A Stylometric Analysis of King Alfred's Literary Works

Paramjit S. Gill
Department of Mathematics and Statistics
Okanagan University College
Kelowna, BC, Canada, V1V 1V7
E-mail: *pgill@ouc.bc.ca*

Tim B. Swartz
Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, BC, Canada, V5A 1S6
E-mail: *tim@stat.sfu.ca*

Michael Treschow
Department of English
Okanagan University College
Kelowna, BC, Canada, V1V 1V7
E-mail: *MTreschow@ouc.bc.ca*

Abstract

For centuries, Alfred the Great was judged to have translated several Latin texts into Old English. Many scholars, however, have expressed doubt whether Alfred could have done all of this work. With the availability of the Old English Corpus in electronic form, it is feasible to subject the texts to statistical stylometric analysis. We approach the problem from a Bayesian perspective where key words are identified and frequencies of the key words are tabulated for seven relevant texts. The question of authorship falls into the general statistical problem of classification where several simple innovations to classical agglomerative procedures are introduced. Our results suggest that one translation that has been traditionally attributed to Alfred (*The First Fifty Prose Psalms*) tends to distinguish itself from texts that are known to be Alfredian.

Key words and phrases: Agglomerative techniques, Bayesian methods, Classification, Dirichlet distribution, Disputed authorship, Entropy, Hierarchical clustering, Multinomial distribution, Old English.

1 INTRODUCTION

After King Alfred the Great defeated the Vikings at the battle of Edington in 878, he turned to strengthening his English kingdom of Wessex that had suffered so greatly under the Viking invasions. Most famous was his program for educational reform. Alfred depicted himself as a philosopher-king, taking up scholarship and making English translations of Latin patristic texts to serve as the basis of education in the English language. Seven translations are associated with his reign. The following three translations internally identify themselves as Alfred's work:

1. Gregory the Great's *Pastoral Care*

2. Boethius's *The Consolation of Philosophy*

3. Augustine's *The Soliloquies*

The other four translations are:

4. Gregory the Great's *Dialogues*

5. Bede's *Ecclesiastical History of the English People*

6. Orosius's *History against the Pagans*

7. *The First Fifty Prose Psalms*

Of these four, only Gregory's *Dialogues* clearly identifies itself as not Alfred's work. Alfred wrote the preface of Gregory's *Dialogues* explaining that he had directed his friends to carry out the translation. Texts 5, 6 and 7 do not identify any translator but have been traditionally attributed to King Alfred. William of Malmesbury, a twelfth century historian, listed Bede's *History* and Orosius's *History* among Alfred's translations. William of Malmesbury also stated that Alfred was working on a translation of the *The First Fifty Prose Psalms* at the time of his death. Old English scholars, however, have come to accept that Alfred could not have translated Bede's *History* because, like the translation of the *Dialogues*, it shows traces of the Mercian dialect (Whitelock 1966). Alfred's alleged translation of Orosius's *History against the Pagans* has also been overthrown on the grounds of extensive differences in vocabulary and sentence structure from the other Alfredian texts (Bately 1980).

With respect to text 7, *The First Fifty Prose Psalms*, Bately (1982) continues to side with tradition, and argues that the translation is due to Alfred. Bately (1982) assessed the authorship of Orosius's *History against the Pagans* and the *The First Fifty Prose Psalms* by analysing the translation of certain words. She noted that the *The First Fifty Prose Psalms* frequently used the same Old English words to translate corresponding Latin words as did the three known Alfredian texts but that Orosius's *History against the Pagans* showed greater differentiation.

In summary, current opinion is nearly unanimous that texts 1, 2 and 3 are Alfredian and that texts 4, 5 and 6 are non-Alfredian. However, the translation of text 7 is less certain. The question arises as to whether a more thorough stylometric analysis would confirm Bately's conclusion that the translation of text 7, *The First Fifty Prose Psalms* is due to Alfred. Stylometry (Holmes and Kardos 2003) allows for a more extensive analysis, not only of contextual words, but also, and more importantly, of non-contextual words.

As we are dealing with Old English translations from the original Latin, we face a special challenge that differs from standard stylometric analysis where the problem is the authorship assignment of original work. It is important to note, however, that the work of translation is itself a kind of authorship that can be subjected to stylistic analysis. The translations considered in this study were initiated at the beginning of Old English prose writing and show the initial development of English prose style. The proem to the translation of Boethius states that Alfred's strategy of translation was variable, sometimes rendering "word for word, sometimes sense for sense." All these translations exhibit an authorial voice that forms the text into the Old English language.

The question of authorship falls into the general statistical problem of classification where objects that are "similar" are grouped together in clusters. The literature on classification and clustering is vast; a selection of textbooks on classification and clustering include Hartigan (1975), Kaufman and Rousseeuw (1990) and Gordon (1999). We note that our problem is not one of discrimination since although we know that texts 4, 5 and 6 are non-Alfredian, we do not know whether any of these texts have common translators (i.e. we do not have full knowledge of the number of underlying clusters nor the membership of all reference objects to clusters).

There are special features in our problem that have lead to simple innovations of the widely used agglomerative algorithm proposed by Lance and Williams (1966). An agglomerative approach is convenient for our application as it allows the use of prior knowledge concerning the authorship of particular texts. For example, we begin by assuming that texts 1, 2 and 3 are Alfredian and we can therefore initiate the agglomerative algorithm with a cluster containing these three texts.

A long-standing difficulty with agglomerative algorithms, and hierarchical clustering methods in general, is the determination of a stopping criterion for merging clusters (Mojena 1977). Without a stopping criterion, agglomerative algorithms continue until all objects are grouped in the same cluster. We use the assumption that texts 4, 5 and 6 are non-Alfredian to determine a relevant stopping criterion. Such an approach may be useful in other clustering applications where some clustering information is known apriori. Fraley and Raftery (1998) use the Bayesian information criterion (BIC) in a model-based clustering approach to determine the number of clusters.

Another innovation in our methods results from viewing the frequencies of the key words as samples from underlying multinomial distributions. This implies that there is uncertainty in the dissimilarity measures between objects (translations in our application). We wish to account for the uncertainty by being able to express probabilities associated with clusters rather than identifying a unique partitioning of objects as is often the case. Although methods of fuzzy analysis (Dunn 1977) provide membership coefficients for individual objects, classical (as opposed to model-based) agglomerative approaches do not provide probability assessments for clustering. Clustering probabilities are sometimes available using Bayesian mixture models (see Liu, Zhang, Palumbo and Lawrence (2003) and the references therein). However, these methods typically rely on Markov chains which often require fine tuning to promote mixing in the Markov chain. In our approach, probability assessments arise from a Bayesian model where posterior probabilities of clusters are obtained via straightforward simulation from the posterior. We anticipate that our simple approach may have value in various clustering applications.

Another nonstandard aspect of our approach is that the variables used for calculating dissimilarities are nonstandard (e.g. neither interval-scaled, binary, ordinal, etc). The variables used in our application constitute probability distributions on a finite set. For this, we propose a dissimilarity measure based on entropy.

A final innovation of our approach pertains specifically to problems of stylometry. In stylometry, a subject matter expert is typically assigned with the problem of choosing non-contextual key words. This is a critical component of the analysis where the incorrect choice of contextual key words leads to less clustering. In many problems of stylometry, the choice of non-contextual key words is regarded as a given. As a by-product of our proposed algorithm, we provide a test of whether key words are non-contextual.

In section 2, we describe the underlying Bayesian model used in our approach. The data (key word frequencies) are assumed to arise from multinomial distributions. The posterior distribution of the underlying parameters is obtained via straightforward simulation. In section 3, we propose two clustering methodologies based on the agglomerative algorithm of Lance and Williams (1966). The second of the two methodologies is more appropriate when the variation within clusters is believed to differ amongst clusters, and it provides a test of whether key words are non-contextual. Both methodologies provide probability assessments for clusters by averaging the partitioning results over the simulation. In section 4, the methods are applied to the seven texts from the Alfredian period and the results are reported. A concluding discussion is provided in section 5.

2 STATISTICAL MODEL

The raw data for this study were obtained from the Dictionary of Old English Corpus (Healey 2000). A principle of stylometric analysis is that authors/translators use high-frequency words unreflectively in their writings. These key words occur regardless of context, and hence, differential rates of usage form a basis for distinguishing authorship. Key words are typically prepositions, conjunctions, articles and common verbs. Clearly, the choice of key words ought to be determined by a subject matter expert. For analysing the seven texts, we generated a list of the 100 most frequent words. We refined the list by omitting all contextual words. We further omitted all words that might depend on the original Latin text and chose those words that were distinctively English and expressive of English style. Table 1 shows the list of the $K - 1 = 17$ individual key words (with the modern English meaning in parentheses). Multiple spellings for many of these words were accounted for and combined. Given the selection of key words, frequency counts of the key words were obtained for each of the $n = 7$ texts using WordSmith Tools (Scott 1998).

Let X_{ik} denote the frequency of key word k in the i-th text, $k = 1, \ldots, K-1$, $i = 1, \ldots, n$ where X_{iK} is the number of non-key words in the i-th text. This gives rise to the model

$$X_i = (X_{i1}, \ldots, X_{iK}) \sim \text{Multinomial}(m_i, p_{i1}, \ldots, p_{iK})$$

where the distributions are independent over $i = 1, \ldots, n$ and $\mathbf{X} = (X_1, \ldots, X_n)$. For the application in question, the texts are comparable in size where the number of words are given by $m_1 = 67639$, $m_2 = 46212$, $m_3 = 15430$, $m_4 = 91006$, $m_5 = 77501$, $m_6 = 47475$ and $m_7 = 19398$. The multinomial distribution is a natural distribution to use in this problem and in many other classification problems. However, the multinomial distribution has been underutilized. For example, we are not aware of any model-based mixture approaches that use the multinomial distribution. In classical partitioning and classical hierarchical approaches, there is no clear choice for a dissimilarity measure between the multinomial vectors $X_1, \ldots, X_n$.

The unknown parameters in the model are the p_{ik}'s, and in a Bayesian approach, it is typical to assign independent flat priors to $(p_{i1}, \ldots, p_{iK})$ for $i = 1, \ldots, n$. This leads to the posterior distributions

$$(p_{i1}, \ldots, p_{iK} \mid \mathbf{X}) \sim \text{Dirichlet}(X_{i1} + 1, \ldots, X_{iK} + 1)$$

where the distributions are independent over $i = 1, \ldots, n$.

The posterior distributions of key word usage are used as "wordprints" of various texts. If two texts have "similar" distributions, then it is likely that the two texts were translated by the same person. Our approach to cluster analysis therefore presents itself: Since it is convenient to sample from Dirichlet distributions, we generate one set of p_{ik}'s and then apply a classification algorithm on the p_{ik}'s to determine the resultant clusters. Repeating the simulation and averaging the results provides posterior probabilities for all clusters. Note that the posterior probability assigned to a specific cluster describes the proportion of time that a given algorithm determines the cluster taking into account the variability in the data. Although we believe that this provides some useful insight, this probability is not the same as the probability that a specific cluster is a correct cluster. To obtain the latter probability, one would require more detailed modelling which describes the variation in texts due to the same author/translator.

3 CLUSTERING METHODOLOGIES

The backbone for the proposed clustering methodologies is the classical agglomerative algorithm due to Lance and Williams (1966) sometimes referred to as unweighted pair-group method of averaging (UPGMA). Like many clustering approaches, a dissimilarity measure is required to distinguish between objects. As the variables $p_{i1}, \ldots, p_{iK}$ constitute a discrete probability distribution, it is sensible to use a measure specifically designed to compare distributions. Entropy (Renyi 1961) has a long and distinguished history in the statistical sciences and is useful for comparing distributions. However, entropy is not a metric as it does not satisfy the symmetry requirement between two objects. We therefore use the entropy-based distance defined between texts i and j

$$d(i,j) = \sum_{k=1}^{K} (p_{ik} - p_{jk}) \log(p_{ik}/p_{jk}).$$

This measure is sometimes referred to as Jeffreys divergence (Jeffreys 1946).

In our problem, we begin with the assumption that texts 1, 2 and 3 are Alfredian. This assumption is easily incorporated by initiating the algorithm with the five clusters $\{1,2,3\}, \{4\}, \{5\}, \{6\}$ and $\{7\}$, where $\{i_1, \ldots, i_j\}$ denotes that texts $i_1, \ldots, i_j$ belong to the same cluster.

The algorithm proceeds by merging two of the clusters together and this procedure continues until one is left with a single cluster containing all of the n objects. In practice, it is useful to stop the algorithm at some point to determine a realistic set of clusters. We use the Lance and Williams (1966) criterion for determining the pair of clusters to be merged in a given step. Let R and Q denote a pair of clusters and let $|R|$ and $|Q|$ denote the number of objects in R and Q respectively. Then a merge is proposed between the pair of clusters for which

$$D(R,Q) = \frac{1}{|R|\ |Q|} \sum_{\substack{i \in R \\ j \in Q}} d(i,j) \tag{1}$$

is the smallest. We note that when $R = \{r\}$ and $Q = \{q\}$ are both clusters with a single object, then $D(R,Q)$ reduces to $D(R,Q) = d(r,q)$. The simplicity of the UPGMA algorithm is one of its main features, and is a primary reason why it is still in active use today.

To terminate the algorithm, we make use of the assumption that texts 4, 5 and 6 are non-Alfredian. Therefore, we stop the algorithm if a merge is proposed that brings any of 4, 5 or 6 together with a cluster containing $\{1,2,3\}$. We impose a second stopping criterion which also takes into account prior knowledge, specifically our knowledge that texts 1, 2 and 3 are Alfredian. Given that texts 1, 2 and 3 were translated by Alfred, this suggests that the distances $d(1,2)$, $d(1,3)$ and $d(2,3)$ are the sorts of dissimilarities that one might expect between texts common to a particular author/translator. Therefore we define

$$d_{\max} = \max\left(d(1,2),\ d(1,3),\ d(2,3)\right)$$

and terminate the algorithm if a merge is proposed where the minimal cluster distance in (1) is such that

$$D(R,Q) > d_{\max}. \tag{2}$$

These stopping rules are intuitive, easy to implement and preserve the simplicity of the UPGMA algorithm. Moreover, the approach may be useful in other clustering applications where some clustering information is known apriori.

In Table 2, we list the 20 partitions that are possible under our proposed algorithm. Simulating the p_{ik}'s from the posterior, determining the resultant cluster, and averaging the results under repeated simulations provides posterior probabilities for each of the 20 partitions. The posterior probability for a particular cluster is then obtained by summing the probabilities over the partitions that contain the cluster.

3.1 Nonconstant Within Cluster Variation

It is possible to imagine situations where the variation within clusters is not the same for all clusters. For example, in the analysis of literary works, it may be the case that one author/translator has a style that varies little while another author/translator may have a style that varies greatly. The stopping criterion (2) assumes that the variation within clusters is the same or comparable for every cluster.

To accommodate the possibility of nonconstant within cluster variation, we divide each of the seven texts into blocks of roughly the same size. We assume that the variation between blocks by an author/translator is the same as the variation between texts by the same author/translator. The use of non-contextual key words is instrumental in forming this assumption. Therefore, in our problem, the basic clustering objects are now blocks (ij) where $i = 1, \ldots, n$ refers to the text and $j = 1, \ldots, n_i$ refers to the block within the i-th text. The division of texts into blocks allows us to assess variation in style amongst authors/translators.

Using the assumption that texts 1, 2 and 3 are Alfredian, we therefore initiate an alternative algorithm with the five clusters

$$
\begin{aligned}
&\{\ (11),\ldots,(1n_1),(21),\ldots,(2n_2),(31),\ldots,(3n_3)\ \},\\
&\{\ (41),\ldots,(4n_4)\ \},\\
&\{\ (51),\ldots,(5n_5)\ \},\\
&\{\ (61),\ldots,(6n_6)\ \},\\
&\{\ (71),\ldots,(7n_7)\ \}.
\end{aligned}
$$

As before, a merge is proposed between clusters R and Q according to (1), and the algorithm terminates if the merge brings any of texts 4, 5 or 6 together with a cluster containing $\{\ (11),\ldots,(1n_1),(21),\ldots,(2n_2),(32),\ldots,(3n_3)\ \}$.

The alternative methodology involves a modification of the second stopping criterion (2). We now terminate the algorithm if a proposed merge between clusters R_1 and R_2 is such that

$$D(R_1,R_2) \quad > \quad \min\left(\text{quantile}_1(q),\ \text{quantile}_2(q)\right) \tag{3}$$

where $\text{quantile}_i(q)$ is the $100q$-th quantile of the entropy-based distances between the objects in cluster R_i, $i=1,2$. In this way, a newly formed cluster of the subclusters R_1 and R_2 has a mean cluster distance $D(R_1,R_2)$ which is comparable to the distances within the tightest of the subclusters. We may regard $q \in (0,1)$ as a tuning parameter of the algorithm where it is becomes more difficult to merge clusters as q decreases. For our applications, we have found that $q = 0.90$ provides an adequate default value. In practice, we may run the algorithm several times using different values of q.

As a by-product of the algorithm, we are able to test the critical assumption that the key words are non-contextual. In stylometry, a subject matter expert is typically assigned with the problem of choosing non-contextual key words, and the choice is often regarded as a given. Clearly, the incorrect use of contextual key words causes greater differentiation between texts and this can lead to less clustering. In our application, we know that texts 1, 2 and 3 are Alfredian, and therefore we can evaluate criterion (3) with respect to the blocks within the text pairs (1,2), (1,3) and (2,3). If any of the three hypothetical merges is disallowed by criterion (3), this suggests that the key words are contextual.

4 ANALYSIS OF THE TEXTS

We apply the first of our clustering methodologies where the underlying clustering objects are the seven texts. Recall that this approach assumes a constant within cluster variation for each of the clusters. We observe that in every simulation, the algorithm proceeds from partition 1 → partition 14 → partition 18 (see Table 2) and then terminates under condition (2). The reason why the algorithm always proceeds in this fashion is that the texts are large leading to Dirichlet posteriors with large parameters. Therefore the p_{ik}'s that are generated from informative Dirichlet posteriors tend to be comparable from simulation to simulation. The terminating partition 18 isolates text 7 *The First Fifty Prose Psalms* from the Alfredian texts $\{1, 2, 3\}$ and also from the non-Alfredian texts $\{4, 5, 6\}$. This provides strong evidence that *The First Fifty Prose Psalms* was not translated by King Alfred.

We now consider the second of our clustering methodologies which we prefer in the given application. It breaks the texts into smaller blocks of roughly the same size. In the order in which the texts are presented, there are 51, 39, 16, 63, 58, 40 and 17 blocks. The breaking of the texts into blocks allows us to account for variation in styles amongst the different translators.

The second clustering methodology allows us to test the important assumption concerning the non-contextuality of key words. We do this by initially separating texts 1, 2 and 3 and observe that after two steps of the algorithm, the three Alfredian texts are clustered together in every simulation. This provides strong evidence that the chosen key words exhibit the required property of non-contextuality. We note further that this was observed under various choices of q in the stopping criterion (3) (e.g. $q = 0.70, 0.90, 0.95$).

Having satisfied ourselves with the choice of key words, the algorithm was run with $q = 0.90$. Over 1000 simulations, the algorithm terminated in partition 1 (79 times), partition 18 (813 times), partition 19 (12 times) and partition 20 (96 times). Summing over partitions 1, 18 and 20, this gives that the clustering algorithm separates the disputed text 7 *The First Fifty Prose Psalms* from the Alfredian texts $\{1, 2, 3, \}$ 99% of the time. This again provides strong evidence that *The First Fifty Prose Psalms* was not translated by King Alfred. We remark that when q is reduced to 0.80, the results are similar in that *The First Fifty Prose Psalms* is separated from the Alfredian texts 100% of the time.

5 DISCUSSION

This paper proposes two clustering schemes that extend the UPGMA algorithm of Lance and Williams (1966) in several directions yet retains the simplicity of the original algorithm. In particular, probability assessments are obtained for partitions and clusters, and stopping rules have been devised which are based on partial knowledge of cluster membership. Although the proposed methods may be useful in various applications, they have immediate application to problems of stylometry. For example, a test for the non-contextuality of key words is proposed as a by-product of the second clustering scheme.

The proposed clustering schemes have been applied to texts whose translations may or may not be due to Alfred the Great. Our stylometric analysis provides strong evidence that *The First Fifty Prose Psalms* was not translated by King Alfred. This assertion contradicts both traditional thought and the conclusions reached by Bately (1982).

6 REFERENCES

Bately, J. ed. (1980). The Old English Orosius. EETS SS 6. Oxford University Press, London.

Bately, J. (1982). Lexical evidence for the authorship of the prose psalms in the Paris Psalter. Anglo-Saxon England, 10, 69-95.

Dunn, J. C. (1977). Indices of partition fuzziness and the detection of clusters in large data sets. In Fuzzy Automata and Decision Processes, edited by M. Gupta, Elsevier, New York, 271-284.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. Computer Journal, 41, 578-588.

Gordon, A. D. (1999). Classification. Chapman and Hall, London.

Hartigan, J. A. (1975). Clustering Algorithms. John Wiley and Sons, New York.

Healey, A. D. (2000). Dictionary of Old English, Old English Corpus. 2000 TEI-P3 conformant version. University of Michigan Digital Library Production Service.

Holmes, D. I. and Kardos, J. (2003). Who was the author? An introduction to stylometry. Chance, 16, 5-8.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. Lon., Ser. A, 186, 453-461.

Kaufman, L. and Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, New York.

Lance, G. N. and Williams, W. T. (1966). A general theory of classificatory sorting strategies: 1. Hierarchical systems. The Computer Journal, 9, 373-380.

Liu, J. S. Zhang, J. L. Palumbo, M. J. and Lawrence, C. E. (2003). Bayesian clustering with variable and transformation selections. In Bayesian Statistics 7, edited by J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, Oxford University Press, 249-276.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. The Computer Journal, 20, 359-363.

Renyi, A. (1961). On the measures of entropy and information. Proc. 4th Berkeley Symp. Math. Stat. and Prob., 1, 547-561.

Scott, M. (1998). WordSmith Tools Manual, version 3.0, Oxford University Press.

Whitelock, D. (1966). The prose of Alfred's reign. In Continuations & Beginnings, edited by E. G. Stanley, Nelson, London, 67-103.

7 ACKNOWLEDGEMENTS

Gill and Swartz were partially supported by research grants from the Natural Sciences and Engineering Research Council of Canada. Gill and Treschow were partially supported by a grant in aid of research from the Okanagan University College.

Table 1: List of the 17 Old English key words and modern English translations (in parentheses).

AC (but)	AND (and)	BID (is)	EAC (also)	HIT (it)
IS (is)	MID (with)	OF (of)	SWA (so)	TO (to)
DA (those, then)	DÆS (of the)	DÆT (that)	WÆS (was)	WID (against)
DONNE (then)	DEAH (although)			

Table 2: List of the 20 partitions that are possible under the proposed clustering algorithms.

1.	$\{123\}, \{4\}, \{5\}, \{6\}, \{7\}$	11.	$\{1237\}, \{46\}, \{5\}$
2.	$\{1237\}, \{4\}, \{5\}, \{6\}$	12.	$\{123\}, \{467\}, \{5\}$
3.	$\{123\}, \{47\}, \{5\}, \{6\}$	13.	$\{123\}, \{46\}, \{57\}$
4.	$\{123\}, \{57\}, \{4\}, \{6\}$	14.	$\{123\}, \{56\}, \{4\}, \{7\}$
5.	$\{123\}, \{67\}, \{4\}, \{5\}$	15.	$\{1237\}, \{56\}, \{4\}$
6.	$\{123\}, \{45\}, \{6\}, \{7\}$	16.	$\{123\}, \{47\}, \{56\}$
7.	$\{1237\}, \{45\}, \{6\}$	17.	$\{123\}, \{567\}, \{4\}$
8.	$\{123\}, \{457\}, \{6\}$	18.	$\{123\}, \{456\}, \{7\}$
9.	$\{123\}, \{45\}, \{67\}$	19.	$\{1237\}, \{456\}$
10.	$\{123\}, \{46\}, \{5\}, \{7\}$	20.	$\{123\}, \{4567\}$