Computational approaches to textual variation in medieval literature

van Dalen-Oskam, Karina
karina.van.dalen@huygensinstituut.knaw.nl
Huygens Instituut, Netherlands

Thaisen, Jacob
thaisen@ifa.amu.edu.pl
Adam Mickiewicz University, Poznań, Poland

Kestemont, Mike
mike.kestemont@gmail.com
CLiPS Computational Linguistics group and Institute for the Study of Literature in the Low Countries (ISLN), University of Antwerp, Belgium

---

Before the age of printing, texts were copied manually only. This was done by scribes – persons who made a copy of a text for their own use or for the use of others. Often, the original is no longer available. All that remain are copies of the text, or copies of copies of copies. We know that scribes made mistakes, and that they changed spellings and wording according to what they thought fit for their audience. And we know that they sometimes reworked the text or parts thereof.

Up till now, these insecurities may have made medieval texts less interesting to work on for digital humanists. However, the complex world of medieval textual copying is a very challenging topic in its own right. Recently, some scholars have tried to develop and apply digital methods and techniques to gain insight in manual text transmission. In this session, they will explain which specific research questions led to their approach, and why traditional methods did not suffice. Then they will describe the digital approach they developed, how they gathered their data, and present the first results. They will sketch the next steps for their research and reflect on which larger questions may come closer to an answer, and which other areas of digital humanities will benefit from this research.

The first paper (by Jacob Thaisen) will focus on how the variability of spelling characteristic of Middle English makes probabilistic models a powerful tool for distinguishing scribes and exemplars. The second paper (by Karina van Dalen-Oskam) goes into vocabulary and frequencies of parts of speech, as a means to get insight in the influence scribes exerted on the texts they copied. The third paper (by Mike Kestemont) aims at erasing or minimizing textual differences in order to assess stability and the persistence of authorial features of manually copied medieval texts.

Probabilistic Modeling of Middle English Scribal Orthographies

Jacob Thaisen
thaisen@ifa.amu.edu.pl
Adam Mickiewicz University, Poznań, Poland

With the Norman Conquest of 1066 written English ceased to be employed for administrative and other official purposes, and the normative spelling conventions established for the West Saxon variety of Old English fell into disuse. When the language eventually regained these crucial domains around three centuries later, a norm for how to spell English no longer existed. The only models available to scribes were the practices of other languages known to them or, increasingly as English strengthened its position, the conventions adopted in the exemplars from which they copied. As a result of the interaction of all these factors, Middle English—the English of the period from the Battle of Hastings to Caxton's introduction of printing from movable type in 1476—is characterized by considerable variation in spelling, even within the output of a single individual. There is nothing at all unusual about one and the same scribe of this period representing one and the same word in more than one way, including very frequent words such as the definite article and conjunctions. Moreover, scribes could use the variability to their advantage in carrying out the copying task,

for example, to adjust the length of lines or speed up the copying process.

The variability of Middle English orthography means it would be misguided to assume that two texts penned by a single scribe necessarily follow, or should follow, identical spelling conventions. They are much more likely to exhibit variation within bounds. Any stylometric attribution of Middle English texts to a single scribe or of portions of a text in a single scribal hand to different exemplars on the basis of spelling must take this nature of the evidence into account. The probabilistic methods known from statistically-based machine translation, spell-checking, optical character recognition, and other natural language processing applications are specifically designed to recognize patterns in "messy" data and generalize on the basis of them. It is the purpose of this paper to demonstrate that this property of these methods makes them adequate stylometric discriminators of unique orthographies.

The methodologies developed in connection with the preparation of *A Linguistic Atlas of Late Medieval English* (McIntosh, Samuels, et al. 1986) separate unique orthographies by manual and predominantly qualitative means; if quantitative data are collected at all, they are subjected only to simple statistical tests. Since texts differ lexically, they are not readily comparable in all respects. The *Atlas* solution is to generate comparability by restricting the investigation to the subset of the respective lexicons of the various texts they may reasonably be expected to share, such as function words. Spelling forms for these words are collected from samples of the texts by selective questionnaire and any pattern present in their distribution detected by visual inspection. The forms are further often analyzed by reference to known dialect markers. The latter translates as the researcher relating the forms to phonological and morphological variables, although there is recognition in the dialectological literature that geographic significance too may characterize other levels of language.

However, it is now practically feasible to estimate the full orthography of which a given text is a sample by building probabilistic models. The reason is that recent years have witnessed an increase in the amount of diplomatically transcribed manuscript materials available in digital form, which makes it possible to abandon the qualitative focus. Scholars are already subjecting the lexical variation present in similar materials to sophisticated computer-assisted quantitative analysis (Robinson 1997, van Dalen-Oskam and van Zundert 2007). Their studies point the way forward.

The building blocks of Middle English orthographies are not individual letters but sequences of letters of varying length which, further, combine with one another in specific ways, with phonograms, morphograms, and logograms existing side by side. Every Middle English orthography has a slightly different set of building blocks, making *n*-gram models a good type of probabilistic model for capturing the distinct properties of each. Such a model is simply an exhaustive listing of grams (letters and letter sequences), each with its own probability and weight.

"Perplexity" expresses how well a given model is able to account for the grams found in a text other than the one from which the model itself is derived. That is, a model – itself a list of grams – is compared with a list of the grams found in another text and the measure simply expresses the level of agreement between the two lists. However, to find out whether the two texts are instances of the same orthography, a better model is a model not of the text from which it is derived but of the orthography which that text is a sample of. This is because the lexis of the text means the probabilities of the grams are not those they have in the orthography. This skew can be reduced by generalizing the model. "Smoothing" refers to the act of (automatically) introducing weights to achieve the best possible generalization.

Chen and Goodman (1998) carry out a systematic investigation of how a range of smoothing techniques perform relative to one another on a variety of corpus sizes in terms of the ability to account for test data. Their data come from present-day English and their basic unit is the word rather than, as here, the letter. They find the technique developed by Witten and Bell (1991) consistently to generalize the least effectively, and that developed by Kneser and Ney (1995), and later modified by Chen and Goodman (1998), consistently to do so the

most effectively. Both weight every (*n*-1)-gram in proportion to the number of different *n*-grams in which it occurs in the training data, i.e. in the text on which the model is based in the present case. The former technique produces the effect that the probability mass is shifted toward those grams which best characterize the training data, making it appropriate if the purpose is to distinguish orthographies within the product of a single scribe. The latter does the opposite, thus more fully capturing the full range of forms accepted by the scribe of the training data; this makes it the better choice if the purpose is to compare texts by a range of scribes in terms of their orthographic similarity.

To demonstrate the adequacy of smoothed models as discriminators of Middle English orthographies, the presenter investigates two corpora by means of the *SRI Language Modeling Toolkit* (Stolcke 2002):

1. The copy of Geoffrey Chaucer's unfinished poem *Canterbury Tales* contained in Cambridge University Library, MS Gg.4.27 [Gg]; the copy is in a single scribal hand.

The Gg text of is divided into equal-sized segments, each of which is subsequently modeled (Witten-Bell smoothing). For every model, its perplexity is computed against every segment other than the segment on which it is based, giving a 19x19 matrix with one blank cell per row. The mean and standard deviation is calculated for each row.

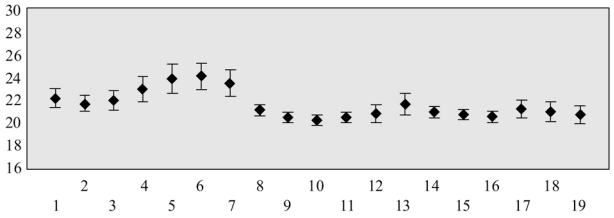

Figure 1

As can be seen in Figure 1, the greatest change in mean perplexity between any two consecutive segments falls between sections 7 and 8, with only that change falling outside the confidence intervals indicated by the whiskers. The hypothesis that two distinct populations are present is confirmed statistically (Mann-Whitney, U = 84, n1 = 7, n2 = 12, P < 0.001, two-tailed): thus, Gg contains two orthographies, their boundary falling around segments 7 and 8. The manuscript contains physical evidence of a change of exemplar late in segment 7 (Thaisen forthcoming).

2. 58 pre-1500 manuscript copies of the *Miller's Tale* and all fifty-eight such copies of the *Wife of Bath's Prologue*, totaling 116 texts; a range of scribes are responsible for these copies.

The Toolkit builds a model of every text and smoothes them (Kneser-Ney modified). For every model, its perplexity is calculated with respect to every text. The perplexities are arranged into two matrices for hierarchical clustering, one for the Miller-based models and another for the Wife-based models.

It is found, firstly, that the two trees are virtually mirror images of one another; secondly, that a Miller text and a Wife text which come from the same manuscript usually appear as sisters and that the cases in which they fail to do so are attributable to a change of scribe or exemplar posited on outside evidence (Thaisen 2009).

These results are sufficiently encouraging to warrant the investment of further resources. They show that probabilistic modeling offers a repeatable, quantified means of measuring the level of similarity between Middle English orthographies and so, also, a tool for separating them. That separation is important not only in authorship attribution and textual criticism, but also in manuscript studies and English historical linguistics. Additional advantages over the *Atlas* methodologies, which focus on dialect rather than textual studies, are the level of exhaustiveness, since all the available data are considered, as well as simple ease, the input being an unlemmatized transcript in plain text format.

References

Chen, S. F. and Goodman, J. T. (1998). *An Empirical Study of Smoothing Techniques for Language Modeling*. *Technical report TR-10-98*. Harvard University. `http://research.microsoft.com/en-us/um/people/joshuago/publications.htm` (accessed 11 March 2010).

Kneser, R. and Ney, H. (1995). 'Improved Backing-off for m-Gram Language Modeling'. *Proceedings of the IEEE International*

Conference on Acoustics, Speech and Signal Processing. , pp. 181-84.

McIntosh, A., Samuels, M. and Benskin, M. (1986). *A Linguistic Atlas of Late Mediaeval English*. Aberdeen: Aberdeen University Press.

Robinson, P. (1997). 'A Stemmatic Analysis of the Fifteenth-Century Witnesses to the Wife of Bath's Prologue'. *The Canterbury Tales Project Occasional Papers: Vol. II*. Blake, N. F. and Robinson, P. (ed.). London: Office for Humanities Communication, pp. 69–132.

Stolcke, A. (2002). 'SRILM: An Extensible Language Modeling Toolkit'. *Proceedings of the 7th International Conference on Spoken Language Processing*. Hansen, J. and Pellom, B. (ed.). Denver: Casual Productions, pp. 901–04.

Thaisen, J. (2009). 'Statistical Comparison of Middle English Texts: An Interim Report'. *Kwartalnik Neofilologiczny*. **56**: 205–21.

Thaisen, J. (forthcoming). 'A Probabilistic Analysis of a Middle English Text'. *Digitizing Medieval and Early Modern Material Culture*. Nelson, B. and Terras, M. (ed.). Tempe: Arizona Center for Medieval and Renaissance Studies.

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

Witten, I. H. and Bell, T. C. (1991). 'The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression'. *IEEE Transactions on Information Theory*. **37**: 1085–94.

Distinguishing medieval authors and scribes

Karina van Dalen-Oskam
karina.van.dalen@huygensinstituut.knaw.nl
Huygens Instituut, Netherlands

We know that medieval scribes (women or men manually copying texts) changed the texts they were copying. Scribes not only made mistakes, but also deliberately changed a text's spelling and wording. We also know that they sometimes changed the content of a text, leaving out episodes and adding others, or for instance changing the moral message. In these cases, scholars may want to describe a text not as a copy of another text but as an adaptation of it. It is not exactly clear, though, when to call a text a copy or an adaptation, and how often scribes chose to adapt instead of 'just' copy a text. It is also uncertain if all copies/adaptations of a medieval text survived and how many were lost (and why). To make things even more complicated, the exact date of most medieval manuscripts is uncertain. And in many cases we do not know the identity of the author of the original text or of the scribes.

1. Research questions

One of the tasks of scholars of medieval literature is to analyse the adaptations in a copy and to try to explain them in a poetical, ethical, or political context, which is of course difficult if the original version of a text is not extant. However, comparing different copies/adaptations of the same text usually presents scholars with enough data to make relevant observations and draw at least some conclusions. Until now, the depth of the analyses was limited to what the human eye and a scholar's amount of research time allowed. However, digital texts and digital text analysis tools can help us to compare texts in many more aspects than was possible up till now.

The questions that are of interest to us are: can we compare manual copies of the same text (semi-)automatically and get insight into the divergences which occur? Can we filter out differences that have to do with language development? Can we filter out the influence of subsequent scribes of a text and focus on those aspects which show us the original author most clearly? If so, could we apply (adapted versions of) authorship attribution tools to medieval texts? Could we also distinguish scribes from each other and are they distinguishable in the same or in a different way from how authors can be distinguished from each other? Can we develop new tools or fine-tune existing tools for scribal measurements? And can these measurements decide if a text is a copy or an

adaptation and if so, how radical the adaptation is?

Up till now, scholars hardly ever tried to systematically answer these questions. The necessary amount of work seemed not proportionate to the possible results as long as there still was enough low-hanging fruit in the close-reading type of analysis of text adaptation. Possibly, scholars in the course of time have been trained to NOT ask these impossible-to-answer questions, although two topics have always had a special place in the humanities: building family trees of manuscripts (stemmatology) and authorship attribution based on traditional, close-reading and simple statistical methods. This shows there has always been a keen interest in new and complex methods when they could possibly answer pressing questions.

2. Data

We would like to introduce two methods which may help scholars to gain insight in the amount of differences between copies of the same text. For this research, we are not interested in mere spelling differences but in more content-related differences. Our area of research is Middle Dutch literature. The first method is a rather simple approach to the vocabulary of all the copies, for which we needed lemmatization of the data, and the second is the comparison of part of speech frequencies in the copies of the same text, which implied a dataset tagged for Part of Speech. A corpus answering these needs was not available yet, so we had to create one first.

Not many Middle Dutch texts are extant in a substantial number of copies. We chose a work by the Flemish author Jacob van Maerlant: the *Rijmbijbel* ('Rhyming Bible'), which is a translation/adaptation of the Medieval Latin *Historia scholastica* written by Petrus Comestor. Van Maerlant finished this work in 1271, and many fragments and fifteen near-complete manuscripts (though not all containing all parts of the text) survive, dating from ca. 1285 to the end of the fifteenth century. One of these manuscripts is available in a good edition; it is also digitally available lemmatized and tagged for parts of speech. Transcriptions of the other manuscripts had to be made for this research. Because of the length of the texts (almost 35,000 lines), we had to work with samples. We chose 5 samples of 200-240 lines from different parts of the text, and transcribed the parallel texts (if available) from all 15 manuscripts, lemmatized the samples and tagged them for parts of speech. The manuscripts are indicated by the letters A, B, C, D, E, F, G, H, I, J, K, L, M, N and O. The lemmas we added to the transcribed texts have the form of the Modern Dutch dictionary entry (or the form the Modern Dutch entry would have had, had the word survived into present-day Dutch). We differentiated between ten parts of speech: noun, proper name, adjective, main verb, copula / auxiliary verb, numeral, pronouns, adverb, preposition, conjunction.

3. Methods

We approached the samples as 'bags of words'. We made use of perl scripts listing all lemmas and parts of speech for each small sample and for the frequency measurements. For each part of speech, in each sample we measured the absolute frequency, the relative frequency, the average of the fifteen samples, the standard deviation, the z-score and the ranking of the manuscript in comparison with the other fourteen manuscripts.

4. First results of the comparison of vocabularies

Table 1 below lists the amount of words (lemmas) each text episode has which do not occur in any of the other copies of the same episode: unique words. The manuscripts A - O are listed in chronological order (although many dates are very approximate). The order from left to right agrees with the order of the episodes in the text itself.

Unique	Eva 'E'	Debora 'D'	Judit 'J'	NT 'M'	Josephus 'T'	Total
C	7	4	2	0	10	23
B	0	0	2	2	3	7
M	3	5	0	3	1	12
A	5	3	1	1	2	12
G	7	6	6	11	8	38
D	0	1	2	3	3	9
L	0	0	0	1	1	2
K	4	5	1	2	2	14
E	8	26	24	10	57	125
F	5	4	3	6	1	19
J	9	7	4	2	5	27
N	3	5	6	8	2	24
H	13	5	9	5	n.a.	32
O	2	5	7	n.a.	n.a.	14
I	13	28	117	n.a.	n.a.	158
Total	79	104	184	54	95	516

Table 1

At a glance we can see that manuscripts E and I show the most unique words. This needs to be investigated: are these scribes the most radical in their adaptations? We will try to push this use of vocabulary analysis further, e.g. measuring the percentage of overlapping vocabulary in the episodes in the different manuscripts.

5. First results of the comparison of PoS frequencies

Figure 2 shows the relative frequencies of nouns for each of the five samples (E, D, J, M, T) for each of the manuscripts (A - O). For this part of speech, we see a big change in the trend for the episode in all of the manuscripts for episode T in manuscript E and for episode J in manuscript I. Graphs for other parts of speech show the same trend.

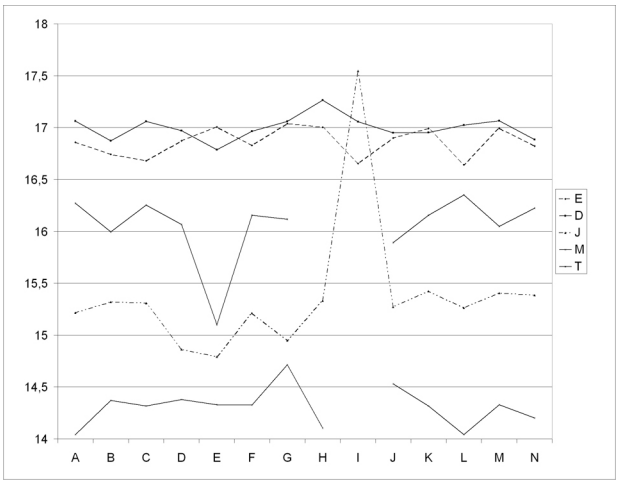

Figure 2

6. Evaluation

Reading the two episodes and comparing them with the other copies of the same text reveals that episode J in manuscript I clearly is a new text. Almost all the rhyme words have been changed, text is added, wording is completely different. The scribe here acted as a new author. This is not the case for episode T in manuscript E. Some rhyme words have changed, some wording is different, but the text is still clearly recognizable as a copy.

7. Conclusions and next steps

It seems simple comparisons of vocabulary and frequencies of parts of speech can pinpoint scribes who did more than copying their exemplar. If this can be confirmed by other experiments, this approach could help to direct scholars to those episodes in texts that may be most rewarding for a closer analysis (e.g. with traditional methods such as close reading). Before this is possible, however, a lot more medieval texts need to be available tagged for headwords and for part of speech. For that, a good tagger for Middle Dutch is highly desirable.

The research questions addressed above are key for getting a better insight into the cultural role of texts and the persons responsible for texts and their transmission, not only in the Middle Ages, but also later. It could help us to find a way to less subjectively compare texts and describe scribal adaptations, and in this way yield insight in the possible kinds of text manipulation throughout the ages.

References

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

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

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

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

Spencer, M. and Howe, C. J. (2002). 'How accurate were scribes? A mathematical

model'. *Literary and Linguistic Computing*. **17**: 311-322.

The Robustness of Rhyme Words in Bypassing Scribal Variation for Medieval Authorship Verification

Mike Kestemont
mike.kestemont@gmail.com
CLiPS Computational Linguistics group and Institute for the Study of Literature in the Low Countries (ISLN), University of Antwerp, Belgium

1. A problem

Modern stylometric approaches can discriminate between authors to a fairly accurate extent. Machine learning techniques, for instance, are able to 'recognize' authors – be they literary or not – based on linguistic features extracted from representative textual samples written by these authors. In recent years, computational studies into authorship issues (such as verification and attribution) have been a popular topic in many research areas. Nevertheless, this kind of research has paid rather little attention to medieval literature. This is especially remarkable since it is precisely in this branch of philology that scholars have to contend with large amounts of texts of which the authorship is unknown or at least disputed. This lack of interest in medieval literature seems due to a variety of factors that all come down to the same basic fact: medieval texts are difficult to automatically process. For example, tools to perform basic actions such as the automatic lemmatization of texts are virtually non-existent for medieval languages, while most stylometric approaches heavily rely on e.g. lemma-frequencies for their feature extraction (cf. Burrows's *Delta*). This is mainly due to the scribal variation that is so typical of medieval manuscripts, as put forward in the introduction to this session proposal. Moreover, medieval texts are rarely extant from autographs and as such, in the majority of the cases, scholars have a hard time assessing which features in manuscriptal copies are *authorial* rather than *scribal*.

2. A solution

Any solution to the problem of authorship attribution for medieval texts has to overcome the difficulties imposed by the (scribal) instability of text in the Middle Ages. Whereas the other two papers in this session focus on the *exploitation* of scribal variation (i.e. textual *in*stability), this paper aims at the exact opposite: *erasing* or *minimizing* these differences in order to assess textual stability and the persistence of authorial features of manually copied medieval texts. In this paper we shall focus on two methods to achieve this goal. Firstly, we shall briefly discuss the MiDL-architecture, a Natural Language Processing system designed for the automated tokenization, lemmatization and part-of-speech tagging of Middle Dutch literary texts. The techniques allow us to get past or 'transcend' superficial scribal variation and focus on the underlying authorial features of texts.

Secondly, we shall report on experiments with rhyme words and pairs – Middle Dutch epic literature was rhymed in about 99% percent of the cases that are currently known to us. This category of words is often claimed to be a very stable factor during the process of text transmission and thus can be expected to be extremely revealing with regard to authorial style. Scribes could not easily alter the rhyme words or rhyme scheme of a text without having to adapt several lines of the text and would often refrain from doing this.

3. A good case study

In authorship related studies, it is often hard to set up an experiment that is entirely 'clean' or 'sterile' from a methodological point of view. If one has to make sure that the *only* difference between two texts is the difference in authorship, one has to keep all other factors (such as gender and education level of the author, topic of the text, ...) as stable as possible over the two texts compared. For the Middle Ages, the poor survival of texts makes it difficult to set up an experiment that fully meets these requirements. For this paper, we shall work on

a single case study that does seem to approach this ideal setup as much as possible: the *Spiegel Historiael*, a Middle Dutch adaptation of the Latin *Speculum Maius*, by Vincentius of Beauvais. This adaptation was initiated by the influential writer Jacob of Maerlant and was later continued by two other authors: Filip Utenbroeke and Lodewijk van Velthem. Of each of these authors near-complete manuscript copies survive of substantial parts of their contribution to the project, called *Partien*. Each of these *partien* is divided in larger units called *books*, which in turn consist of smaller *chapters*. In this study we shall focus on the first *partie* by Maerlant (31K lines in 532 chapters), the second by Utenbroeke (41K lines in 461 chapters) and the fifth by Velthem (27K lines in 387 chapters). These chapters will be our main comparison unit. What makes this *Spiegel historiael* such an interesting case is that comparing these texts for authorial differences indeed keeps many other factors rather constant, such as level of education, gender, genre, etc.

4. Preprocessing

As a starting point, we shall briefly discuss the architecture which we have developed for the preprocessing of our texts: the MiDL-system (joint work with Walter Daelemans & Guy de Pauw of the Antwerp Computational Linguistics group, CLiPS). The MiDL-system performs tokenization, lemmatization and Part-Of-Speech tagging for Middle Dutch literary texts. The technology we present is optimized for this specific material but should scale well to other medieval languages (or any resource-scarce language characterized by a lot of spelling variation). In this contribution we shall focus on the *corpus-Gysseling* (CG), a corpus that was digitized and semi-automatically annotated at the Institute for Dutch Lexicology (INL). More specifically we shall report on results with the so-called 'literary part' of this corpus (ca. 600K running tokens) that contains all Middle Dutch literature, surviving from manuscripts dated between 1200 and 1300AD. The main issue we will discuss is lemmatization, as we will argue that this step is actually the key to all subsequent operations (such as e.g. PoS-tagging or shallow parsing).

Lemmatization refers to the process whereby natural language tokens are assigned a 'lemma'. The basic purpose of doing this – in any language or research domain – is that it enables the generalization 'about the behaviour of groups of words in cases where their individual differences are irrelevant' (Knowles & Mohd Don 2004:69). Hence, lemmatization can be considered a problem of mapping many-to-one: similar tokens are mapped to the same 'abstract representation' that, as such, comes to subsume 'all the formal lexical variations which may apply' (Crystal 1997). There exists an obvious parallel with the lexicographer's activity of grouping words under the same 'dictionary headword' (Ibid.; Knowles & Mohd Don 2004:70). When it comes to medieval languages, the main issue that is to be dealt with are historical spelling variants (HSV). When compared to the problem of lemmatization in modern languages, it adds a level of complexity:

Modern
LEMMA X = {$token^1$, $token^2$, ..., $token^{n-1}$, $token^n$}X
Medieval
LEMMA X = {$token^1$={$variant_1{}^1$, $variant_2{}^1$}, $token^2$={$variant_1{}^2$, $variant_2{}^2$} ..., $token^{n-1}$={$variant_1{}^{n-1}$, $variant_2{}^{n-1}$}, $token^n$={$variant_1{}^n$, $variant_2{}^n$}}

The main purpose of lemmatization, as such, lies with a form of token normalization that allows us to transcend superficial spelling variations.

5. Experiments

In this paper we shall focus on lexical features (n-grams of lemmata) and shallow morpho-syntactic features (n-grams of PoS-tags). Our main research emphasis will be on rhyme words. We will present the results of leave-one-out validation on our data as following. Using a machine learning algorithm, we will do experiments on *m* samples by Maerlant, *n* samples by Utenbroeke and *l* samples by Velthem (with the sample size set to individual chapter entities). During each fold, we will each time 'leave out' one chapter by one author (e.g. one by Utenbroeke) and train on the chapters that are left (e.g. *m* Maerlant-samples, *n-1* Utenbroeke-samples and *l* Velthem-samples). We will 'test' each time the accuracy of our algorithm after training in terms of accuracy: 'Can the learner correctly identify by which author the omitted sample was written?'

References

Biemans, J.A.A.M. (1997). *Onsen Speghele Ystoriale in Vlaemsche. Codicologisch onderzoek naar de overlevering van de Spiegel historiael van Jacob van Maerlant, Philips Utenbroeke en lodewijk van Velthem, met een beschrijving van de handschriften en Fragmenten*. Leuven: Peeters.

Crystal, D. (1997). *A Dictionary of Linguistics and Phonetics*. Oxford: Oxford University Press.

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

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

Knowles, G. and Mohd Don, Z. (2004). 'The notion of a "lemma". Headwords, roots and lexical sets'. *International Journal of Corpus Linguistics*. 69-81.

Luyckx, K. and Daelemans, W. (2008). 'Authorship Attribution and Verification with Many Authors and Limited Data'. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*. Manchester, 2008, pp. 513-520.