between 50 and 100. The results for this feature set were astonishing: in leave one out testing their predictive power on its own reached some 86%, leaving all other features behind them and jumping to the first place in the ranking. This surprising result might be of importance for the study of contemporary texts as well, should there be a resemblance between scribal interventions and those of, for instance, newspaper editors. It is often noted that 'there is a general worry, with newspapers that the texts of the authors are often changed by editor(s)' (p. 5 in [8] or p. 485 in [12]). This approach might shed a new light on these matters.

We subsequently did similar leave one out testing on chi square compressed feature vectors (50 features) of the top five scoring feature sets, attaining an average result of some 95% for samples of 75 lines. By adjusting the sample size, as with the scribes, better results could be obtained (see figure 3). In the case of samples of 80 lines, an overall 97,62% could even be observed. These results make one thing clear: the fact that these textual shares of both author A and author B have been 'filtered' by one copyist, even to such a large extent that the copyist can easily be distinguished from his scribal counterpart in the manuscript, does not stand in the way of our discriminating both authors. Because of this result, we tried to answer one final question: would it be possible to pinpoint the location of authorial change more precisely in the text? To answer this question, we applied a windowing technique, similar to the one adopted by Van Dalen-Oskam and Van Zundert 2007. We used samples of various sizes to window, line by line, through the area where the change in authors supposedly took place. In the case of 100 lines samples, the first window thus would cover lines 7500 to 7600, the second 7501 to 7601 and so forth. We applied various window sizes: from our leave one out experiments we knew that larger window sizes (e.g 80 lines) would yield more accurate results but on the other hand, larger windows would result in a less precise indication of the take over. The tiny black pipes in figure 4 indicate for each window size where the classification system detected the second author.

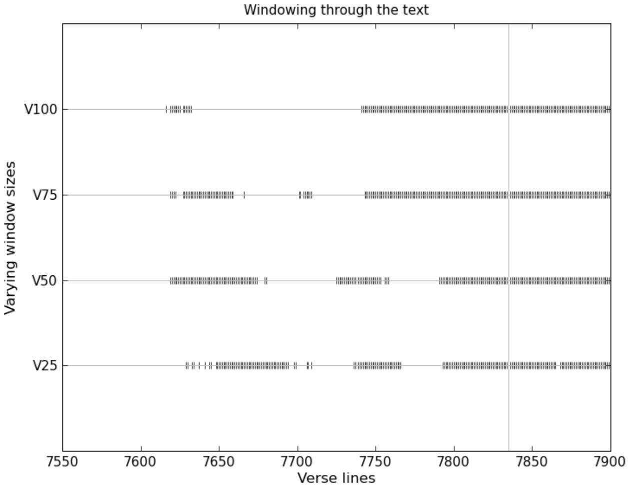

Figure 4: Windowing through the Walewein, predicting where author B started

All window sizes agree that somewhere before line 7800, author B has taken over. However, it is fascinating that all window size experiments also detect author B (at least once) before the final take over. Given the fact that our classification system did not prove flawless during previous testing, chances exist that misclassification occured. Nonetheless, given the observation that during leave one testing the samples between 70 and 100 lines, all had scores in the upper nineties, something else might be the case, as the amount of supposed misclassifications is far more numerous than one could expect. Possibly author B intervened in A's text, before the final take over. As mentioned, we also know that B had no problem with adding a short prologue to A's text. Moreover, Van Dalen-Oskam and Van Zundert 2007 also found evidence that supports this hypothesis. A feature extraction that is more focused on detecting authorial (rather than scribal) style might sharpen our view on these matters in the future.

4 Conclusion

In this study we focused on scribal variation in medieval manuscripts. We showed that lazy machine learning techniques in combination with intelligent feature extraction, are capable of discriminating medieval scribes. Even when scribes would have altered texts only on a superficial level, it is possible to tell them mutually apart by mere linguistic means. Needless to say, the success of the applied technique is highly dependent on both the specificity of the task and the amount of reliable training data available. These results have far-reaching implications, as they suggest that scribal interventions might compromise our possibilities to model the original authorship of manually copied medieval text material. Further experimentation showed that, at least for the time being, there is no need to worry. The fact that material from two distinct authors was 'filtered' by the same copyist, even to such a large extent that the copyist could be readily distinguished from his scribal counterpart in the manuscript, did not heavily withstand our discriminating both authors. On the basis of the texts studied here, one could be inclined to think that medieval scribes only ‘appropriated’ texts on a shallow, superficial level, leaving authorial features intact on deeper levels. This result might also be of interest for research into e.g. modern newspaper articles when trying to detect editorial interventions. However, it would not be safe to conclude this contribution without the commonplace that future research is necessary to further investigate this claim.

5 Acknowledgements

Mike Kestemont gratefully acknowledges the generous support from the Research Foundation - Flanders. Both authors would like to thank W. Daelemans, F. Willaert, K. Luyckx as well as the anonymous reviewers, whose comments all helped to improve on this paper.

References

[1] J.F. Burrows. Delta: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287, 2002.

[2] G. Claassens and D. Johnson, editors. *Dutch romances. 1: Roman van Walewein*. Brewer, Cambridge, 2000.

[3] W. Daelemans, J. Zavrel, K. Van der Sloot, and A. Van den Bosch. Timbl: Tilburg memory based learner, version 6.1, reference guide. Technical report series, ILK Research Group, 2007.

[4] Walter Daelemans and Antal van den Bosch. *Memory-Based Language processing*. Cambridge University Press, 2005.

[5] K. Van Dalen-Oskam and J. Van Zundert. Delta for middle dutch: Author and copyist distinction in walewein. *Literary and Linguistic Computing*, 22(3):345–362, 2007.

[6] M. Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the Twentieth International Conference on Computational Linguistics (COLING 2004)*, pages 611–617, Geneva, 2004. Coling 2004 Organizing Committee.

[7] J. Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270, 2007.

[8] Kim Luyckx and Walter Daelemans. Shallow text analysis and machine learning for authorship attribution. In *Computational Linguistics in the Netherlands 2004: Selected papers from the Fifteenth CLIN Meeting*, pages 149–160, 2005.

[9] Kim Luyckx and Walter Daelemans. Authorship attribution and verification with many authors and limited data. In *Proceedings of the Twenty-Second International Conference on Computational Linguistics (COLING 2008)*, pages 513–520, Manchester, UK, 2008. Coling 2008 Organizing Committee.

[10] Maria C. Sherwood-Smith. *Studies in the reception of the ‘Historia Scholastica’ of Peter Comestor*. The Society for the Study of Medieval Languages and Literature, Oxford, 2000.

[11] M. Spencer and C.J. Howe. How accurate were scribes? a mathematical model. *Literary and Linguistic Computing*, 17(3):311–322, 2002.

[12] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4):471–495, 2000.

[13] Jacob Thaisen. Overlooked variants in the orthography of british library, additional ms 35286. *Journal of the Early Book Society*, 11:121–143, 2008.

[14] H. Van Halteren, R.H. Baayen, F.J. Tweedie, M. Haverkort, and A. Neijt. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77, 2005.

[15] G.U. Yule. *The Statistical Study of Literary Vocabulary*. Cambridge, 1944.