Finding Fingerprints in the Fabric of Verse: Unearthing Style in Old English Poetry Using Machine Learning

Finding Fingerprints in the Fabric of Verse:

Unearthing Style in Old English Poetry Using Machine Learning

Ravindra N. Mynampaty

A Thesis in the Field of Linguistics

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

November 2017

Abstract

This study hypothesizes that poetic styles exist in Old English literature and that these styles, whether they pertain to an individual or a school of poets, can be identified and categorized by examining verse syntax, particularly the usage of auxiliary verbs and verbals. Further, this thesis attempts to determine what common features exist in the handful of poems currently accepted as having been composed by the poet Cynewulf and what other works share this style.

Scholars have previously undertaken such studies of style by investigating word-order (e.g., Bliss 1980) and by meticulously cataloging and analyzing data about auxiliaries, verbals, their types, and the environment in which they occur (e.g., Donoghue 1987). By and large they have found that classifications, and therefore styles, of Old English poems are not clear-cut.

This thesis builds upon such previous work and analyzes the same data by adopting an unsupervised machine learning approach which uses computer algorithms to cluster 19 poems into groups based on their syntactic features, and to determine if these groups represent styles based on the shared characteristics of their constituent poems.

This analysis did not find distinctive styles in the clusters that emerged from the larger corpus that was studied, indicating that Old English poetry was not monolithic but consisted of a range of styles throughout its history. In the narrower context of Cynewulfian authorship, the results were more promising and suggest that an additional poem can be considered for inclusion in his canon.

'I wish life was not so short,' he thought. 'Languages take such a time, and so do all the things one wants to know about.'

- J. R. R. Tolkien, *The Lost Road*

To Nausheen

Acknowledgments

My sincere gratitude to my thesis director, Prof. Daniel Donoghue, who counseled me patiently over the past year, was so generous with his time, and for trusting me with all those shoeboxes of index cards with the auxiliary-verbal data. It has been my honor to work with this verb-*hord* he painstakingly gathered so many years ago.

My thanks also go to my research advisor, Dr. Stephen Shoemaker, for his guidance and encouragement. I am a better writer for having taken his proseminar.

I have had the privilege of studying with so many excellent teachers at Harvard—the names Fortson, Nevins, and Witzel come to mind—who teach for spreading knowledge for its own sake. They instilled in me a great appreciation for human languages and for the field of Linguistics as a way of understanding the human mind. I am indebted to my colleague James Zeitler, whose expertise sharpened my understanding of machine learning which was crucial for completing this thesis.

To my family who supported and encouraged me throughout the ALM process—no easy feat for them given that this degree pursuit has been a 16-year labor of love of mine—I love you so much. My parents Prasad and Padmavathi were always there for me. Nausheen, you are the sweetest one no doubt. Aditya and Padma, you are my treasure, and I know you will achieve great things in your lives—may all your dreams come true.

Table of Contents

List of Tables

List of Figures

Chapter I

Introduction

All human languages have *syntax,* which is "a system of rules and categories" that form the basis of how sentences are formed (O'Grady 730). The words that comprise these sentences can be divided into a handful of groups or "syntactic categories." Those words, which carry substantial semantic meaning, fall into the lexical categories such as nouns, verbs, adjectives, prepositions, and adverbs. Non-lexical or functional categories are determiners, degree words, qualifiers, conjunctions, and auxiliary verbs (O'Grady 184-185). The last mentioned of these functional classes, that is, auxiliary verbs (or "auxiliaries," abbreviated as "aux"), will play the pivotal role in the analysis of Old English poetic data and the development of the resulting arguments presented in later chapters.

The main verb in a sentence or clause, also known as the "verbal," carries the primary semantic weight of the action being described. Baker (2012) explains that in Old English, verbals can be any one of three forms of a verb: (a) the infinitive which is the basic form of the verb that one typically uses in dictionary lookups, (b) present participle which is an adjective-like verb usually signaling ongoing, repeated, or habitual action, or (c) past participle which "expresses the state that is consequent upon an action having been completed" (26). Auxiliaries help such verbals by adding functional meaning and are thus also referred to as "helping verbs." Examples of auxiliaries in Modern English

are *will, can, may, must, should,* and *could*. Some forms of Old English auxiliary verbs are *willan* ("wish," "be willing," "desire," "intend"), *magan* ("be able to," "can," "may"), *bēon* ("be"), and *habban* ("have," "hold," "possess"). This thesis strives to identify poetic styles in Old English literature, using a data set of auxiliaries and a methodology described in Chapter III. Owing to the lower semantic weight of auxiliary verbs, one can speculate that they are used somewhat subconsciously because of personal preference by poets and, as a side-effect, offer clues about their individual styles of composing verse. Parallels with conjunctions would be the classic case of "since" vs. "because" or possessive phrase constructions like "the book's cover" vs. "the cover of the book." In these examples there is no semantic difference between the two options, a writer might simply pick one due to being more partial to one usage. If such choices are made consistently they could play a part in the defining the "style" of a particular writer or school of writers.

Alan Bliss (1980) studied the word order of Old English clauses in *Beowulf* that contain an auxiliary verb, which he defines as "a finite verb used with a dependent infinitive or past participle." (159) This was part of his endeavor to establish "styles" of Old English verse syntax which is a topic he writes, on which scholars lack consensus. Contributing to this lack of consensus, Bliss says, are factors such as the nature of poetry, in which the artistic choice to preserve archaic word orders to allow "rhetorical" constructions is not uncommon. Additionally, and most importantly, poetic metrical concerns perhaps had a large role in what syntax could be used in a given work. Bliss argues that these conditions makes it very challenging to categorize Old English verse into distinct styles of poetry even when one examines so simple a component of syntax as

word order. He generalizes three patterns of word order in Old English prose, as the following phrases show:

SVO:	Se cyning besæt þæt fæsten.
VSO:	Þa besæt se cyning þæt fæsten.
SO...V:	Þa se cyning þæst fæsten besæt.

Where each pattern has the following components:

S: Subject, e.g., *Se cyning* ("the king")
V: Verb, e.g., *besæt* ("besieged")
O: Object, e.g., *fæsten* ("place")

In the examples above, word order assists in establishing the meaning of words such as *þa* which can mean either "then" or "when." The VSO phrase is a principal clause starting with "then," while the SO…V phrase is a subordinate clause beginning with "when." There are many instances, however, when only one or two of these three components appear. This can happen with intransitive verbs which do not take an object, resulting in the pattern SV which could be either SV[O] or V[S]O (where the square brackets denote a component that does not appear explicitly in the verse). Additionally, in poetry, the subject is often left out since it can be inferred by virtue of occurring in a previous line, in which case the pattern could be [S]VO or V[S]O. In both these situations determining the pattern is fraught with uncertainty. To avoid this ambiguity Bliss chooses to work with the word order of phrases containing auxiliary verbs. Using this approach, Bliss' paper finds several "constraints controlling the word order" in *Beowulf* in auxiliary verb clauses (178).

Donoghue (1987) extended the study of the Old English auxiliary well beyond *Beowulf* by examining 19 poems and documenting in detail all occurrences of auxiliary-

verbal phrases with the goal of "discovering new facts (or recovering old ones) about the techniques of Old English verse." (2) The poems included in his study are: *Andreas, Beowulf, Christ and Satan ("Chr Sat"), Christ I, Christ II, Christ III, Daniel, Elene, Exodus, Genesis A, Genesis B, Guthlac A, Guthlac B, Juliana, The Battle of Maldon ("Maldon"), Meters of Boethius ("Met Boe"), Metrical Psalms ("MPsalms"), The Phoenix,* and *Solomon and Saturn ("Sol Sat").* Donoghue also included a 20th poem, *The Fates of the Apostles ("Fates")* when he examined the question of Cynewulfian authorship of a subset of the 19 works.

Donoghue hypothesized that such a study would advance our understanding of Old English syntax and help determine various poetic styles in an objective way. Such a data-driven statistical analysis (using 16 different attributes of auxiliaries in each of the 19 poems) he writes, would result in groupings of poetic style (if any such indeed exist) which are more rigorous than looser categories such as oral vs. written, pagan vs. Christian, epic vs. lay, and classical vs. debased (3).

Donoghue provides the following word-order examples of auxiliary-verbal clauses in Old English and the environments in which they occur:

SvOV:	Se cyning hæfde þæt fæsten beseten.	(Common)
vSOV:	Þa hæfde se cyning þæt fæsten beseten.	(Demonstrative)
SO...Vv:	Þa se cyning þæt fæsten beseten hæfde.	(Conjunctive)

Where:

S: Subject, e.g., *Se cyning* ("the king")
O: Object, e.g., *fæsten* ("place")
v: Auxiliary, e.g., *hæfde* ("had")
V: Verbal, e.g., *beseten* ("besieged")

Donoghue describes how the additional advantages of using the auxiliary become apparent through these examples since the template S-O-V is consistent in each with the

variable being *v* (the auxiliary), which is slotted in at different locations in the template depending on the type of the clause (5). Even if the object is absent it is still possible to tell the three templates apart. Consider the patterns Sv[O]V, vS[O]V, and S[O]...Vv: in each case it is possible to deduce the slot of the object and thus recognize the pattern. (By contrast, a missing subject complicates matters since the distinction between [S]vOV and v[S]OV cannot be ascertained.) There are only two patterns of the auxiliary and verbal: vV and Vv, so the occurrence of the subject and object are not of much concern. The ubiquity and abundance of auxiliaries in Old English poetry provides a substantial amount of data on which empirical analyses can be conducted. The "metrical status" of auxiliaries is also quite predictable; that is, they are unstressed except in the very particular case of being "displaced from the initial metrical dip of the clause." (Donoghue 7) This allows us to consider stress as a factor in determining categories. Another benefit of using auxiliaries as a barometer of style is that their stress and template position are subject to four constraints (Kuhn's two laws, Sievers' law, and Bliss' rule) that can be used as additional factors of analysis (Donoghue 15).

This approach of the "test of the auxiliary" is not without its limitations. Donoghue identifies three of these (21). First, it does not necessarily follow that two poems using auxiliaries in the same fashion have common authorship. Second, since the approach is evidence-based, it is imperative that a critical mass of data be taken into account, which eliminates shorter poems from the data set being analyzed on the grounds that a shorter poem provides an insufficient sample size. Finally, the topic or genre of the poem may play a significant role in the auxiliaries that occur in it. A saint's life, for example, can be quite different in tone and lexicon from a poem in a profane genre, such

as *The Battle of Maldon*.

Donoghue finds that Old English poems could not be assigned cleanly to distinct categories and that "...the evidence from auxiliaries is too heterogeneous to identify schools or general styles of Old English verse…" (101). He concludes, however, that they do function as "syntactic markers" and that poets leveraged them in constructing half and full-lines (101).

It is certainly challenging to pin down a precise definition of the word "style." Ohmann offered this generic description: "Style is a *way* of doing *it."* (1964) Calder in 1979 compiled a set of essays by various authors that discuss style in Old English poetry but he writes that these writers each investigated just "...one aspect of Anglo-Saxon style..." and none of them tried to "...treat the field as a whole." (57) This thesis aims to build upon the foundation of the data compiled in Donoghue's work discussed above in an effort to possibly arrive at a more precise definition of the word "style" and thereby bring an element of certainty to this collection of poetry which has a "...long, often unfocussed, and sometimes confused tradition." (Calder 57) The hypothesis is that a deeper investigation of auxiliaries applying more advanced methods of statistical analysis to features of these words and the syntax of the verses they appear in, will reveal stronger distinctions among the same 19 poems. This includes identifying patterns of possible common authorship/borrowing, changes in style, and variation in individual styles of Old English verse, as one might expect to find in a literary tradition spanning a few centuries.

As this study will show, the outcome of this endeavor is a surprising one, especially to those accustomed to the romantic/modernist notion of the poet as an individual whose distinctiveness is grounded in his own highly personal literary style.

Chapter II

The Data: Auxiliaries and their Attributes

This chapter describes the 16 features (also referred to variously as "attributes," "variables," or "dimensions") of the poems used in Donoghue's study. The first variable is quite simply the average number of auxiliaries that occur per 100 lines of each poem expressed as a percentage. The second and third variables are the percentage of auxiliaries in *a-clauses* and *b-clauses*. This bears a brief explanation. A characteristic feature of Old English poetry is its alliteration, which is "a repetition of the same sound at the beginning of two or more stressed words in a line." (Terasawa 3) Each line itself comprises two verses or half-lines occurring on either side of a "syntactical boundary" called a *caesura* (Baker 123). In modern publications of Old English literature, this *caesura* is visually depicted with a space. The clauses that begin before and after this syntactic boundary are referred to as the *a-clause* and *b-clause* respectively. Consider for example the opening lines of what is undoubtedly the most popular Old English work, i.e., the epic *Beowulf* presented below in which a space visually divides each line into two halves which are sometimes referred to as the on-verse and off-verse.

Hwæt! We Gardena in geardagum,
þeodcyninga, þrym gefrunon,
hu ða æþelingas ellen fremedon.
Oft Scyld Scefing sceaþena þreatum,
monegum mægþum, meodosetla ofteah,
egsode eorlas. [*Beowulf* 1-6]

The fourth attribute Donoghue defines, *light auxiliaries* (9-10), are those that Bliss called "monosyllables" but in fact fall into the following categories:

(a) True monosyllables (e.g., *sceal*, *mōt*)
(b) Disyllables containing an unstressed prefix (e.g., *onginð*, *forlǣt*)
(c) Those which have a short first syllable (e.g., *hafað*, *mægen*, *sculon*)
(d) Those preceded by a negative proclitic (e.g., *ne sceal, ne onginð, ne hafað*)

The fifth attribute, Donoghue's *heavy auxiliaries* (10) correspond to Bliss' disyllables, which are:

(a) Disyllables with a long first syllable (e.g., *mōton*, *sceoldon*)
(b) Disyllables preceded by an unstressed prefix (e.g., *ongunnon*)
(c) Either of the two above preceded by a negative proclitic (e.g., *ne mōton*, ne *sceoldon*, *ne ongunnon*)

The sixth, seventh, and eighth variables are the percentages of the three general groups of auxiliaries used by Donoghue in his study: *modals* (*cunnan*, *magan*, *mōtan*, *sculan*, and *willan*), *passive periphrastics* (*bēon*, *weorþan*, *standan*, and *licgan*), and *others* (e.g., quasi-auxiliaries such as *habban* and *onginnan*, accusative-and-infinitive constructions such as *hēt...beran*, and verbs of motion followed by infinitives of motion such as *gewāt...nēosan*) which are neither of the other two groups (27). The ninth variable is the percentage of auxiliaries that occur in initial clauses of the poems. The 10th variable is the number of bracketing patterns for every 100 initial auxiliaries. According to Donoghue, bracketing patterns are stylistic devices and "can be considered to be elaborate formulas" denoted by the pattern "v...A||AV," where v and V represent the auxiliary and verbal, and A represents the alliterating word in the line (43). An example

of such a bracketing pattern is shown below:

> Wæs se grimma gæst Grendel hāten,
> mǣre mearcstapa. [*Beowulf* 102-103a]

The rest of the variables deal with the order of auxiliaries (v) and verbals (V) and their stress patterns in principal and dependent clauses. A principal clause begins with a noun, pronoun, adjective, or adverb, while a dependent clause begins with a relative pronoun or a conjunction (Donoghue 48). Three stress patterns are possible for each of these types of clauses: vV, v́V, and Vv́ yielding six total variables 11 through 16. (When an acute accent mark appears above a symbol, it indicates that the word represented by that symbol is stressed.) The 16 attributes of auxiliaries used for analysis are described in the following list and summarized in tables 1 and 2. (The shorthand text presented in parenthesis at the end of each list item corresponds to the column heading in tables 1 and 2.)

1. Percentage of auxiliaries in the poem (aux).
2. Percentage of auxiliaries that occur in a-clauses (a-clause).
3. Percentage of auxiliaries that occur in b-clauses (b-clause).
4. Percentage of auxiliaries that occur with light syllables (light).
5. Percentage of auxiliaries that occur with heavy syllables (heavy).
6. Percentage of auxiliaries that are modals (modal).
7. Percentage of auxiliaries that are passive periphrastics (passive).
8. Percentage of auxiliaries that are "other," i.e., neither modals nor passives (other).
9. Percentage of auxiliaries that are clause-initial (initaux).
10. Percentage of bracketing patterns per 100 initial auxiliaries (bracket).

11. Percentage of auxiliaries in principal clauses with a stress pattern vV (prin_vV).
12. Percentage of auxiliaries in principal clauses with a stress pattern v́V (prin_v́V).
13. Percentage of auxiliaries in principal clauses with a stress pattern Vv́ (prin_Vv́).
14. Percentage of auxiliaries in dependent clauses with a stress pattern vV (dep_vV).
15. Percentage of auxiliaries in dependent clauses with a stress pattern v́V (dep_v́V).
16. Percentage of auxiliaries in dependent clauses with a stress pattern Vv́ (dep_Vv́).

Note that Donoghue identified two other variables in his study, referring to them as “Doubtful Principal” and “Doubtful Dependent” clauses, which he analyzed separately. These “doubtful” clauses are neither confidently principal nor dependent, and since some doubt attends their status it is better not to include them among the other clauses. For this reason, these two variables were omitted and not used in this thesis.

Table 1
Auxiliaries and Attributes 1-8

Poem	aux	a-clause	b-clause	light	heavy	modals	passives	others
Andreas	19.5	57	43	60	40	39	26	35
Beowulf	18.6	54	46	51	49	44	19	37
Chr Sat	22.5	63	37	55	45	57	17	16
Christ I	18.9	54	46	58	42	43	41	18
Christ II	18.5	52	48	72	28	53	29	17
Christ III	18.2	49	51	58	42	56	27	26
Daniel	17.3	70	30	52	48	33	30	36
Elene	18.9	40	60	65	35	27	43	30
Exodus	14.8	71	29	45	55	33	32	34
Genesis A	19.3	42	58	62	38	38	20	42
Genesis B	35.5	38	62	51	49	54	14	32
Guthlac A	21	52	48	47	53	60	22	18
Guthlac B	17.1	26	74	63	37	47	30	23
Juliana	19.2	39	61	69	31	39	21	40
Maldon	28.3	78	22	48	52	51	16	33
Met Boe	19.6	59	41	62	38	58	21	21
MPsalms	12.2	81	19	58	42	41	46	13
Phoenix	13.9	39	61	65	35	30	57	13
Sol Sat	16.8	64	36	75	25	61	18	21

Source: Donoghue, Daniel. *Style in Old English Poetry: The Test of the Auxiliary*. New Haven: Yale University Press, 1987. Print.

Table 2

Auxiliaries and Attributes 9-16

Poem	initaux	brackets	prin_vV	prin_ýV	prin_Vý	dep_vV	dep_ýV	dep_Vý
Andreas	27	25.3	50	37	13	4	40	57
Beowulf	24	24.1	38	38	23	5	41	53
Chr Sat	25	17.1	54	23	23	24	22	55
Christ I	8	28.6	47	40	13	11	39	50
Christ II	25	45	67	30	3	8	39	54
Christ III	25	19.4	59	26	15	17	39	45
Daniel	26	23.5	41	28	31	4	32	64
Elene	21	13.2	41	41	18	4	27	70
Exodus	37	53.1	63	31	6	0	39	61
Genesis A	27	19.1	39	45	17	12	46	42
Genesis B	19	7.3	66	18	16	27	31	43
Guthlac A	24	17.1	42	31	27	5	36	59
Guthlac B	22	4.8	46	43	11	3	45	53
Juliana	19	11.5	39	42	19	2	47	51
Maldon	24	36.4	70	24	6	22	24	54
Met Boe	16	26.4	53	36	11	6	56	38
MPsalms	29	50.6	39	40	21	17	54	29
Phoenix	19	16.7	40	45	16	3	47	50
Sol Sat	19	25	56	32	12	43	43	14

Source: Donoghue, Daniel. *Style in Old English Poetry: The Test of the Auxiliary*. New Haven: Yale University Press, 1987. Print.

Chapter III

Methodology

This study uses an *Unsupervised Machine Learning* approach to analyze the data discussed in the preceding section. In his 1987 study, Donoghue approached this problem by examining and trying to make sense of raw counts and percentages, and he did not use advanced statistical methods. Therefore, his data could potentially benefit from a more rigorous analysis. In this thesis, the data are subjected to three types of machine learning cluster analysis in order to identify groupings among the Old English poems under consideration. Machine learning implies that software algorithms can be applied to allow computers to learn from datasets. The adjective “unsupervised” indicates that this approach can be used to extract hidden patterns and structures in previously uncategorized or unlabeled input data (Albalate 3) and can be contrasted with “supervised” learning methods where training sets or examples are used to teach a computer program how to perform certain functions.

Using the attributes of the poems that have been captured in tables 1 and 2, a multi-dimensional analysis is conducted to see if, from a statistical standpoint, some poems are similar to or different from others. Specifically, the data are subjected to three types of cluster analyses in order to identify groupings (or "clusters") among the Old English poems under consideration. The analysis is conducted using the statistical and graphing "R" programming environment, which is open-source software used extensively by data scientists and statisticians for data analysis. (The program source code used in

this study is presented in Appendix A and B.) Once the clusters have been algorithmically determined, the goal is to ascertain if these groups of poems can be characterized by style, genre, date, or any other plausible facet.

Note that although 16 attributes appear in tables 1 and 2, some of them are complements of each other. For any given poem, *Solomon and Saturn* for example, the values of the a-clause (64%) and b-clause (36%) add up to 100%, as do the values for light (75%) and heavy (25%) syllables. Each of these two pairs therefore constitutes a complementary set. The ramification is that it is possible to deduce the value of one member of any such pair if the value of the other member is specified. Therefore, the information carried by one of them is always redundant. For this reason, the value of only one member of each pair—it does not matter which one—needs to be used as input for analyzing this data. In this study, the values of a-clause and light syllables will be used as input to the software, and consequently b-clause and heavy syllable values will be ignored.

In general, if a set comprises *n* number of variables, and if it is possible to determine all the information expressed by those *n* variables from only *n -1* of them, then for all practical purposes only *n -1* variables need to be used for analysis. This means that the data can be further simplified when working with other sets of attributes in tables 1 and 2 where the values add up to 100%. Taking *Solomon and Saturn* as the example once again, we can see that the set of modals (61%), passives (18%), and others (21%) add up to 100%. Similarly, the sets of prin_vV (56%), prin_v́V (32%), and prin_Vv́ (12%) total 100% as do dep_vV (43%), dep_v́V (43%), and dep_Vv́ (14%). In each of these sets, it is possible to calculate the value of any one member if the values of the other two are

known, so one only needs the values of any two to capture the complete information about that set. It is again possible to drop one of the members of each set for computational purposes since the information carried by any one of them is superfluous. Therefore, this study will drop the attributes others, prin_Vv́, and dep_Vv́. The outcome of this exercise is that 11 variables are used as software input although effectively the information of all 16 variables in tables 1 and 2 is being utilized.

The first method employed to analyze the data is called *Hierarchical Clustering* using the distance matrix method. This involves taking the initial data set with each item, also referred to as an "observation," in its own cluster. These clusters are iteratively combined until there is only one remaining. Since this is a combinatorial method, it can be more specifically described as an *agglomerative* procedure (Kabacoff 370). *Hierarchical clustering* provides a good jumping off point for this type of analysis since one of its outputs is a *dendrogram*, which is a tree-like branching diagram where each observation being analyzed (in this case each poem) appears individually at the bottom. At each higher level, some poems get combined into clusters based on how similar they are to each other. This dendrogram provides a hint of a possible family tree of the observations in the data set.

Figure 1 shows an example dendrogram of how 18 countries can be grouped based on four variables: per-capita income, literacy rate, infant mortality, and life expectancy which was data tabulated by Wikibooks (2016). At any given value of "Height" on the Y-axis, which represents the dissimilarity of the observations (in this case, the countries), a horizontal "cut" yields clusters of countries as branches of the tree at that level of dissimilarity. Three clusters are illustrated in figure 1, with the cut being

made at a “Height” of approximately 18,000, and show how the countries are grouped together based on the four variables. (A cut at 30,000 would have identified only two branches.) As one might expect, the developed nations and developing nations appear in different branches. Even within the “developed” cluster, there is a sub-branch comprising Italy and Lithuania, indicating that, based on the variables used, these two countries are more similar to each other than they are to Australia, the United Kingdom, Japan, Germany, and Greece.

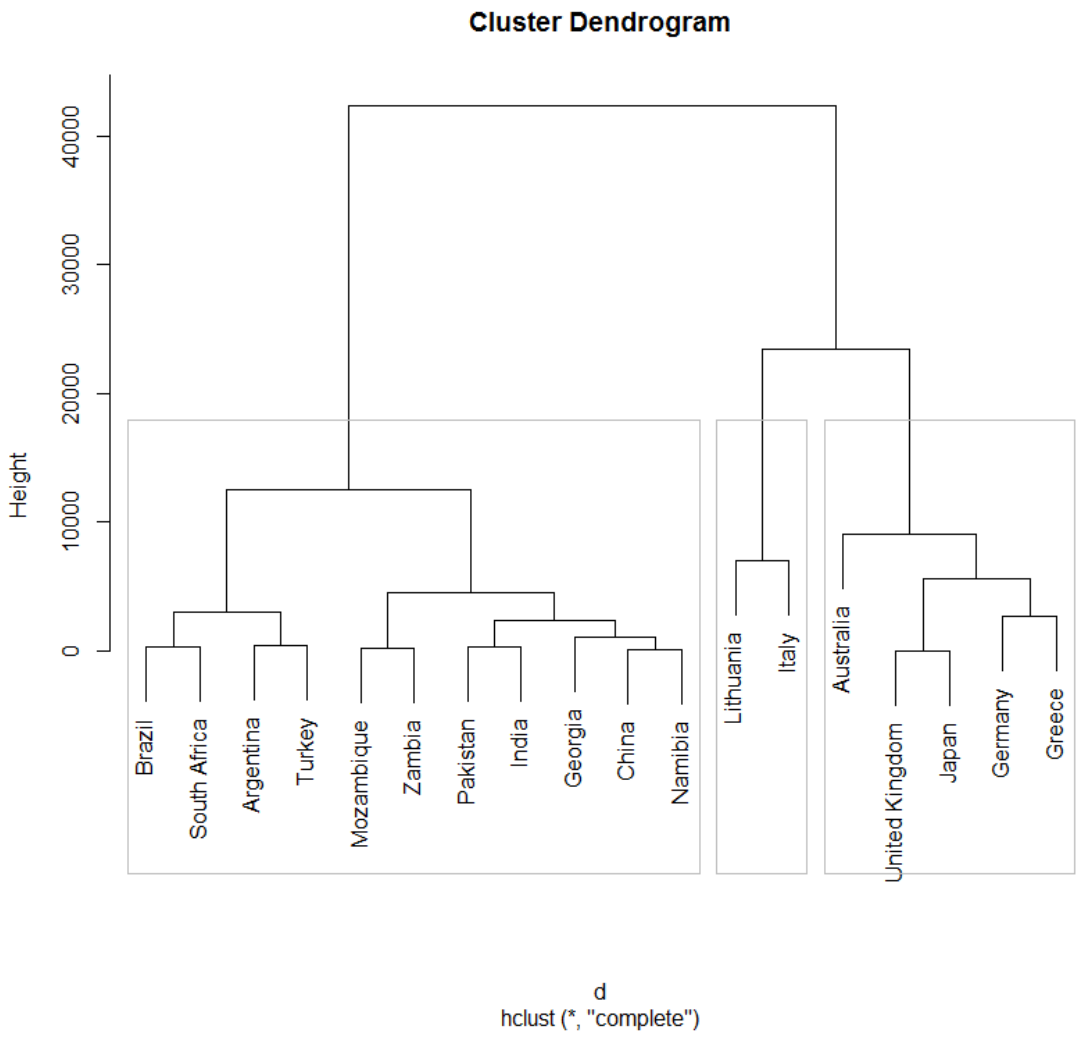

Figure 1. Example dendrogram generated from hierarchical clustering

The second method is *k-means Clustering* in which a random value of *k* is initially used to select *k* "means" or "centroids." The mean, or centroid, is the center of each cluster but in multiple dimensions, or more precisely, as many dimensions as there are in the data set. Each observation (which in this context is a poem) is then assigned to its nearest centroid followed by generating a new set of centroids as the average of all observations in a given cluster. These steps are run iteratively until the clusters into which the observations are assigned are relatively stable or for a predetermined number of iterations (Kabacoff 378). Note that in this method it is possible for an observation to change clusters, unlike *agglomerative hierarchical clustering* in which "once an object is allocated to a group, it cannot be reallocated" as the number of clusters decreases (Mardia 369).

A *k-means* clustering is typically run several times with increasing values of *k* (starting from k=2), but the algorithm does not by itself provide any clues as to how many clusters might actually exist in a given data set. It is possible to make a judgment of the number of clusters by leveraging the Sum of Square Errors (SSE) value computed by this approach, which tends to decrease to zero as the value of *k* increases. SSE is therefore a convergence criterion that can be used to determine how many iterations of *k-means* should be run. When SSE is graphed against the corresponding value of *k*, an "elbow" pattern is sometimes revealed. This is the value of *k* for which the SSE has been reduced to a low value and even if *k* is increased the SSE does not decrease significantly. In other words, the value represented by the elbow is probably the right number of clusters (Poulson 2013).

Another rule-of-thumb with this type of analysis is to aim for high entropy in

cluster sizes; that is, there should not be a large variation in the number of observations in each group. Figure 2 illustrates a *k-means* analysis for the sample data of countries for a k value of 3. The three ellipses indicate which countries group together. It should be noted that this diagram illustrates multi-dimensional data on a 2-dimensional graph; hence on paper it appears that ellipse 1 and 3 are intersecting when in fact the data behind them are not. Figure 3 illustrates how an "elbow" can be discerned at a k value of 3.

The *k-means clustering* approach has some limitations due to its underlying assumptions, one of which is the presupposition that the clusters are spherical with all the data points in a given cluster being equidistant from its centroid. Therefore this method works well for some data sets and not others (Wikipedia 2017).

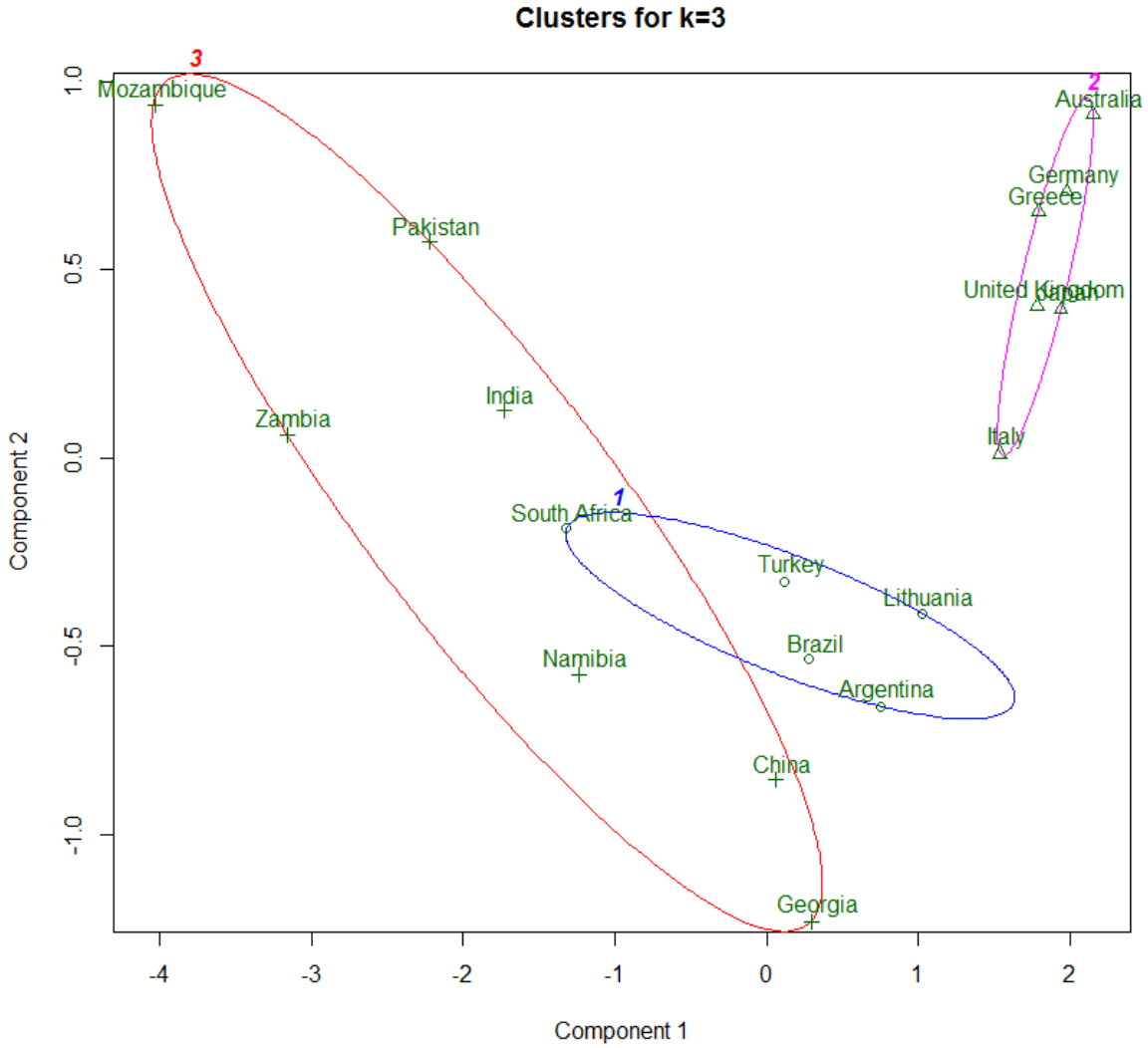

Figure 2. Example clusters generated by k-means using a k value of 3

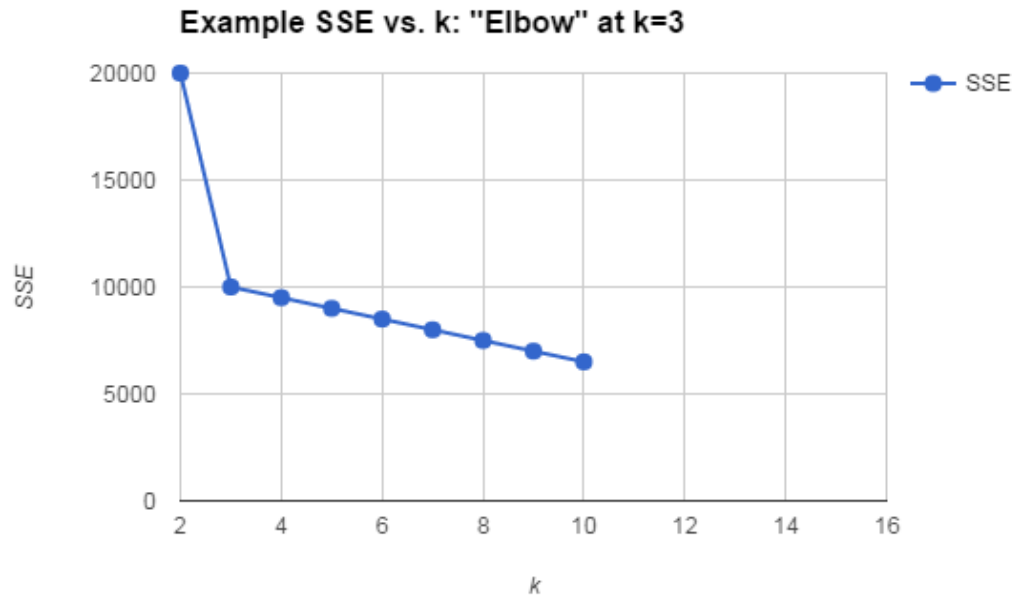

Figure 3. k-means clustering “elbow”

The third method used in this thesis is called *Principal Component Analysis* (*PCA*). In this approach a multi-dimensional data set can be reduced to fewer dimensions by combining the variables/attributes into "components" to determine which components have the strongest influence on how the data can be grouped into clusters (Hothorn 347-348). *PCA* can thus help in making the data easier to explore. For example, by reducing the data into two components it can be plotted on a 2-dimensional graph which is easier for humans to process. *PCA* is quite different from the two clustering methods described earlier in that it attempts to group the columns of the data in tables 1 and 2 and deals with these features of the poems in aggregate, while *hierarchical* and *k-means* clustering attempt to group the rows (Poulson 2013). Therefore, this study uses the two clustering methods to identify potential groups among the poems and then uses *PCA* to try and understand which attributes of the poems have the greatest impact in how the poems are

distributed across the groups. Figure 4 illustrates this on a "biplot," with PC1 and PC2 being the two components with the highest importance in that order. The four variables are shown with a directional arrow or "vector" indicating how they correlate to PC1 and PC2. For example, infant mortality moves in a negative direction with respect to PC1. This indicates that the lower a country's infant mortality the higher it scored on PC1. Life expectancy exhibits the opposite behavior. Note that this diagram makes it appear that PC1 and PC2 are equally important since the X and Y-axes have the same scale. However, as has been stated earlier, PC1 has more influence on how the data is distributed. For this reason, India and Pakistan are not as far apart as the diagram may suggest at first glance. The difference between them, especially along the X-axis (PC1) is quite small. Japan and the United Kingdom are plotted practically on top of each other indicating that they are much more similar along both the PC1 and PC2 axes.

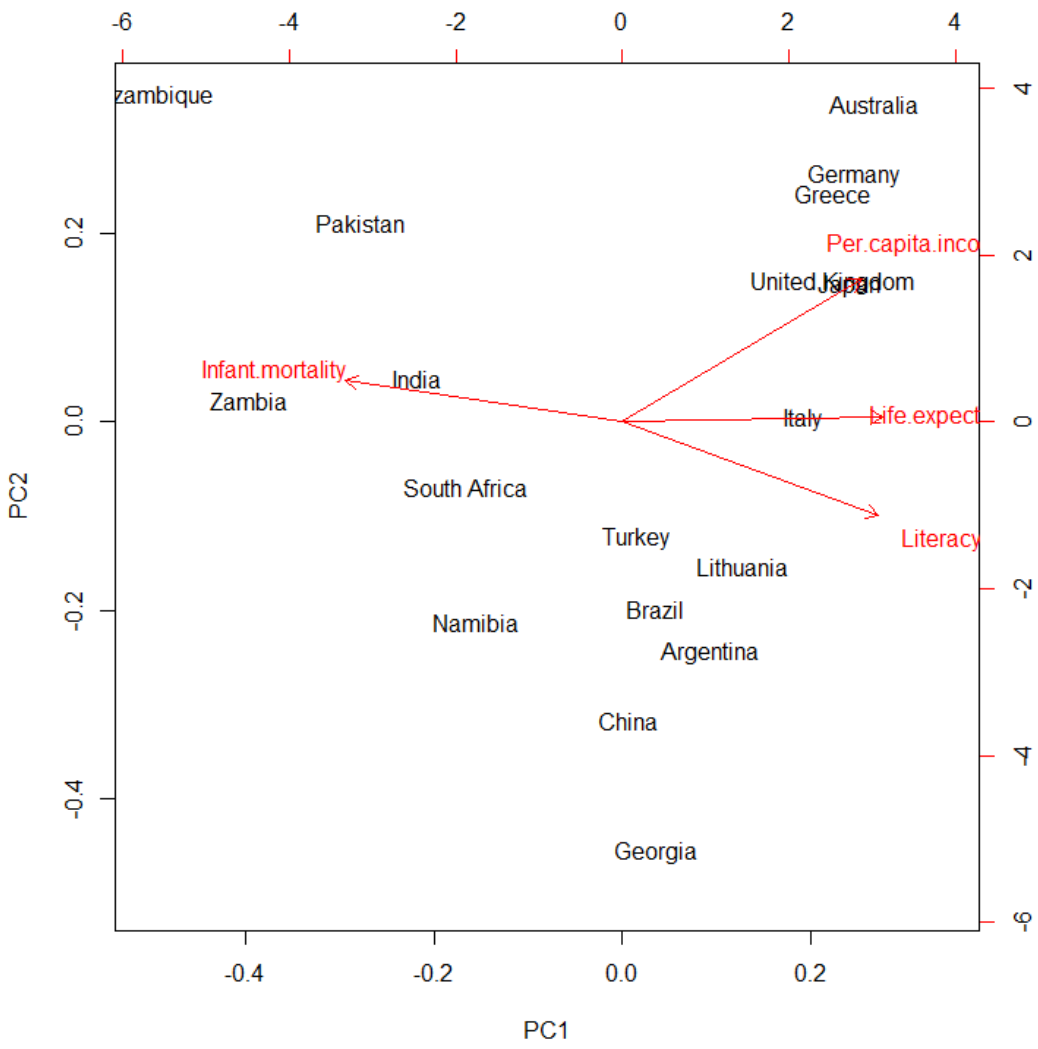

Figure 4. Principal Component Analysis of 18 countries

It is worth noting explicitly at this juncture that any clusters or groups identified by the above techniques do not necessarily mean that the observations under consideration in a given cluster are generally similar or related to each other, whether they are the 18 countries in the example data set above, or the 19 Old English poems that will be analyzed in later chapters. Rather, they are similar to each other only based on the values of the attributes used in the study. That is, if additional or a different set of attributes are used, it is conceivable that the poems could be grouped in entirely different clusters. As a trivial example, if a different set of variables—say average yearly temperature and distance from the equator in degrees latitude—were used to cluster the 18 countries, it is very likely that Australia and Brazil would have been grouped together as would Japan and China.

Chapter IV

Results

The results of the clustering analysis using each of the three techniques are described in this chapter.

Hierarchical Clustering

This algorithm was run multiple times in an attempt to group the 19 poems in 2 to 10 clusters. (Although 10 would be a large number of groups for only 19 poems it is nevertheless useful to use multiple iterations in order to judge upper limit beyond which further grouping becomes unproductive.) Table 3 displays the group number assigned to the poem by the computation for 2-10 clusters. For example, *Andreas, Beowulf, Christ I,* and *Daniel* are always in group 1 regardless of the total number of clusters. *Genesis A* on the other hand switches clusters: it is in group 1 for a 2, 3, and 4-cluster computation, in group 3 when the poems are categorized into 5 or 6 clusters, group 4 for 7 clusters, group 5 for 8 and 9 clusters, and group 7 for 10 clusters. *Solomon and Saturn* is remarkable in that while it is group 2 for the smaller number of clusters (2-5), it is in a different group for each of the larger clusters (7-10).

Table 3
Hierarchical Clusters

	Number of Clusters								
	2	3	4	5	6	7	8	9	10
Andreas	1	1	1	1	1	1	1	1	1
Beowulf	1	1	1	1	1	1	1	1	1
Chr Sat	2	2	2	2	2	2	2	2	2
Christ I	1	1	1	1	1	1	1	1	1
Christ II	2	2	2	2	2	3	3	3	3
Christ III	1	1	1	1	1	1	4	4	4
Daniel	1	1	1	1	1	1	1	1	1
Elene	1	1	1	3	3	4	5	5	5
Exodus	1	3	3	4	4	5	6	6	6
Genesis A	1	1	1	3	3	4	5	5	7
Genesis B	2	2	4	5	5	6	7	7	8
Guthlac A	1	1	1	1	1	1	4	4	4
Guthlac B	1	1	1	3	3	4	5	5	7
Juliana	1	1	1	3	3	4	5	5	7
Maldon	2	2	2	2	2	2	2	2	2
Met Boe	1	1	1	1	1	1	4	4	4
MPsalms	1	3	3	4	4	5	6	8	9
Phoenix	1	1	1	3	3	4	5	5	5
Sol Sat	2	2	2	2	6	7	8	9	10

The hierarchical clusters can be visually represented as well, as shown by figure 5 which displays a tree-structure of the clusters and how the poems combine into groups from the bottom to top. Based on the attributes used for this analysis, this illustrates how *Andreas* and *Beowulf* are similar to each other since they appear in the same branch and at the same level in the tree. That is, they combine quite early in the iterative process. *Genesis B*, in contrast, does not combine into a group until the last-but-one level indicating that it is quite dissimilar to the other poems in that branch.

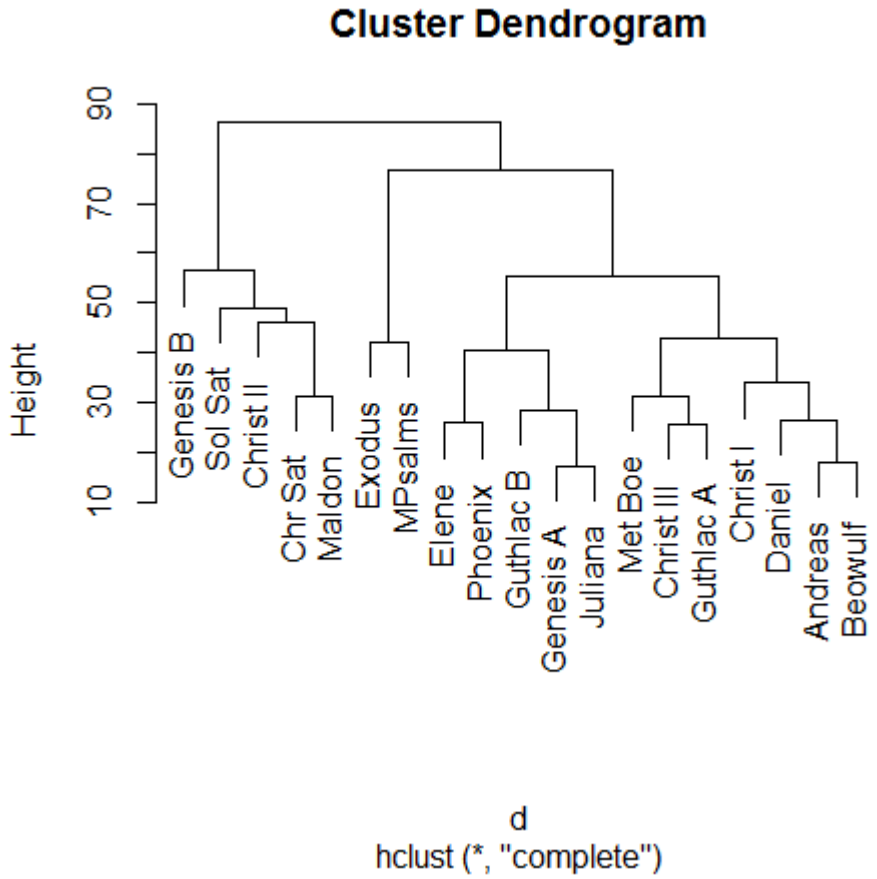

Figure 5. Hierarchical cluster dendrogram

The next step is to determine a reasonable number of clusters. Based on the dendrogram generated above, at first glance it seems four or five clusters would be a good number to use. The rationale for this choice is that for a fewer number of clusters (that is, two and three) the number of poems in each would be quite large and for a higher number of clusters (six or more) there would be too few poems in a given cluster. (Consider the extreme case of 10 clusters: this would result in no less than five poems each in a group of one by themselves—hardly a cluster!) Four or five clusters seems to be in the Goldilocks zone. Figures 6 and 7 visually show the dendrograms and groups for these two cases.

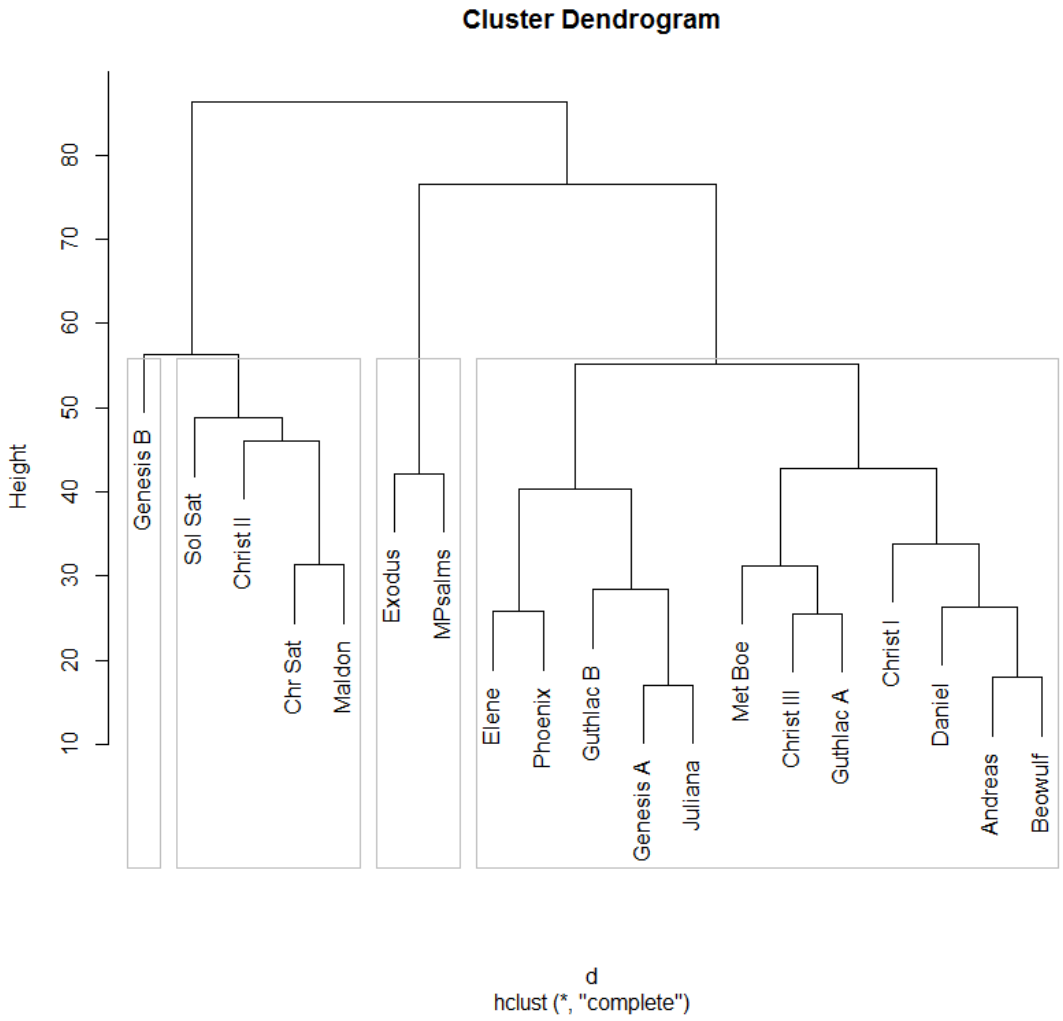

Figure 6. Hierarchical cluster dendrogram: Four clusters

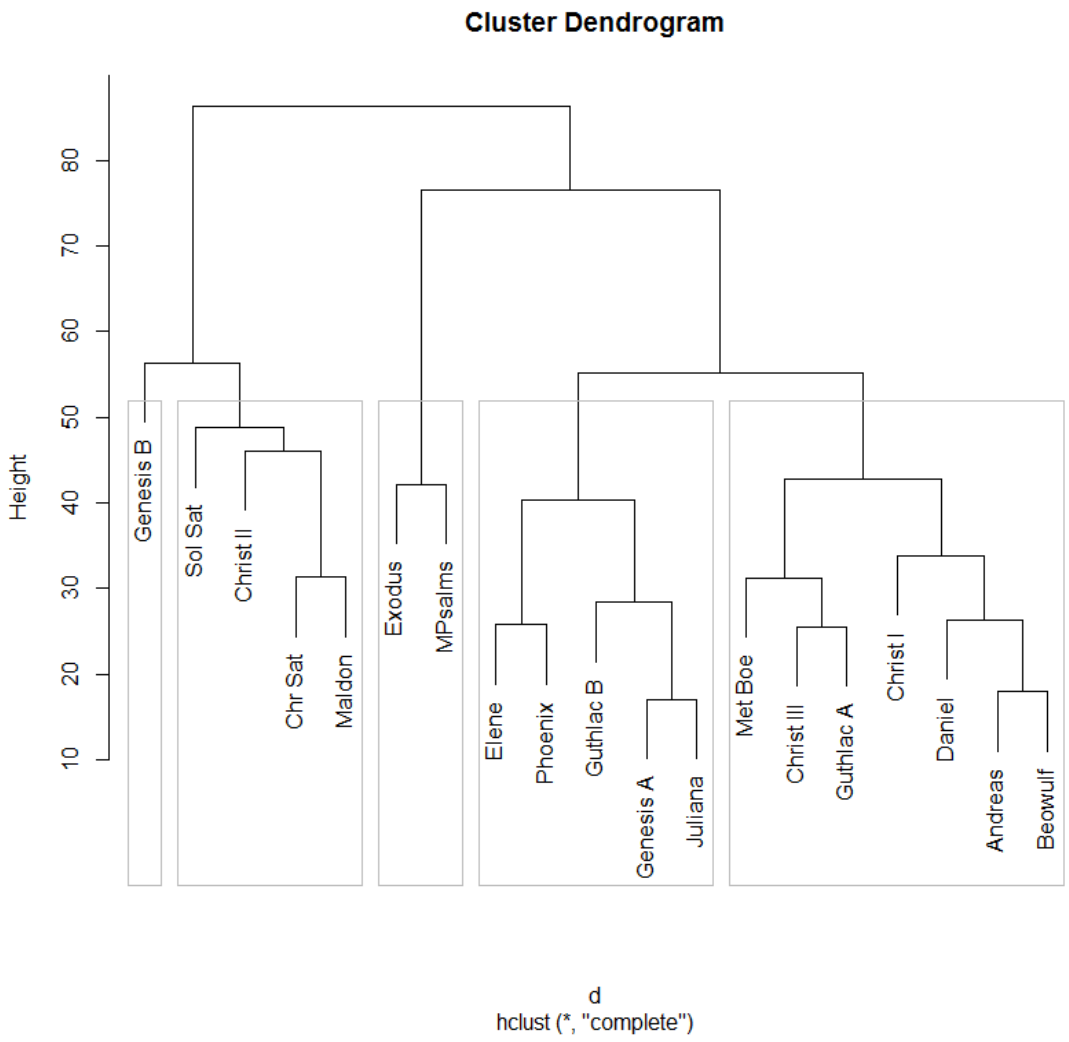

Figure 7. Hierarchical cluster dendrogram: Five clusters

k-means Clustering

This computation was run with values of *k* from 2-15 with the intention of identifying where the aforementioned "elbow" occurs. The data generated is captured in table 4 and a graphical representation is shown in figure 8.

Table 4
Variation of Sum of Squared Error (SSE) as the Value of k Increases

k	SSE
2	15708
3	11936
4	9636
5	8276
6	7099
7	5966
8	4880
9	3992
10	3244
11	2679
12	2147
13	1679

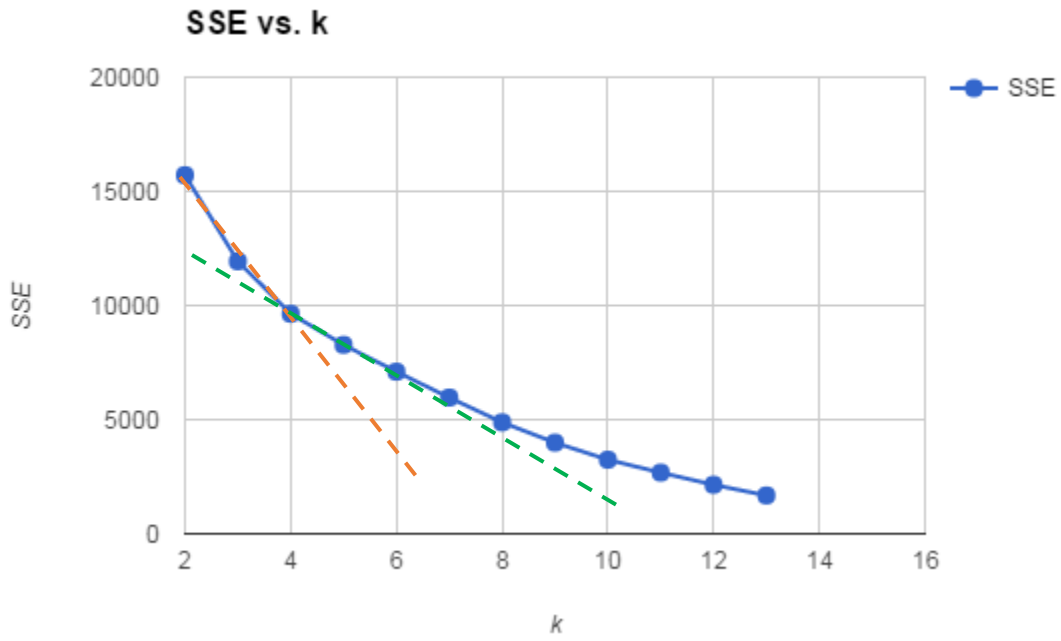

Figure 8. Graph of Sum of Squared Error (SSE) vs. k

Although choosing a reasonable value of *k* is a judgment call, as can be seen from figure 8, the "elbow" seems to occur in the graph at a *k* value of four, and an examination of the data in table 4 indicates that the relative decrease in SSE is not substantial for values of *k* in excess of four: it took only two steps of *k* for SSE to decrease from 15,708 to 9,636 (a difference of approximately 6,000), while it took another five increases of *k* for SSE to decrease from 9,636 to 3,992 (also a difference of approximately 6,000). Hence the curve flattened out significantly for a *k* value beyond four. (Note, however, that the goal of high entropy has not been achieved since the poems are not evenly distributed among the groups for any number of clusters.) Table 5 indicates how many observations occur in each group for the various cluster sizes.

Table 5
Cluster Size Variation

No. of Clusters	Cluster sizes
2	5, 14
3	5, 5, 9
4	2, 5, 5, 7
5	2, 3, 4, 5, 5
6	2, 2, 3, 3, 4, 5
7	1, 2, 2, 2, 3, 4, 5

Since four would be a good choice to use for the value of *k,* it is worthwhile examining how the poems are grouped using this computation and, as a secondary check, comparing it with a *k* value of five. This allows us to see if the poems get substantially regrouped beyond the "elbow." Tables 5 and 6 show how the poems are grouped for values of *k* of four and five, and figures 9 and 10 display these clusters graphically and indicate that the bulk of the poems remain in the same groups in both these cases. For this reason, this study will use a *k* value of four for the purpose of interpreting the results.

Table 6
k-means: Four Clusters of Sizes 2, 5, 5, 7

Cluster	Poems						
1	Andreas	Beowulf	Christ I	Christ II	Daniel	Guthlac A	Met Boe
2	Exodus	MPsalms					
3	Elene	Genesis A	Guthlac B	Juliana	Phoenix		
4	Chr Sat	Christ III	Genesis B	Maldon	Sol Sat		

Table 7
k-means: Five Clusters of Sizes 2, 3, 4, 5, 5

Cluster	Poems				
1	Andreas	Beowulf	Christ I	Daniel	Guthlac A
2	Exodus	MPsalms			
3	Elene	Genesis A	Guthlac B	Juliana	Phoenix
4	Chr Sat	Genesis B	Maldon		
5	Christ II	Christ III		Sol Sat	

Tabulating the data in this fashion makes it appear that all poems in a given cluster are equally similar to one another, when in reality they are not. The cluster plots of figures 9 and 10 help illustrate this since they display the poems on a coordinate system, making it possible to infer to some extent that some poems are further from the *k-means* centroid than others. However, as mentioned in Chapter III, these are plots of multi-dimensional data on a 2-dimensional graph and so are limited in indicating degrees

in dissimilarity. The latter are more apparent in *hierarchical clustering* diagrams such as figure 5 which shows how *Genesis B* is dissimilar to other poems in the same cluster.

Note that the cluster number generated by *k-means* (first column in tables 5 and 6) and *hierarchical* (first column in table 3) methods is essentially a label indicating groups, and there is no correspondence in labels across methods. For example, the *k-means* cluster label "4" in table 7 is not the same as the *hierarchical* cluster label "4" in table 3. In order to use a consistent group-labeling convention across the two methods, the clusters will be relabeled using letters from the Roman alphabet as shown in table 8. (Since we are comparing the five *hierarchical* with four *k-means* clusters, the entry matching letter "E" for *k-means* is not applicable.)

Table 8
Relabeled Clusters Across k-means and Hierarchical Methods

	Cluster Number	
Cluster Label	Hierarchical	k-means
A	1	1
B	4	2
C	3	3
D	2	4
E	5	N/A

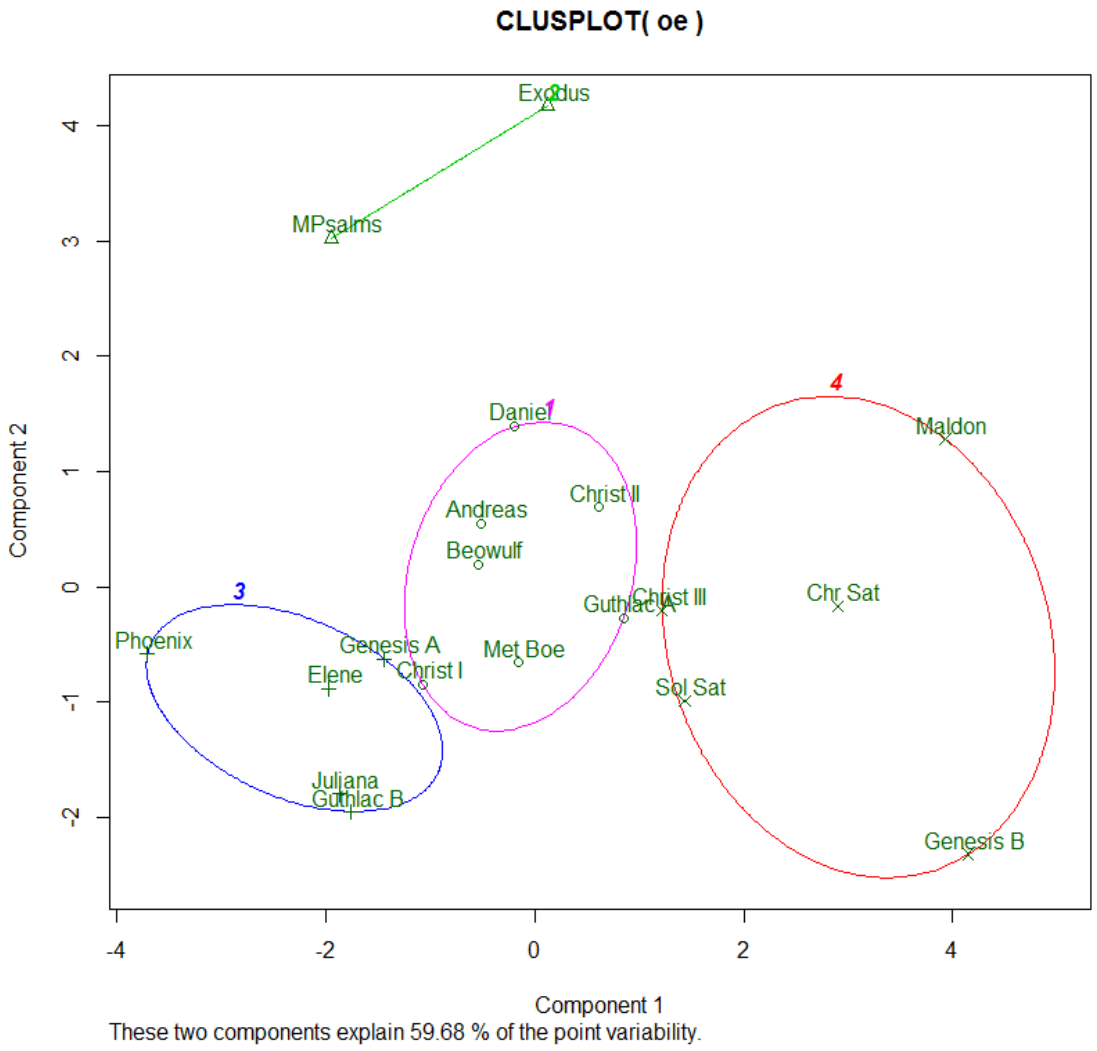

Figure 9. k-means clusters using a k value of 4

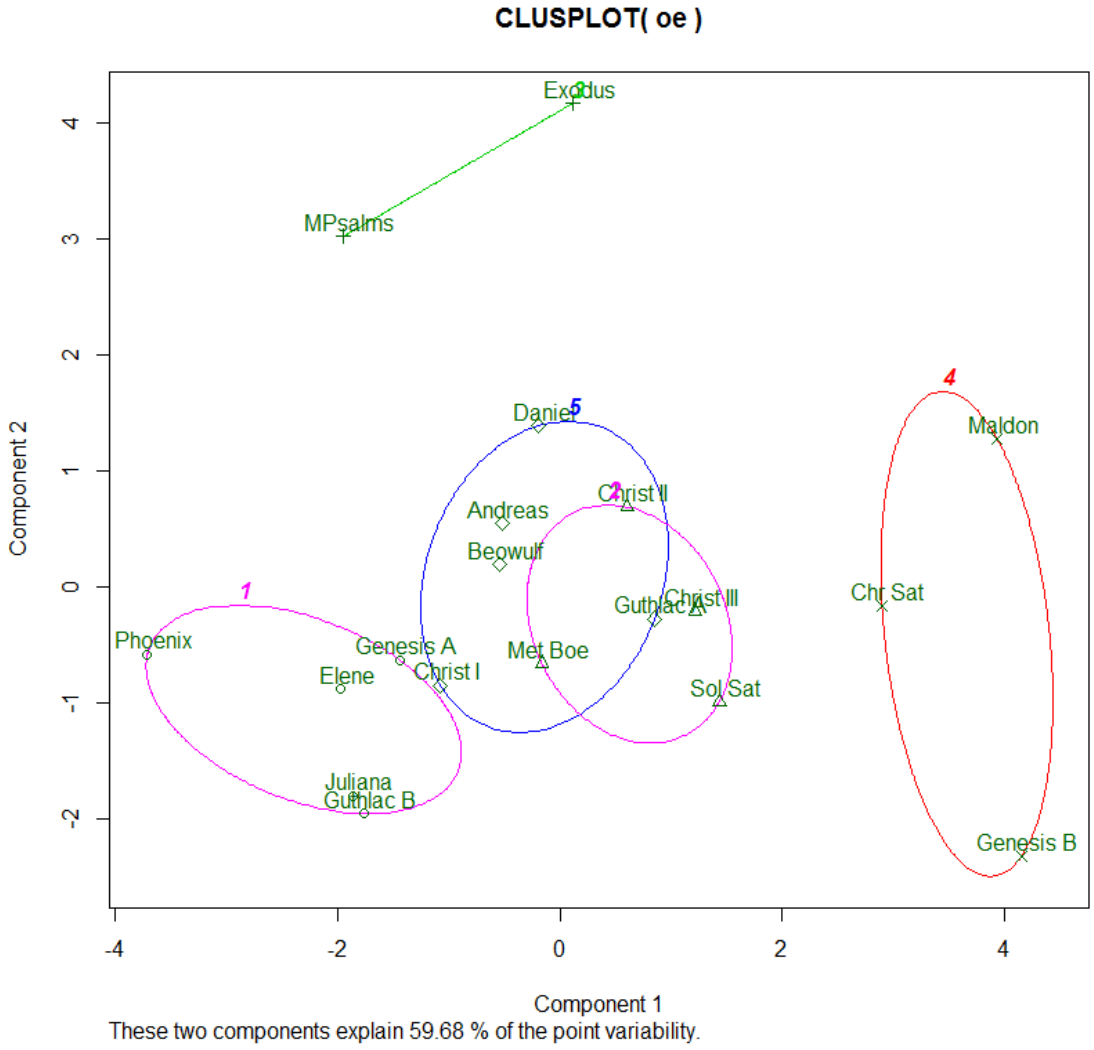

Figure 10. k-means clusters using a k value of 5

Principal Component Analysis (PCA)

A PCA computation on the data generated 11 components (PC1-PC11), summary statistics of each are displayed in table 9, with the standard deviation (SD) being the indicator of the importance of that component. That is, a higher value of standard deviation indicates higher importance. As for proportion of overall variation, PC1 has 0.37 (37%), PC2 has 0.23 (23%), and PC3 has 0.14 (14%). PC4 to PC10 have 7% or less. Selecting a cutoff is a judgment call (Poulson 2013). In this case it appears that the first two or three components have the largest influence—they cumulatively account for 0.74 (= 0.37 + 0.23 + 0.14) or 74% of the variation in the data. The values decrease rapidly from PC4 to PC11, which are therefore less important components.

Table 9
Summary PCA Statistics: Importance of the Components

Component	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Standard deviation	2.02	1.57	1.25	0.91	0.80	0.75	0.57	0.52	0.35	0.28	0.20
Proportion of Variance	0.37	0.23	0.14	0.07	0.06	0.05	0.03	0.02	0.01	0.01	0.00
Cumulative Proportion	0.37	0.60	0.74	0.81	0.87	0.92	0.95	0.98	0.99	1.00	1.00

The importance of each component can be represented by the scree plot shown in figure 11. A visual examination of this diagram reveals that PC1 is the most influential component while PC2 and PC3 have a much lower effect. PC4 to PC11 are less consequential for this analysis.

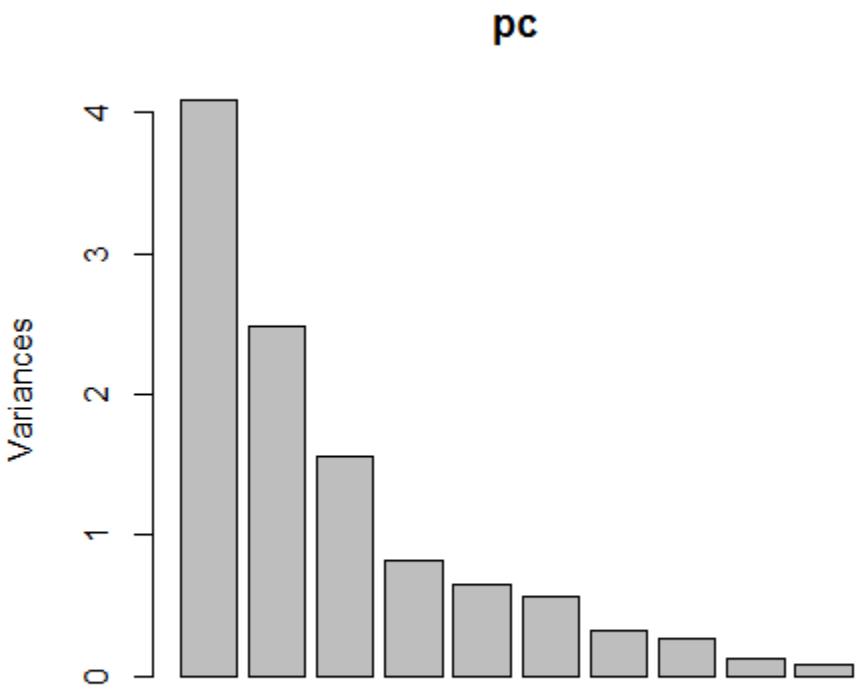

Figure 11. Scree plot of principal components

Consider now what constitutes each of the principal components, and the weights of each attribute for a given component, the data for which is shown in table 10.

Table 10
Relative Weights of Attributes in PCA

Rotation	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
aux	-0.38	0.25	-0.25	0.05	0.00	0.23	-0.47	0.29	0.04	0.56	-0.23
a.clause	-0.16	-0.50	0.21	0.09	-0.46	-0.08	-0.21	-0.23	0.28	-0.05	-0.52
light	0.20	0.28	0.49	0.21	0.44	-0.31	-0.19	-0.33	0.27	0.29	-0.10
modals	-0.33	0.17	0.38	-0.33	-0.11	0.16	0.54	-0.20	-0.29	0.32	-0.20
passives	0.36	-0.15	-0.02	0.59	-0.02	0.17	0.46	0.34	0.02	0.29	-0.23
initaux	-0.04	-0.46	-0.18	-0.39	0.44	-0.48	0.13	0.26	-0.03	0.23	-0.21
brackets	-0.02	-0.56	0.28	0.10	0.08	0.27	-0.24	-0.10	-0.30	0.33	0.51
prin_vV	-0.38	-0.10	0.15	0.22	0.57	0.36	-0.02	0.07	-0.10	-0.47	-0.28
prin_v́V	0.46	0.08	0.07	-0.10	-0.05	-0.01	-0.33	0.01	-0.70	-0.08	-0.41
dep_vV	-0.31	0.11	0.42	0.24	-0.23	-0.49	-0.05	0.53	-0.21	-0.14	0.16
dep_v́V	0.30	0.00	0.43	-0.46	-0.01	0.35	-0.09	0.49	0.36	-0.07	-0.01

Note that the absolute value associated with each attribute is what needs to be considered, not whether it has a positive or negative sign. For example, in the case of PC1, the weight of the aux is -0.38 and that of passives is 0.36. This indicates that both have similar weight in constituting this component even though one is a negative value. A negative value represents a negative correlation; thus a poem with a higher number of auxiliaries in b-clauses is lower on PC1. Conversely, a poem with a higher number of auxiliaries in a-clauses is higher in PC1. Another point of note is that all the attributes have a part to play in terms of their contribution to the principal components. This can be

discerned by examining the data under the column headings PC1, PC2, and PC3, which were determined to be the most influential. Although the absolute value (ignoring the sign and shown inside pipe symbols '|' below) of an attribute may be quite low in one of the columns, it is quite significant in one of the other two columns. Take for instance bracketing, which has a low absolute value of |0.02| under PC1 but much higher values of |0.56| and |0.28| under PC2 and PC3 respectively. Although the percentage of bracketing patterns in a poem has low impact on PC1, this attribute cannot be ignored because it exhibits a fair bit of influence on PC2 and PC3. As shown in table 9, PC2 and PC3 together account for 0.37 (0.23 + 0.14) or 37% of the variance, which is a significant fraction.

Figure 12 displays how this data might be plotted using PC1 and PC2 as the X and Y-axes. Since the direction of the vectors for brackets and initial auxiliaries is almost parallel to the Y-axis, it implies that these two attributes really only influence PC2. Principal and dependent clauses on the other hand have a greater impact on PC1 based on their directionality in this figure.

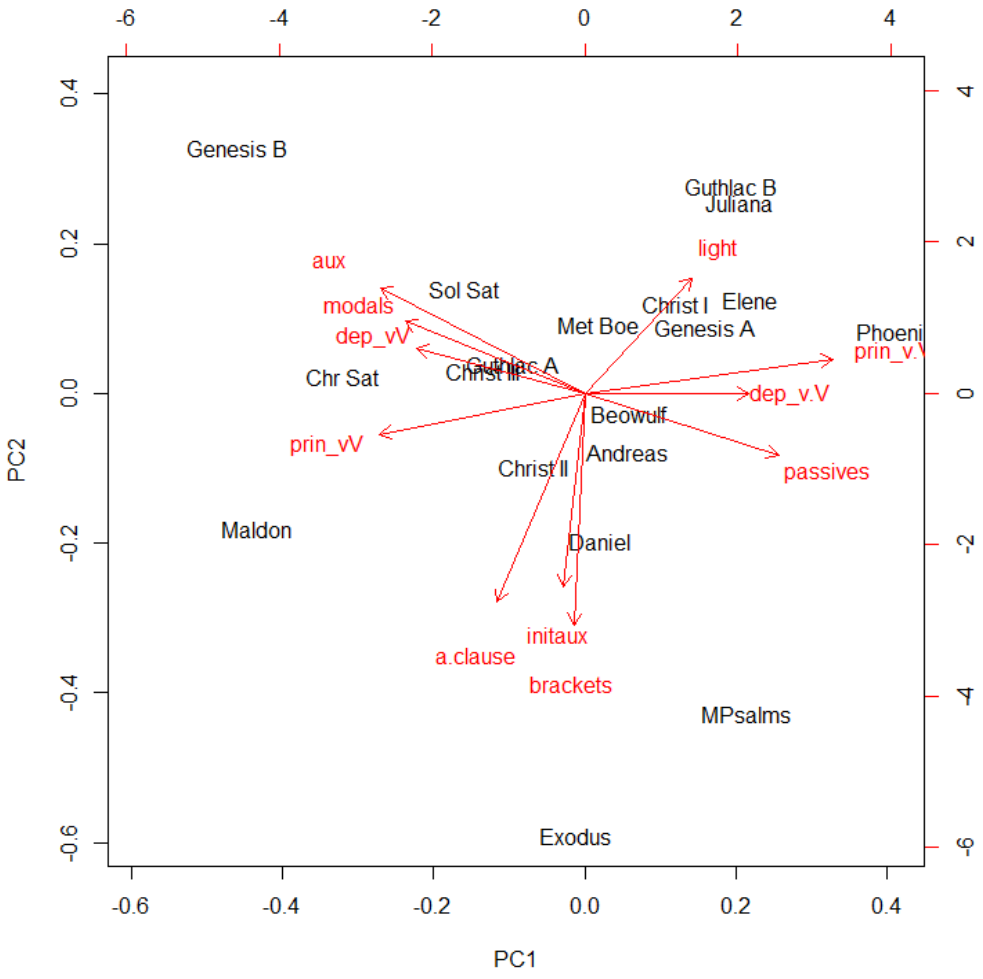

Figure 12. PCA output for 19 poems

Interpretation of Results

The *hierarchical* 5-cluster and *k-means* 4-cluster scenarios yielded similar clusters as indicated in table 11. Only three poems displayed an inconsistency between the two methods in terms of which clusters they appeared in. *Christ II* (*hierarchical:* Cluster D, *k-means:* Cluster A) and *Christ III* (*hierarchical:* Cluster A, *k-means:* Cluster D) essentially interchanged clusters for the two methods, while *Genesis B* grouped with *Elene, Guthlac B, Juliana,* and *Phoenix* for *k-means*, but appeared by itself in a separate cluster for *hierarchical*. Even in the latter case, *Genesis B* combines with the same poems at next level of agglomeration at a slightly larger value of "Height" (approximately 58), indicating that it is quite similar to this group—in the same family tree so to speak—and actually very dissimilar to every other poem since it combines with the other branches only at the final step at the top of the diagram (see figure 6).

Having two methods converging on a similar outcome holds some promise that this methodology is of value, although by no means does it guarantee that the results are determinate. With these results at hand, it is now possible to examine each cluster in order to understand what factors they have in common outside of testing the auxiliary. That is, whether or not the output of clustering is consistent with any other attributes of the poems. For the purpose of such comparisons additional aspects of the poems such as length (number of lines), source, and main topic were tabulated in table 12. (Also included in this table is *Fates of the Apostles* since this poem is part of the data set in the later analysis of Cynewulfian authorship.) Since absolute dates are difficult to

establish in Old English literature, a relative chronology is captured in this table based on Fulk (1992) who used the broad categories of "Caedmonian" (including *Beowulf*), "Cynewulfian" (including *Andreas*), and "Alfredian and later." These have been generalized as "Early," "Middle," and "Late" respectively in table 12. Such broad categories of time have also been used by other Old English scholars, such as Wilhelm Bode, who classified kennings as "pre-Cynewulf," "Cynewulf," and "later" (Calder 1979). Hence there is precedent for using this approach for categorizing this literature.

Table 11
Hierarchical vs. k-means Clusters

	Cluster Label	
Poem	Hierarchical	k-means
Andreas	A	A
Beowulf	A	A
Chr Sat	D	D
Christ I	A	A
Christ II	***D***	***A***
Christ III	***A***	***D***
Daniel	A	A
Elene	C	C
Exodus	B	B
Genesis A	C	C
Genesis B	***E***	***D***
Guthlac A	A	A
Guthlac B	C	C
Juliana	C	C
Maldon	D	D
Met Boe	A	A
MPsalms	B	B
Phoenix	C	C
Sol Sat	D	D

Table 12
Additional Attributes Outside of Auxiliary Verb Data

Poem	Topic	Chronology	Close translation	No. of lines
Andreas	Saint's life	Middle		1722
Beowulf	Heroic	Early		3182
Chr Sat	Biblical	Middle		729
Christ I	Religious	Middle		439
Christ II	Religious	Middle		427
Christ III	Religious	Early		798
Daniel	Biblical	Early		764
Elene	Saint's life	Middle		1321
Exodus	Biblical	Early		590
Fates	Religious	Middle		122
Genesis A	Biblical	Early		2219
Genesis B	Biblical	Middle		617
Guthlac A	Saint's life	Early	Latin	818
Guthlac B	Saint's life	Middle	Latin	561
Juliana	Saint's life	Middle	Latin	731
Maldon	Historical	Middle		325
Met Boe	Philosophical	Late	Latin	1750
MPsalms	Religious	Late	Latin	5040
Phoenix	Religious	Middle	Latin	677
Sol Sat	Wisdom	Late		506

Cluster A contains the following poems: *Andreas*, *Beowulf*, *Christ I*, *Daniel*, *Guthlac A*, and *Meters of Boethius.* Depending on which clustering method is used *Christ II* or *Christ III* can also be incorporated into this cluster. At first glance, the appearance of *Andreas* and *Beowulf* in the same cluster is quite encouraging, since the influence of the latter on the former has been noted by many authors. As Riedinger (1993) observed "…the *Andreas* poet borrowed frequently and methodically from *Beowulf*." (283) Also, Clark (2014) offered the opinion that *Andreas* was "a poem almost certainly influenced by *Beowulf*." (226) It is conceivable that it is this influence and borrowing, whatever its extent, which accounts in part for these two works to be grouped together in this clustering exercise. Borrowing is also consistent with the relative chronology since *Beowulf* is the earlier of the two.

Christ I, Christ II, Christ III, Guthlac A, and Daniel fit in this cluster in that they are all chronologically either "early" or "middle." The first four also have in common the fact that they appear in the Exeter Book. This co-occurrence is probably not that meaningful in and of itself since the Exeter book is after all the largest collection of Old English literature, so there is always a chance that any two books would appear in it. However, one interesting point is that *Guthlac A* appears immediately after the three *Christ* books. Could the person who put together this anthology used this order deliberately due to similarities in style or authorship? Another possibility is that *Guthlac A* was placed after Christ III in the manuscript that the Exeter scribe copied, and the (presumed) motivation for copying them consecutively rests with the scribe of the exemplar, not the Exeter scribe. The variable from tables 1 and 2 that seems to have a strong impact on this cluster is the a-clause which ranges in value from 49 (*Christ III*) to

59 (*Met Boe*) for the bulk of them, with *Daniel* being the outlier at 70. Poems outside this group have a-clause values of 42 and under or 63 and over. (Recall that a-clause is part of a complementary set, so this association could have been alternatively stated using the other part of the set, in this case the b-clause. This statement would apply to any other complementary sets as well.)

Cluster B comprises *Exodus* and *Metrical Psalms*. No commonality is apparent here, and the differences are quite stark. The former is chronologically early, and the latter is quite late. In fact, these two works are substantially different in terms of length, with the latter containing 10 times as many lines as the former. The variable values that link these poems are aux (14.8 and 12.2), a-clause (71 and 81), and, most strongly, the number of bracketing patterns (53.1 and 50.6). In fact, these are the only two poems that have bracketing pattern values of greater than 50; the next highest value is only 36.4 (*Maldon*).

Elene, *Genesis A*, *Guthlac B*, *Juliana*, and *Phoenix* make up cluster C. If the subject matter is taken into consideration, two subgroups can be identified within this cluster. *Genesis A* of course tells the story of the beginning of mankind from an Old Testament perspective, and describes events in the Garden of Eden, the fall of the angels, Noah's flood, etc. (Anlezark 2011). *The Phoenix* also references many of these Biblical themes so could potentially fit alongside *Genesis A* in that regard. The obvious common thread that runs through the other three poems is that they all deal with Saint's lives, which can explain why they group together. Two poems in this cluster 3 both have a "heroic" theme in common: the *Genesis A* poet "...only treats the heroic aspects of the story" when describing Abrahamic events (Doane 1978), and there is a heroic, warrior-

like element of the eponymous main character of *Elene.* The variable with the strongest association with this cluster is the a-clause, which varies from 39 (*Phoenix*) to 42 (*Genesis A*), with *Guthlac B* as the odd one out at 26. All other poems have a-clause values greater than or equal to 49 except *Genesis B* which is at 38.

Christ and Satan, *Genesis B*, *Battle of Maldon*, and *Solomon and Saturn* constitute the last group, cluster D. This group is quite heterogeneous; its only commonality is the relative chronology—all four are "middle" or "middle-to-late." Passives have the strongest association with this cluster, with values varying from 14 (*Genesis B*) to 18 (*Solomon and Saturn*). Although all other poems are at 19 or above, this distinction is not that marked, since there are several with values in the range of 19-22.

Overall, the clustering analysis of the 19 poems produced mixed results, with no clear boundaries demarcating the groups that were generated. Within each group the relative chronology varies, as does the subject matter. The number of lines in the poems has a wide range within each cluster. Only within subgroups in the few cases described above is there any sort of conformity. For these reasons, it should be noted that the analysis thus far does not reveal distinctive schools of style and/or authorship.

Chapter V

The Cynewulfian Authorship Question

To this day we do not know the names of most Old English poets; the handful that are known include the Venerable Bede, Cædmon, and Cynewulf, as noted by Bjork (2013), who in tracing the history of Cynewulf observes that the poet's name famously appears at the end of four poems (*Christ II, Elene, Fates of the Apostles,* and *Juliana*), spelled variously in runic letters as "Cynwulf" or "Cynewulf." In fact, no two of his signatures are exactly alike (Niles 285). Bjork cites Stodnick's (1997) speculation that there was no actual person named Cynewulf (viii), but goes on to write that if such a person existed he must have lived "between the mid-eighth and the late tenth centuries" (x), and opines that in addition to the four signed poems, *Guthlac B* is also likely to have been written by this poet (xi).

In his 1987 study that provides the basis for this thesis, Donoghue also tried to apply the test of the auxiliary in order to establish whether or not the four signed poems of Cynewulf have a style distinct from other Old English poetry or if they indeed share the same style. If such a distinction exists it would add to the evidence that they were composed by the same author. He approached this problem by comparing the auxiliary verb data for each of the signed poems to several others which were included "to provide non-Cynewulfian points of contrast," (108) examining a total of 14 poems. The sets of attributes used for analysis were further refined by including only those least influenced

by subject matter of the poem, which narrowed the list to the following six (108):

1. Number of auxiliaries per 100 lines of the poem.
2. Percentage of auxiliaries that are clause-initial.
3. Number of bracketing patterns per 100 initial auxiliaries.
4. Percentage of auxiliaries that occur in a-clauses vs. b-clauses.
5. Distribution of the three word-orders for principal clauses.
6. Distribution of the three word-orders for dependent clauses.

This chapter tackles the same questions with the goal of answering them by applying machine learning clustering methods to the same raw data. As Donoghue observed, "The exact size of the [Cynewulfian] canon has dramatically grown and shrunk since the discovery of the runic signatures." (106) While the loose current consensus among scholars is that Cynewulf is the author of the four signed poems and *Guthlac B*, the possibility of including other poems in the canon remains. This study may shed some light on this point by identifying additional potential candidates to be added to the Cynewulfian canon.

Hierarchical Clustering

The 14 poems under study were grouped into 2 to 10 clusters by running the computer program several times. The results are captured in table 13, with the four signed Cynewulfian poems marked as bold italics. As with the previous analysis of 19 poems, these results also show consistency for some poems (*Andreas*, *Beowulf*, and *Daniel* are always or almost always in the same group) and a fair bit of variance for others (*Phoenix*

lands in as many as six groups).

The “family tree” dendrogram generated by *hierarchical clustering* (figure 13) shows three or four branches to be a reasonable number. Figures 14 and 15 display these groupings.

Table 13
Hierarchical Clusters: Cynewulf

	Number of Clusters								
Poem	2	3	4	5	6	7	8	9	10
Andreas	1	1	1	1	1	1	1	1	1
Beowulf	1	1	1	1	1	1	1	1	1
Christ I	1	1	1	1	2	2	2	2	2
Christ II	***1***	***2***	***2***	***2***	***3***	***3***	***3***	***3***	***3***
Christ III	1	1	1	1	2	4	4	4	4
Daniel	1	1	1	1	1	1	1	1	5
Elene	***2***	***3***	***3***	***3***	***4***	***5***	***5***	***5***	***6***
Exodus	1	2	2	2	3	3	6	6	7
Fates	***2***	***3***	***4***	***4***	***5***	***6***	***7***	***7***	***8***
Guthlac A	1	1	1	1	1	1	1	1	1
Guthlac B	2	3	3	3	4	5	5	8	9
Juliana	***2***	***3***	***3***	***3***	***4***	***5***	***5***	***8***	***9***
Maldon	1	2	2	5	6	7	8	9	10
Phoenix	2	3	3	3	4	5	5	8	9

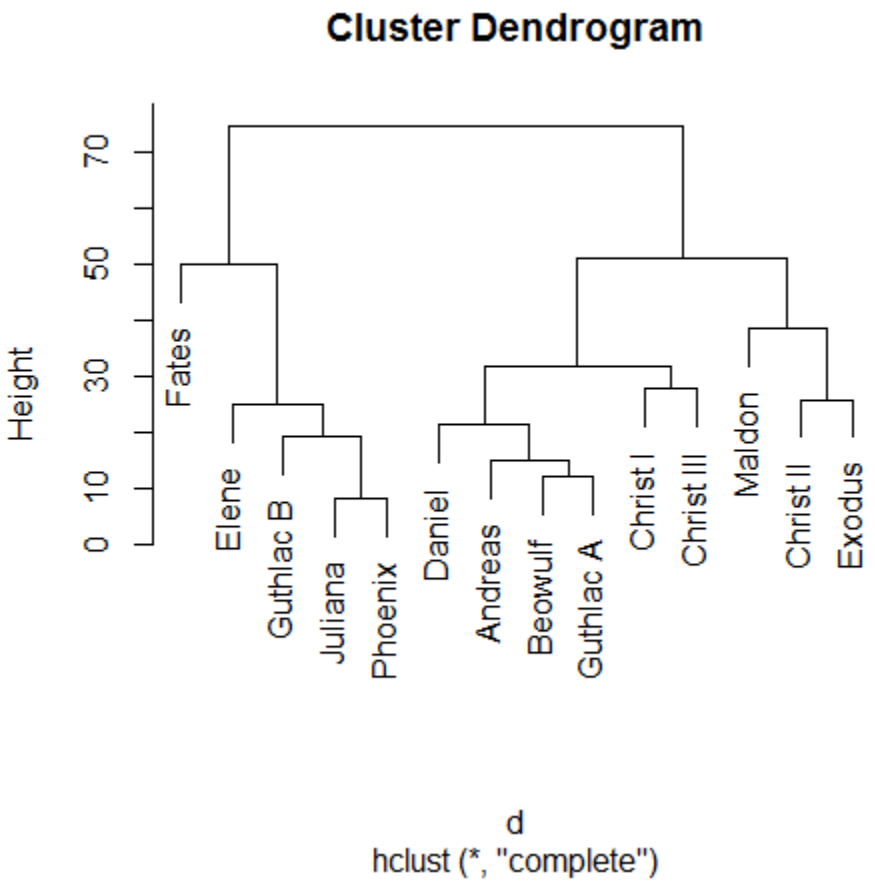

Figure 13. Hierarchical cluster dendrogram

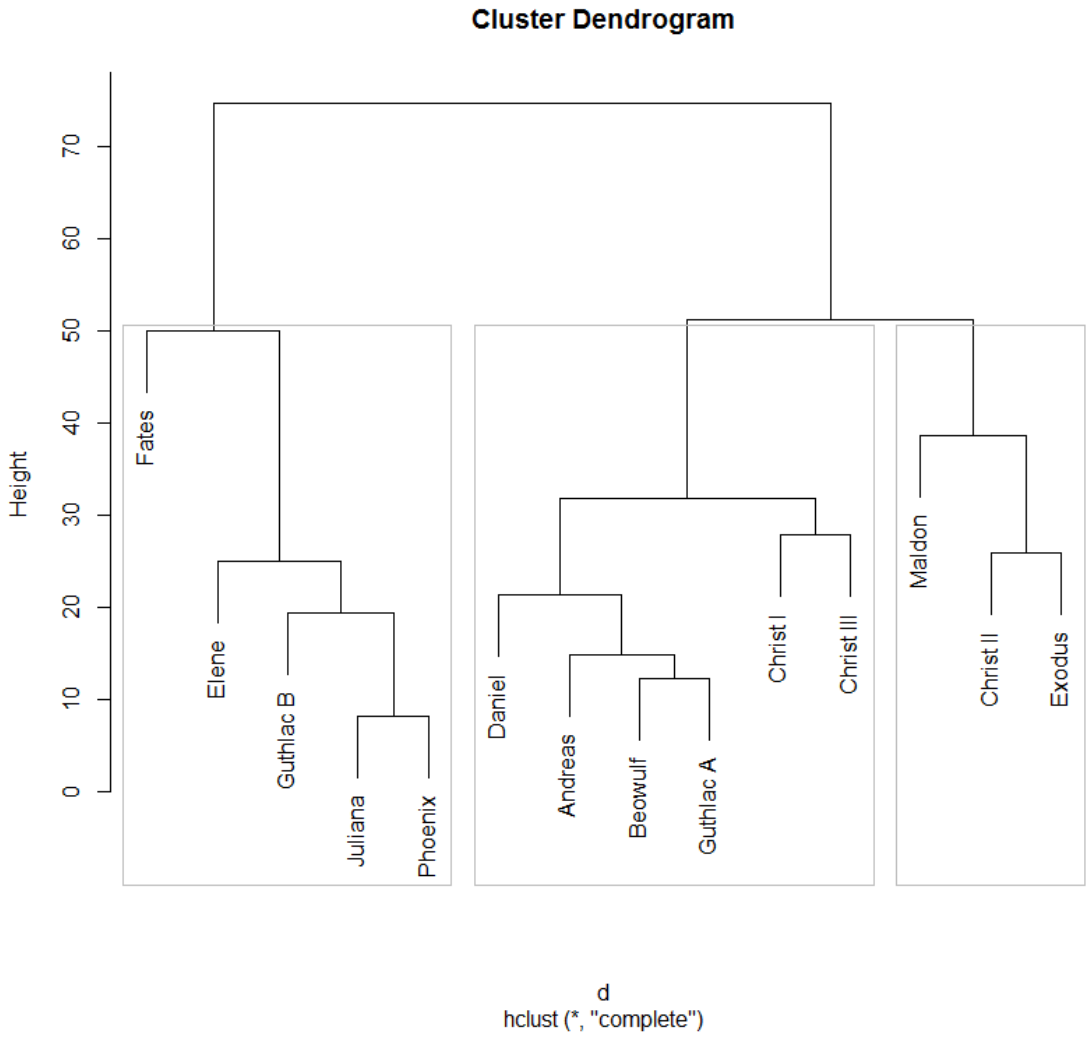

Figure 14. Hierarchical cluster dendrogram: Three clusters

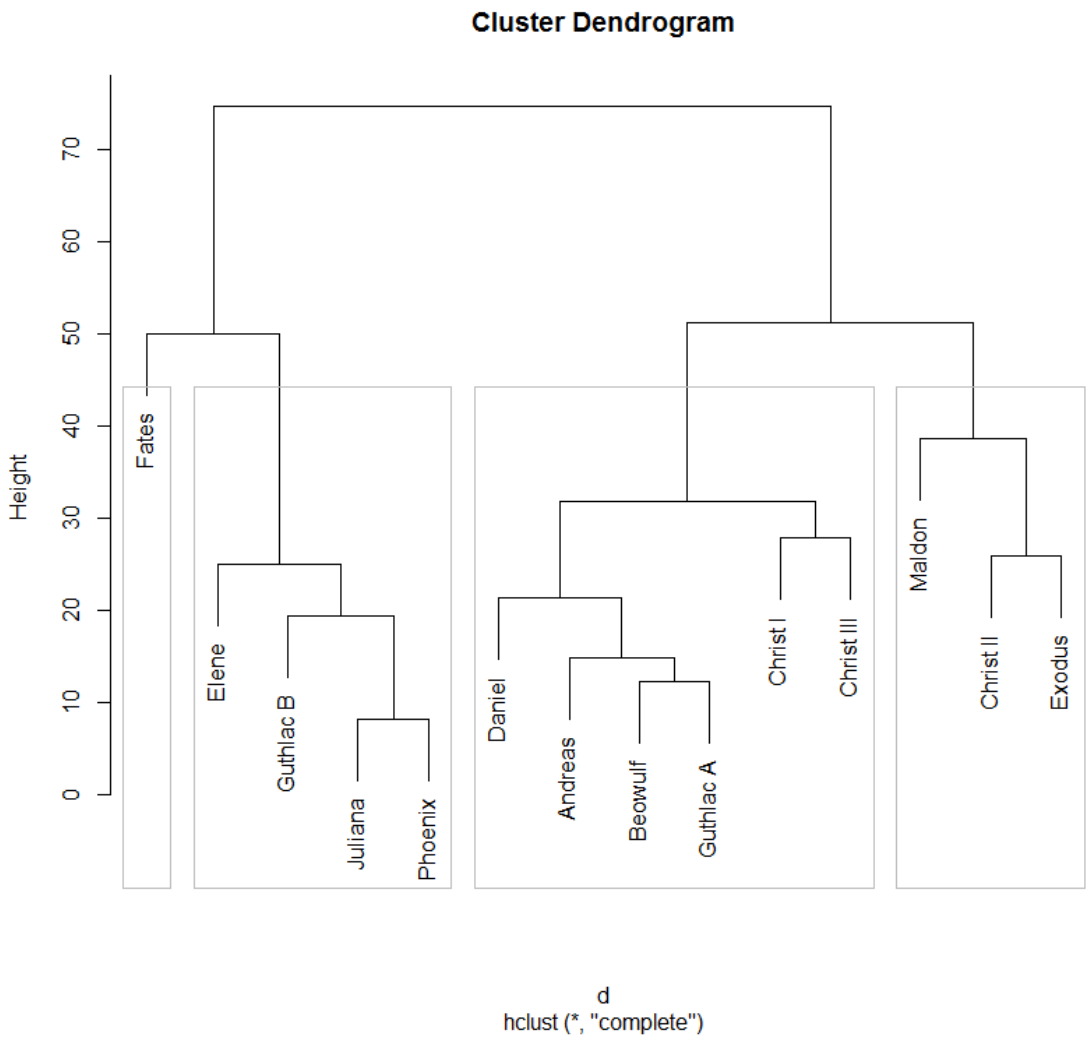

Figure 15. Hierarchical cluster dendrogram: Four clusters

k-means Clustering

This algorithm was run for values of *k* from 2-10 in order to identify the "elbow." The output is captured in table 14 and shown graphically in figure 16. Table 15 shows how the poems are grouped for the various cluster sizes.

Table 14
Variation of Sum of Squared Error (SSE) as the Value of k Increases

k	SSE
2	6190
3	4503
4	2980
5	2158
6	1706
7	1340
8	1007
9	692
10	420

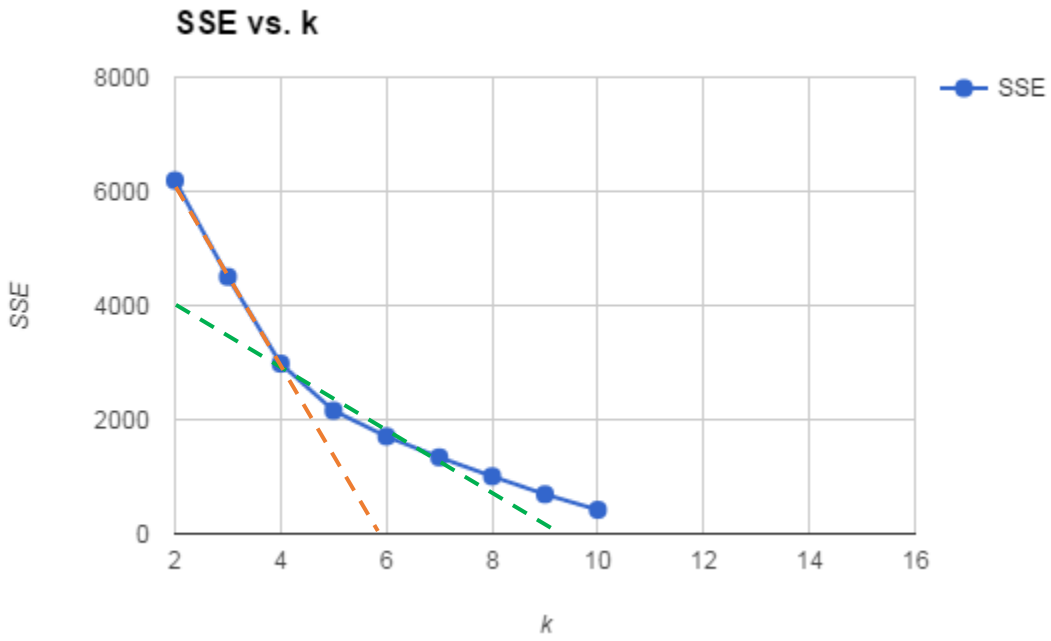

Figure 16. Graph of Sum of Squared Error (SSE) vs. k

Visually one can discern an "elbow" at a value of *k*=4 and this is borne out by computing the SSE difference in table 14. SSE decreased by 1687 between *k*=2 and *k*=3, and by 1523 between *k*=3 and *k*=4. Beyond a *k* value of four the rate of change of SSE is slower, for example it is only 822 between *k*=4 and *k*=5 with the curve further flattening out after that. Hence, a *k* value of four is reasonable for this computation, and table 16 illustrates how the poems cluster in this scenario. Table 17 captures the clusters for *k*=5 with only one poem, *Maldon*, moving to an entirely new group lending confidence to using *k*=4 for interpreting the results. Figures 17 and 18 show the cluster plot for *k* values of 4 and 5 respectively.

Table 15

Cluster Size Variation

No. of Clusters	Cluster sizes
2	4, 10
3	3, 4, 7
4	1, 3, 4, 6
5	1, 1, 2, 4, 6

Table 16

k-means: Four Clusters of Sizes 1, 3, 4, 6

Cluster	Poems					
1	Andreas	Beowulf	Christ I	Christ III	Daniel	Guthlac A
2	Christ II	Exodus	Maldon			
3	Elene	Guthlac B	Juliana	Phoenix		
4	Fates					

Table 17

k-means: Five Clusters of Sizes 1, 1, 2, 4, 6

Cluster	Poems					
1	Andreas	Beowulf	Christ I	Christ III	Daniel	Guthlac A
2	Christ II	Exodus				
3	Elene	Guthlac B	Juliana	Phoenix		
4	Fates					
5	Maldon					

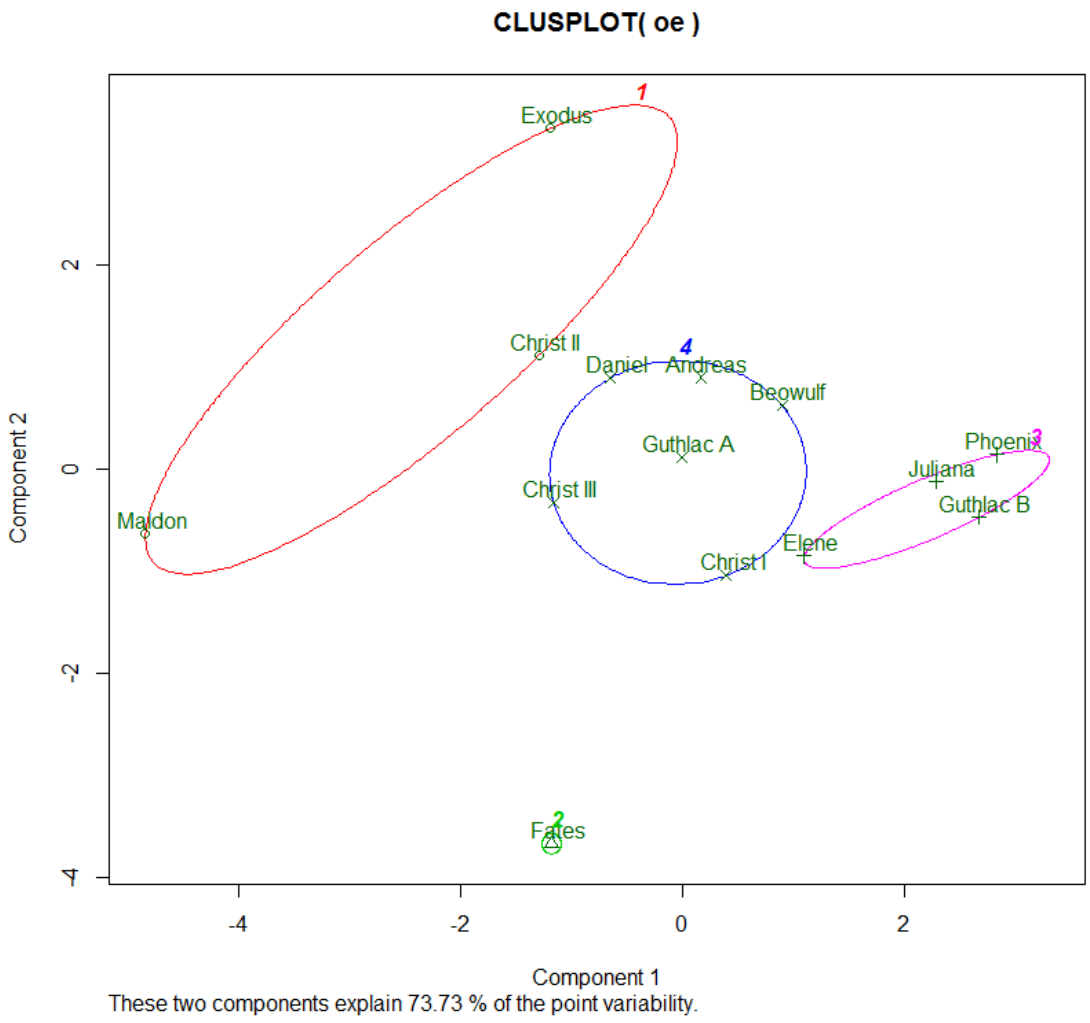

Figure 17. k-means clusters using a k value of 4

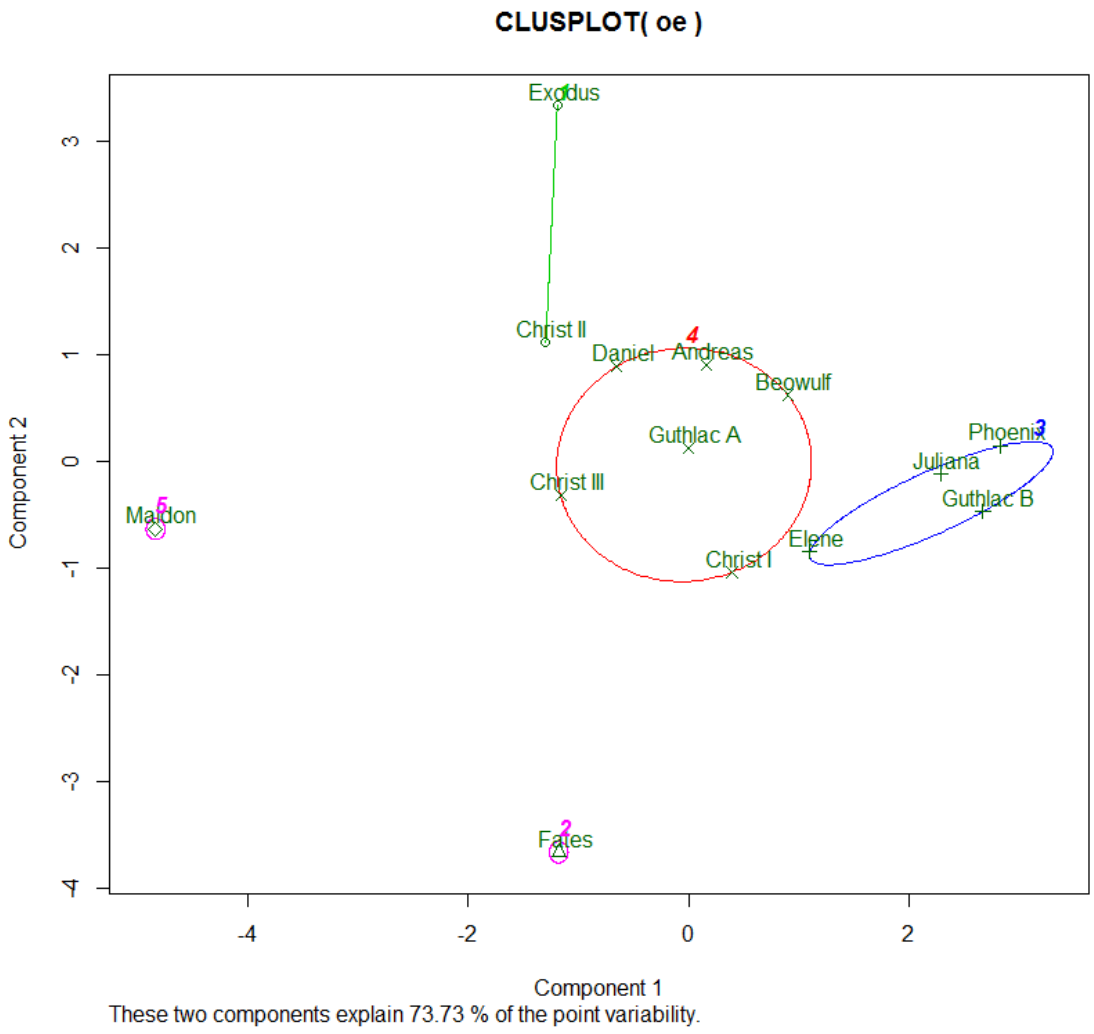

Figure 18. k-means clusters using a k value of 5

Principal Component Analysis (PCA)

PCA calculations yielded eight components (PC1-PC8) shown in table 18. PC1 and PC2 have relatively high values of standard deviation (SD). Together they account for 74% of the cumulative proportion, with PC1 at 0.47 (47%) being much higher than PC2 at 0.27 (27%). PC3 through PC8 each are at 10% or less. The scree plot in figure 19 visually indicates that PC1 has the most influence and dwarfs PC2. PC3-PC8 are less important in this regard so a cutoff after the first two components is reasonable.

Table 18
Summary PCA Statistics: Importance of Components

Importance	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.93	1.47	0.88	0.73	0.64	0.51	0.27	0.23
Proportion of Variance	0.47	0.27	0.10	0.07	0.05	0.03	0.01	0.01
Cumulative Proportion	0.47	0.74	0.83	0.90	0.95	0.98	0.99	1.00

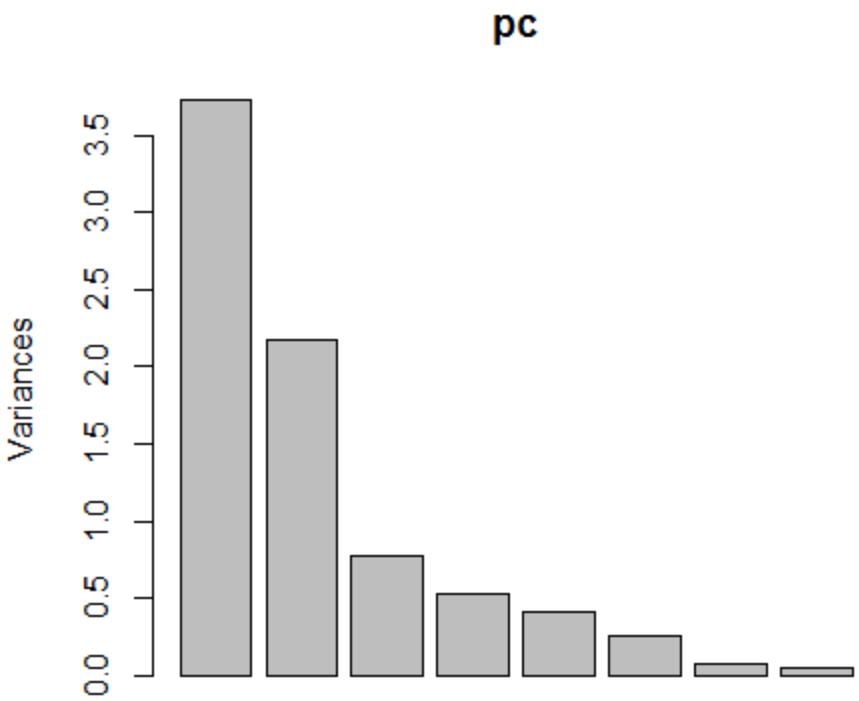

Figure 19. Scree plot of principal components

Table 19 displays the makeup and weights of the principal components' attributes and figure 20 plots this data in two dimensions. This is a very busy plot, as the poems clustered closely together which forces the labels to overlap. The data in this table and figure indicate that again all the attributes have a significant influence on at least PC1 or PC2 and therefore cannot be ignored in this analysis.

Table 19
Relative Weights of Attributes in PCA

Rotation	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
aux	-0.32	-0.12	0.79	-0.41	0.01	-0.23	-0.04	0.18
initaux	-0.09	0.60	0.13	0.06	-0.58	-0.21	0.33	-0.34
brackets	-0.27	0.50	-0.15	-0.22	0.51	-0.16	-0.45	-0.34
a_clause	-0.42	0.24	0.11	0.37	0.44	0.30	0.54	0.22
prin_vV	-0.41	-0.06	-0.55	-0.37	-0.13	-0.38	0.22	0.41
prin_ýV	0.47	-0.09	0.04	-0.14	0.43	-0.50	0.52	-0.24
dep_vV	-0.36	-0.43	-0.10	-0.29	-0.04	0.33	0.24	-0.65
dep_ýV	0.35	0.34	-0.02	-0.64	0.01	0.54	0.17	0.20

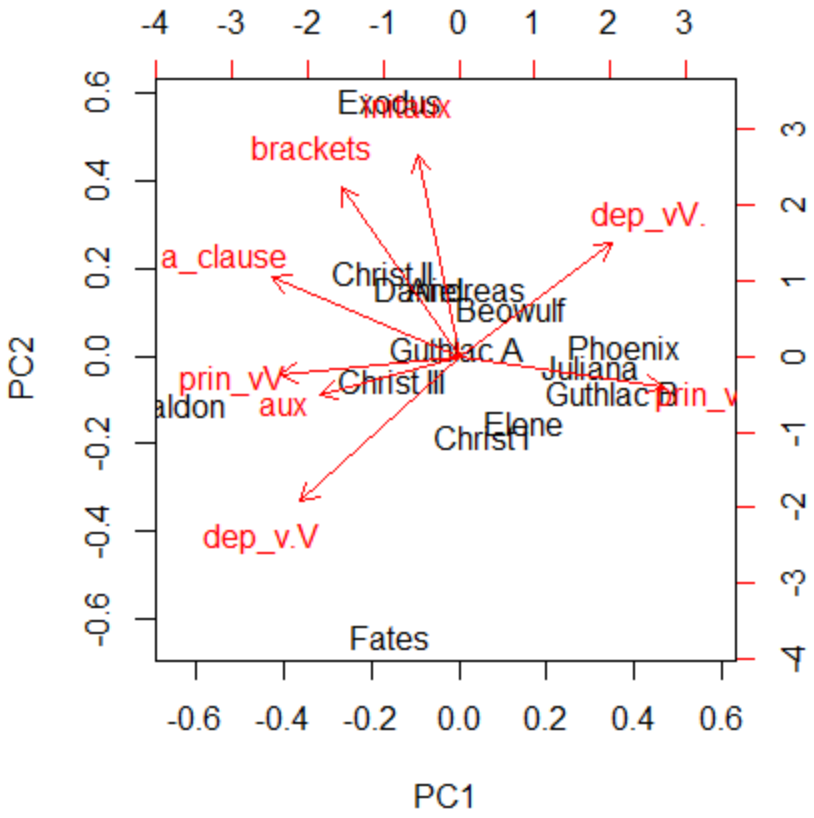

Figure 20. PCA output for 14 poems

Additional Scenarios

Two other permutations of the data were computed as an additional check. The first used the same six attributes but included all 20 poems to see if expanding the input data caused any new patterns to emerge. The *hierarchical clustering* output for four clusters is shown in figure 21 and the *k-means clustering* plot for *k* value of 4 is presented in figure 22, which was compared with figure 15 in order to identify similarities and differences.

While there are some discrepancies, such as the sub-cluster of *Fates* and *Genesis B* switching to an entirely new branch, the compositions of most of the groups remain the same. *Juliana* and *Phoenix*, for instance, merge with *Guthlac B* and *Elene* as before. The noticeable consistency of the order in which the poems combine as well as in the "height" along the Y-axis at which they merge within a group is a strong indicator of their similarity. The cluster comprising *Guthlac A*, *Beowulf*, *Andreas*, and *Daniel* is exactly the same as in figure 15, as is the one with *Exodus*, *Christ II*, and *Maldon*.

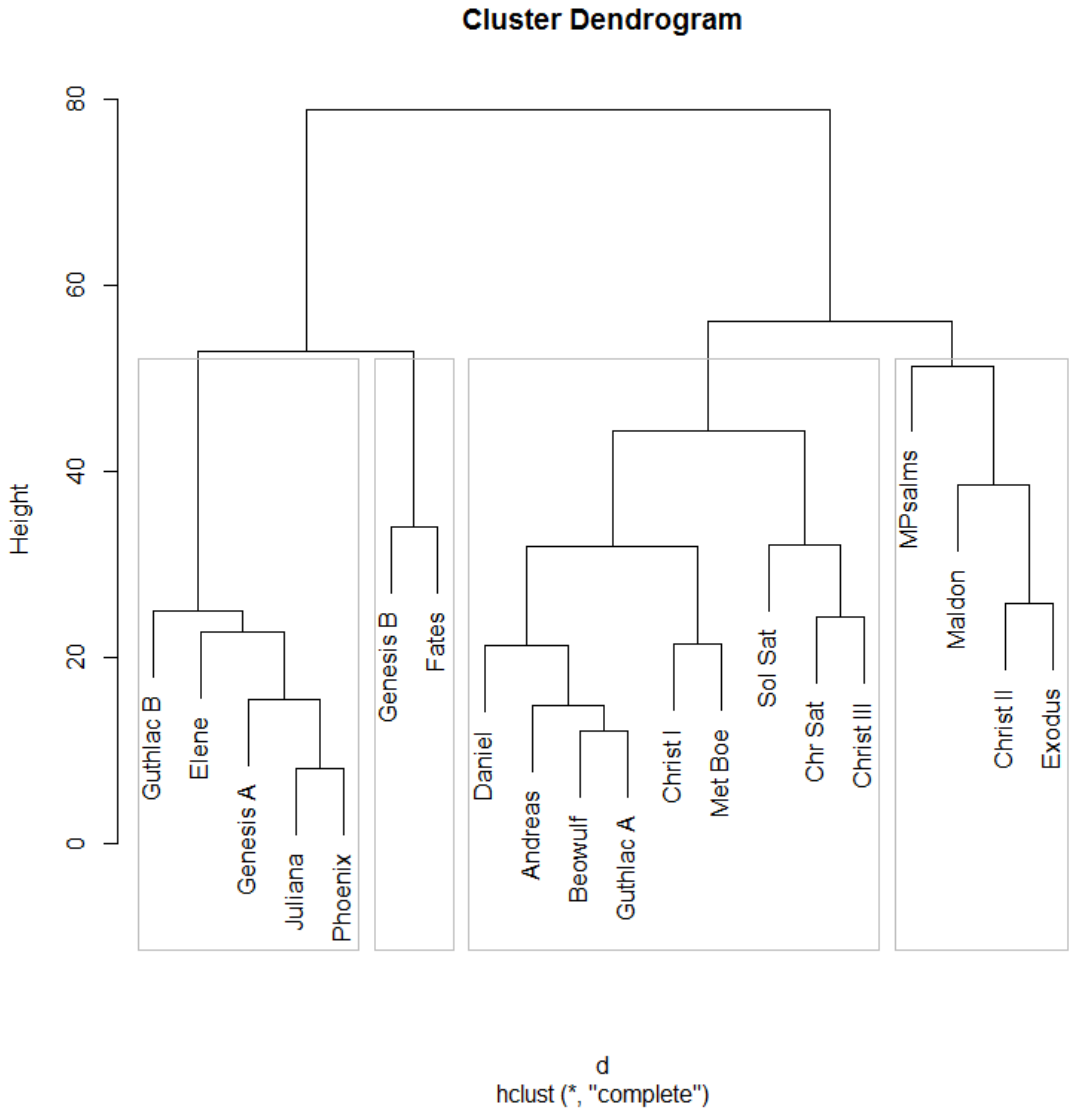

Figure 21. Dendrogram: 20 poems using six attributes with 4 clusters

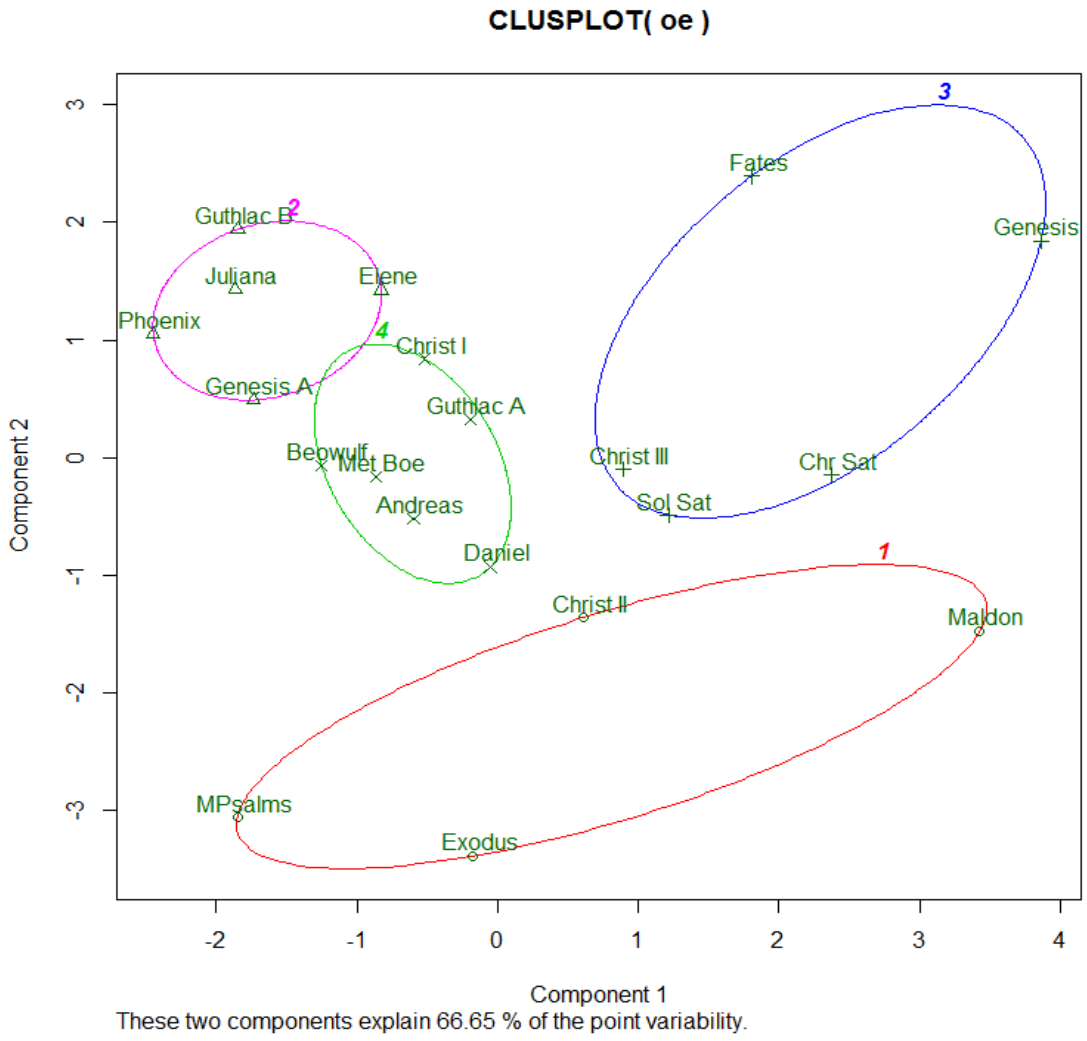

Figure 22. k-means: 20 poems using six attributes with 4 clusters

The second additional scenario excluded *Fates* and thus analyzed the other 19 poems using six attributes. This was done to evaluate whether *Fates,* a very short poem of only 122 lines, contributed value to the analysis in terms of the level of impacting groupings. The outcome of *hierarchical clustering* is shown in figure 23, which indicates that the hierarchy is exactly the same as for the 20-poem scenario of figure 21 except that *Fates* simply drops out of its slot. So it appears that *Fates* has no effect on how the other poems are grouped.

These results of these two additional permutations demonstrate that the clusters that remained the same were relatively stable when several more poems outside of the Cynewulfian canon were thrown into the mix. Therefore, the output of the original scenarios (tables 12, 15, and 17) can be used to interpret the results and this is discussed in the next section.

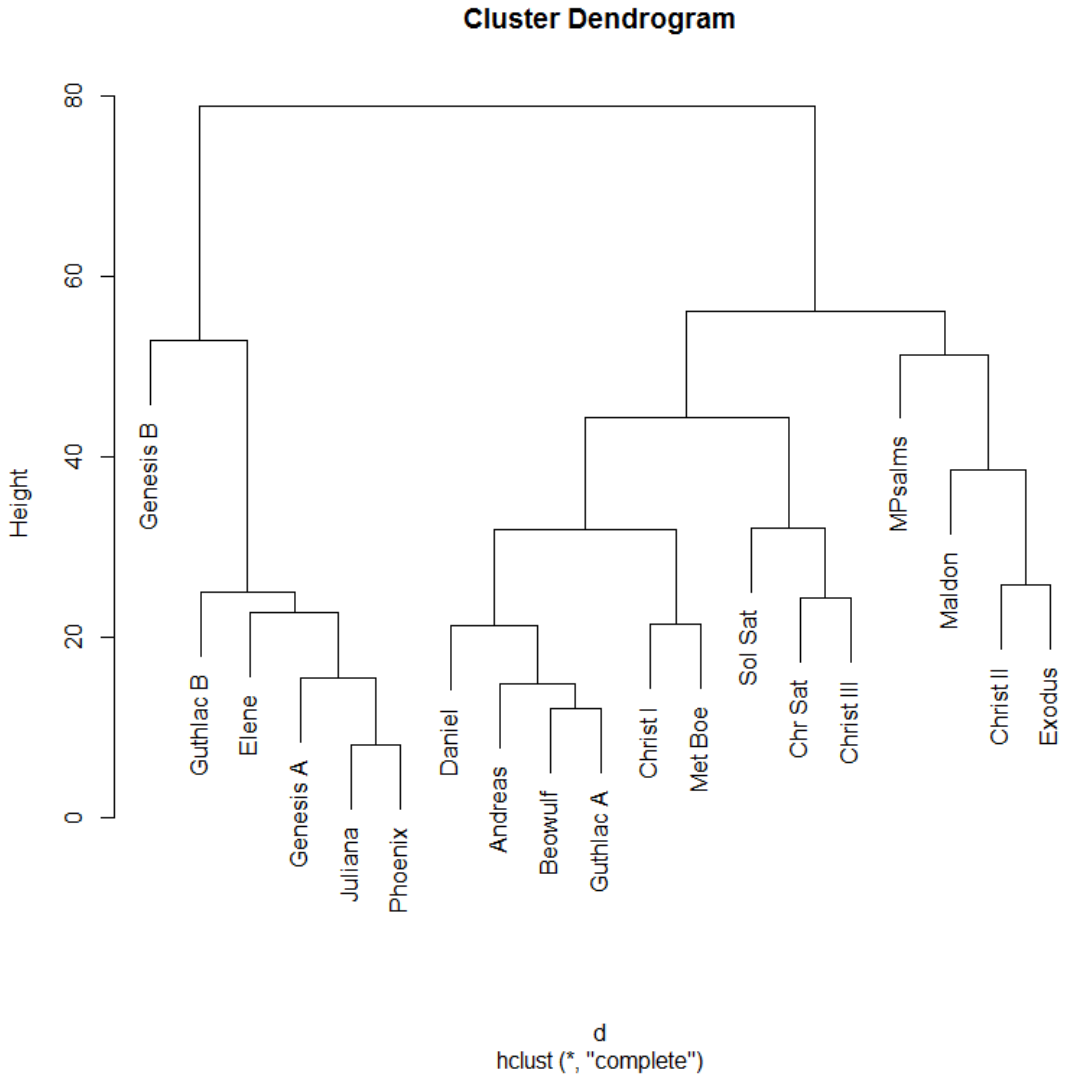

Figure 23. Dendrogram: 19 poems using six attributes

Interpretation of Results

For both methods, the data clustered into four groups in an identical manner (see table 20). Cluster relabeling is not required here since by chance the labels for the two methods are an exact match. Such consistency between two algorithms is encouraging as it indicates that this approach is sound.

Table 20
Hierarchical vs. k-means: Four clusters

	Cluster Number	
Poem	Hierarchical	k-means
Andreas	1	1
Beowulf	1	1
Christ I	1	1
Christ II	2	2
Christ III	1	1
Daniel	1	1
Elene	3	3
Exodus	2	2
Fates	4	4
Guthlac A	1	1
Guthlac B	3	3
Juliana	3	3
Maldon	2	2
Phoenix	3	3

Fates is the lone member of Cluster 4, however it is possible to examine each of the other three groups in turn to see which attributes, if any, have a strong association with a given cluster. Cluster 1 comprises *Andreas*, *Beowulf*, *Christ I*, *Christ III*, *Daniel*, and *Guthlac A*. The variable which corresponds well with this group is brackets with values varying from 17.1 (*Guthlac A*) to 28.1 (*Christ I*). Other poems are at 16.7 and lower or 36.4 and higher. To a lesser extent, a-clause also can be linked to this cluster, with values ranging from 49 (*Christ III*) to 70 (*Daniel*). *Fates* and *Christ II* slip into this group if only a-clause values are considered, as they are at 50 and 52 respectively. Other poems are at 40 and below or 71 and above.

Cluster 2 consists of *Christ II*, *Exodus*, and *Maldon*. Once again, brackets is the connecting attribute at 36.4, 45, and 53.1 respectively. If *Fates* (67) is discounted, the attribute prin_vV also sets this group apart, as these three poems have values of 63, 67, and 70. All others are at 59 or less. The co-occurrence of *Exodus,* which is an early poem with a religious theme, and *Maldon,* which is a late poem with a secular theme, is unexpected and thus noteworthy. (Breeze 1999 has remarked that the word *laerig,* which is derived from a Welsh loan word and has the meaning "shield rim," occurs only in these two works, but this hardly seems relevant in the context of poetic style.) The clustering analysis of the 19 poems in earlier chapters placed these poems in two entirely different sub-clusters which did not combine until very late (see figure 7). The loss of information caused by using fewer variables has thus proven to be quite significant in this case.

Cluster 3 is made up of *Elene*, *Guthlac B*, *Juliana*, and *Phoenix*. This group stands out by the sheer number of individual attributes which associate strongly with it. These attributes are four in number and include initaux (varying from 19 to 22), brackets (4.8 to

16.7), a-clause (26 to 40), and prin_ýV (41 to 45). Other poems fall well outside these ranges.

One point of similarity that jumps out from the output is how strongly in lockstep *Elene* and *Juliana* are in the *hierarchical clustering* data captured in table 13. When this set is clustered into 2 to 8 groups, these two poems occur in exactly the same groups (2, 3, 3, 3, 4, 5, 5, 5, and 6), showing variation only when the number of clusters is 9 or 10, while the other two signed poems are not as closely aligned. This data correlates well with Donoghue's conclusion that "It is easy to see the hand of a single poet in [*Elene* and *Juliana*]" (108).

In table 13, *Fates* and *Christ II* display a pattern that differs from *Elene* and *Juliana,* placing them in a completely different group for the *k* value of four (group 2 for *Christ II* and group 4 for *Fates*). *Christ II* appears to be much closer to *Exodus* and *Maldon* than it is to the other signed poems, while *Fates* is closer but still diverges to a lesser degree. This is also revealed in the hierarchical family tree of figures 13-15 which show that *Fates* is clearly in the same branch as *Elene* and *Juliana* (albeit combining with them at a higher level) whereas *Christ II* is in an altogether separate branch. This suggests that perhaps *Christ II* is not as Cynewulfian in style as one might expect and therefore bears further investigation.

The surprise in this cluster is *Phoenix,* which is usually considered to be non-Cynewulfian, but follows the cluster distribution pattern of *Guthlac B*, *Elene,* and *Juliana* in table 13, and they all co-occur in the same group (Cluster 3) for *k-means* cluster sizes of three and four. To explain this unexpected similarity, one must reach deeper into the raw data used as input into the clustering algorithms. Donoghue offers one possible

explanation, writing "The low proportion of auxiliaries throughout *Phoenix* and the very high proportion of *b* clauses in *Guthlac B* set them apart from *Elene* and *Juliana*." (212) He goes on to remark that since the end of *Guthlac B* is missing in the manuscript, it leaves open the possibility that Cynewulf's runic signature might have appeared there (212). Additionally, as the reader may recall, Bjork (2013) actually considered *Guthlac B* to have been written by Cynewulf (xi) based on the work of, among others, Orchard (2003). In fact, based on his analysis of shared formulae among the poems in question, Orchard links Cynewulf's signed poems to most strongly to *Andreas*, but also tentatively to *Phoenix*, *Christ III*, and *Guthlac A* and *B,* writing especially about the last mentioned "...on the basis of the existing evidence the notion that Cynewulf wrote *Guthlac B* is an attractive one." (294)

The findings of this study, however, are at odds with some of Orchard's conclusions. There is no support in the clustering output of a link, however tenuous, between the four signed poems and *Christ III* and *Guthlac A*. The machine learning algorithms consistently place *Christ III, Guthlac A,* and *Andreas* in a different cluster than the Cynewulfian works.

Orchard also proposes that the *Andreas* poet borrowed "extensively" from, and therefore this poem has a close connection with, *Christ II*, *Fates*, *Elene*, and *Juliana* (294). Clustering analysis results suggest otherwise. In none of the three methods used is there a hint that *Andreas* can be linked to any of these four poems. If *Andreas* borrowed from *Beowulf* in a very significant manner (as discussed in Chapter IV), and these two poems have exhibited a strong affinity to one another by appearing in the same group or sub-branch in all the clustering scenarios examined in this thesis, why then does *Andreas*

not also cluster with the four signed poems? Style based on auxiliaries and verbals might offer a clue in this situation. Borrowing of formulas is a deliberate, conscious act but it appears that the more subconscious choices made by the *Andreas* poet in the manner that he used auxiliaries and verbals were not affected by the large extent of borrowing from *Beowulf*. In other words, his style, a trace of which he leaves behind in word order, i.e., his subconscious auxiliary-verbal usage, appears to have trumped whatever formulas in *Beowulf* from which he may have taken inspiration. The borrowings identified by Orchard have few auxiliary-verbal constructs, which supports the presumption that their usage is subconscious and thus not a component of formulas.

The clusters identified by the machine learning algorithms and presented earlier in this chapter support the case for considering *Phoenix* and *Guthlac B* as strong candidates to be included in the Cynewulfian canon, and Orchard's conclusions are compatible with this. As the dendrogram in figure 15 shows, *Phoenix* combines with *Juliana* very early as we move up the Y-axis, *Guthlac B* merges with these two at the next level of agglomeration even before *Elene* joins this cluster, and *Fates* does not join this family until much later. The *k-means* clustering data in table 16 and figure 17 also support the argument that *Phoenix* and *Guthlac B* are closer to *Elene* and *Juliana* than is *Fates.* Finally, the *PCA* output in figure 20 plots *Phoenix*, *Juliana*, and *Guthlac B* practically on top of each other (indicating they are highly similar to each other), *Elene* is slightly further away while *Fates* falls at quite a distance. Based on these results, one can reasonably argue that *Phoenix* and *Guthlac B* are Cynewulfian, if not in authorship, at least in style.

Chapter VI

Conclusions

At first glance, the results discussed thus far appear to be inconclusive. This study hypothesized that there would be clear boundaries in style and that poems will fall into distinctive categories. Given that neat breaks between clusters of poems were not identified, one is tempted to take Shippey at his word that "Any piece of Old English verse is liable to resemble others…" (1972). After all, why else would these poems not be easily grouped by historical period, genre, author, and so on? Some modern scholars have come to the conclusion that "the Old English metrical tradition was not stable, but constantly evolving, as all living traditions tend to be" (Bredehoft 109), implying that poetic meter changed over time. Yet this study has analyzed poems composed over a period of a few centuries, finding no notable differences in style, but rather a general similarity across the data set. Therefore, the findings discussed in the previous chapters lend themselves to the conclusion that the Old English poetic tradition is actually a range or a mixture of styles.

As such, this is a vindication of the pre-modern poetic tradition in that the results confirm a continuity over a relatively long span of time, which may well confound the expectations of a modern individual living in the 21st century. This is consistent with Shippey's observation "Old English verse is strangely homogenous over a long period; this inner consistency is the result of a mode of composition not present in the modern

world…" (112). By contrast with modern society, even so mundane a gesture as signing a poem carried far more weight for Old English poets. Although Cynewulf added his name to his works, there was more to it than conceit, that runic signature moved beyond the expression of authorship to other, more substantive ends such as asking for prayers for his soul. In this context Bragg observes of Cynewulf that "...he, too credits God for his talent demonstrates a deeply held notion of the relative inconsequence of the individual artist." (31)

This study has found that Old English poetry manifested a range of styles in every historic period. This range is somewhat compatible with Foley's examination of Old English, ancient Greek, and South Slavic oral poetry, in which he writes, "Genres do leak in traditional verse…" and "Features from one poetic form turn up in others." (76) Dealing specifically with oral poetry of these three languages, Foley found that "Old English poetry illustrates the most widespread leakage between and among its traditional verse genres" and that "various genres annex diction and narrative patterns that…are shared at the level of the larger poetic tradition." (102) Similar "leakage" in the 19 poems analyzed in this study (which are not restricted to just oral poetry) might account for some attributes that occurred across identifiable clusters. (This will become evident a little later in this discussion through the presentation of heat map visualizations to describe the poems and clusters.) This study has shown as part of the *PCA* outputs that all the features (also referred to as attributes, variables, or dimensions) used for this type of machine learning analysis play a non-trivial role and that none may therefore be disregarded. By way of a hypothetical illustration, imagine for a moment that, by some stroke of good fortune, in a discovery as fortuitous as the finding of the Dead Sea Scrolls,

a treasure trove of hitherto undiscovered Old English poems were unearthed. If these verses were to be analyzed using the auxiliaries, one would have to use all of the features identified in Donoghue's 1987 study in order to do full justice to such an analysis. This assertion becomes clear upon examining the data in tables 9 and 18, and the accompanying discussions in the present thesis. To recapitulate that discussion, in the *PCA* computations performed in this study, the poem attributes that were used had a significant value in at least one of the most influential principal components. For example, in table 19, PC1 and PC2 were determined to be the most important components, and the attribute prin_vV has a low value of -0.06 under PC2 but has a fairly high value of -0.41 under PC1. For this reason, prin_vV must be included in any clustering analysis, whether of the set of poems examined in this study, or of any new poems that might be discovered in the future.

Figure 24 represents an attempt to illustrate this argument by approaching it in a different direction. It displays a colorized "heat map" view combined with the previously discussed *hierarchical clusters,* depicted as rows (i.e., along the Y-axis, effectively rotated 90 degrees counterclockwise from earlier diagrams). Colored rectangles in heat maps indicate the poem corresponding to that row and the color indicates the value of each attribute for that poem. The color key displays the mapping of values to colors along a spectrum: dark red signifies lowest values of attributes, orange slightly higher ones, yellow indicates average values, light green stands for moderately high numbers, and dark green represents the highest values. If a large number of adjacent blocks for a given attribute had the same color for a particular branch or sub-branch the inference would be that all poems in the branch in question typically have the same or similar values. If, for

instance, the sub-branch of *Exodus* and *MPsalms* were green for the attribute "aux," but the same attribute showed red for every other poem, we could state that *Exodus* and *MPsalms* had high percentages of total auxiliaries and that this was a distinctive feature of that cluster. In fact, no such clear distinctions are apparent in these two figures, yet these heat maps can still be used tell a story about the clusters in a general descriptive fashion. Based on figure 24, the clusters identified in Chapter IV can be described as follows:

Cluster A comprising *Andreas*, *Beowulf*, *Daniel*, *Christ I*, *Christ III*, *Guthlac A*, and *Met Boe* consists of poems having:

(i) Mostly very low values of dep_vV

(ii) Moderately low values of aux, passives, initaux, brackets

(iii) Moderately low to average values of prin_ýV and dep_ýV

(iv) Average values of modals

(v) Average to moderately high values of prin_vV

(vi) Mostly moderately high values of a-clauses and light syllables

Cluster B comprising *Exodus* and *MPsalms* consists of poems having:

(i) Low to moderately low values of dep_vV

(ii) Moderately low values of aux

(iii) Moderately low to average values of modals, passives, initaux, prin_ýV

(iv) Average values of brackets

(v) Average to moderately high values of light syllables, prin_vV, dep_ýV

(vi) High values of a-clauses

Cluster C comprising *Genesis A*, *Juliana*, *Guthlac B*, *Phoenix*, and *Elene*, consists of poems having:

(i) Mostly low values of dep_vV

(ii) Moderately low values of aux

(iii) Mostly moderately low to average values of brackets

(iv) Moderately low to moderately high values of modals, passives, initaux

(v) Mostly average values of a-clauses, prin_vV, prin_v́V, dep_ v́V

(vi) Moderately high values of light syllables

Cluster D comprising *Maldon*, *Chr Sat*, *Christ II*, *Sol Sat*, and loosely *Genesis B* (since it combines very late with this cluster) consists of poems having:

(i) Low to average values of dep_vV

(ii) Mostly very low values of aux

(iii) Moderately low values of passives, initaux, prin_v́V

(iv) Moderately low to average values of brackets, dep_ v́V

(v) Average to moderately high values of light syllables

(vi) Mostly moderately high to high values of a-clauses

(vii) Moderately high values of modals, prin_vV

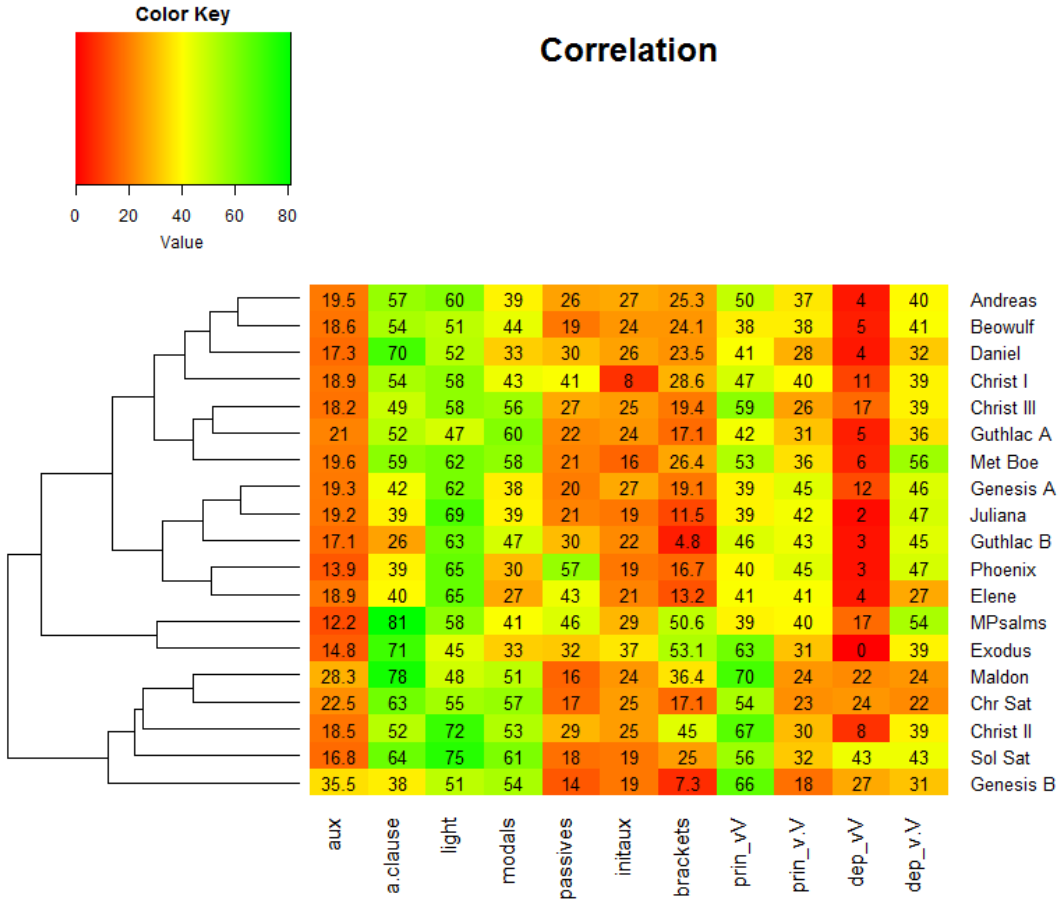

Figure 24. Heat map for 19 poems

Figure 25 displays a similar heat map for the 14 poems used in the Cynewulfian analysis and the clusters identified in Chapter V can be described as follows:

Cluster 1 comprising *Andreas*, *Beowulf*, *Daniel*, *Christ I*, *Christ III*, and *Guthlac A* consists of poems having:

(i) Low to moderately values of dep_vV

(ii) Moderately low values of aux, initaux, brackets

(iii) Moderately low to average values of prin_ýV

(iv) Mostly average values of dep_ ýV

(v) Moderately high values of a-clauses and prin_vV

Cluster 2 comprising *Exodus, Christ II* and *Maldon* consists of poems having:

(i) Low to moderately low values of dep_vV

(ii) Moderately low values of aux, prin_ýV

(iii) Moderately low to average values of initaux, dep_ ýV

(iv) Average to moderately high values of brackets

(v) Moderately low to moderately high values of

(vi) Moderately high to high values of a-clause, prin_vV

Cluster 3 comprising *Elene, Juliana, Phoenix, Guthlac B,* and loosely *Fates* (since *Fates* is in a group by itself but it eventually combines with this group) consists of poems having:

(i) Mostly low values of dep_vV

(ii) Mostly moderately low values of aux, initaux, brackets

(iii) Moderately low to average values of prin_ýV

(iv) Average to moderately high values of prin_vV

(v) Moderately low to moderately high values of a-clauses, dep_ ýV

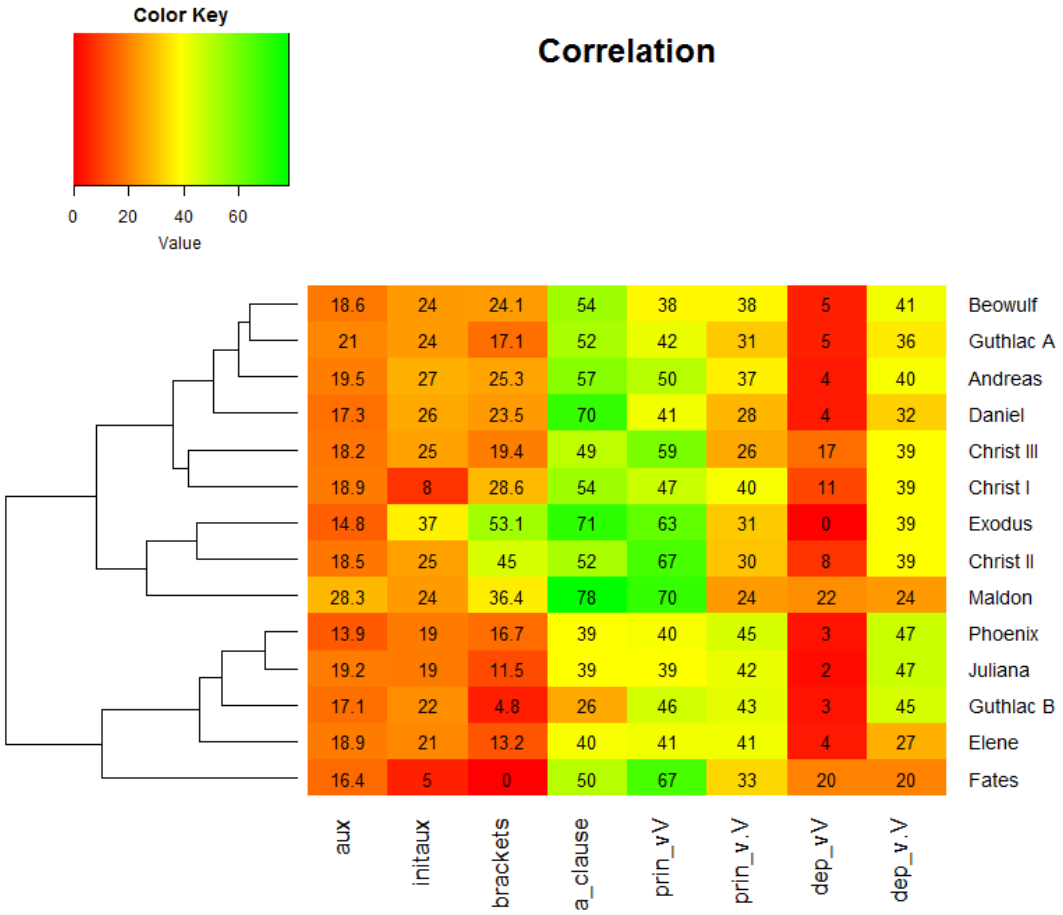

Figure 25. Heat map for 14 poems: Cynewulfian analysis

How might it be possible to improve upon the methodology used in this thesis and what avenues could future researchers of studying style pursue? One possibility is to incorporate and use additional attributes similar to the ones shown in table 12 (e.g., topics, chronology, number of lines) which might help sharpen the output and draw less fuzzy boundaries between clusters of poems and thereby between styles. It is also possible, and probably worth the effort, to apply this clustering approach to much more granular data. Instead of using aggregate auxiliary verb data, using specific verbs and specific kinds of verbs (such as verbs of being) might yield more distinct categories of poems. While purely unsupervised machine learning can be useful to support other methods of determining style, it seems it is too much to expect the output of such programs to clearly delineate categories or styles of poetry by itself.

Where this machine learning approach shows promise is in its ability to provide researchers a tool for describing clusters of Old English poems. The case in point is the description that was developed using the heat map in figure 25 for Cluster 3 (*Elene, Juliana, Phoenix, Guthlac B,* and loosely *Fates*). The raw data behind the descriptions can potentially be used to train software to identify other poems which could have been composed by Cynewulf or by other poets working in the Cynewulfian "style." Such an approach would expand upon this study by integrating a *supervised machine learning* component into the analysis. If used in this fashion, these descriptions and associated data can be thought of as a unique marker for this style—essentially another form of a Cynewulfian signature. Better yet, since this marker was not placed explicitly by the author but is a quality immanent in the poems, it might be more appropriate to refer to it as a Cynewulfian "fingerprint."

Appendix A

R Source Code: Clustering

```
#---------------------------------------------------------
# R program used to cluster data and visualize results
# Code based on Barton Poulson's Lynda.com course
#    entitled "R Statistics Essential Training."
#---------------------------------------------------------
rm(list = ls())  # Clean up

#---------------------------------------------------------
# Load data
#---------------------------------------------------------
oe <- read.csv("~/Desktop/R/input_data.csv", header = TRUE)
rownames(oe) <- oe[,1]  # Use poem names for row names
oe[,1] <- NULL          # Remove poem names as variable
oe                      # display the loaded data

#---------------------------------------------------------
# Hierarchical clustering
#---------------------------------------------------------
d <- dist(oe)  # Generate the distance matrix
d              # Display the distance matrix
c <- hclust(d)  # Get clusters
plot(c)         # Plot the Dendrogram

# Put observations in groups
# Need to specify either k = groups or h = height
# Do several levels of groups at once
# "gm" = "groups/multiple"
gm <- cutree(c, k = 2:10) # generate 2-10 groups
gm # Display the groups

# Draw boxes around clusters for various values of k, only k==4 shown
rect.hclust(c, k = 4, border = "gray")



#---------------------------------------------------------
```

```
# k-means clustering, compute for k value of 2-10, only k==4 shown
#-------------------------------------------------------
set.seed(1)
km <- kmeans(oe, 4, iter.max=10, nstart=25)
km
km$centers
km$tot.withinss
table(km$cluster)

# Graph based on k-means
require(cluster)
clusplot(oe,  # data frame
         km$cluster,  # cluster data
         color = TRUE,  # color
         #shade = TRUE,  # Lines in clusters
         lines = 3,  # Lines connecting centroids
         labels = 2
)  # Labels clusters and cases

#-------------------------------------------------------
# Principal Component Analysis
#-------------------------------------------------------
# Principal components model using default method
pc <- prcomp(oe,
             center = TRUE,   # Centers means to 0 (optional)
             scale = TRUE)    # Sets unit variance (helpful)

# Get summary stats and generate the Scree plot
summary(pc)
plot(pc)

# Get standard deviations and how variables load on PCs
pc

# See how cases load on PCs and generate the Biplot
predict(pc)
biplot(pc)
# ------------------------ End Program ------------------------------
```

Appendix B

R Source Code: Heat Maps

```
#------------------------------------------------------------
# R program used to generate heat maps
# Code based on Sebastian Raschka's tutorial
# http://sebastianraschka.com/Articles/heatmaps_in_r.html#c-
customizing-and-plotting-the-heat-map
#------------------------------------------------------------
# Install and load any packages necessary
install.packages("pheatmap")
library(pheatmap)
if (!require("gplots")) {
  install.packages("gplots", dependencies = TRUE)
  library(gplots)
}
if (!require("RColorBrewer")) {
  install.packages("RColorBrewer", dependencies = TRUE)
  library(RColorBrewer)
}

# Load the data
data <- read.csv("~/Desktop/R/input_data.csv", header = TRUE)
rnames <- data[,1]                            # Use poem names for row
names
mat_data <- data.matrix(data[,2:ncol(data)]) # Transform column 2-5
into a matrix
rownames(mat_data) <- rnames                  # assign row names

# Plot the heat map
my_palette <- colorRampPalette(c("red", "yellow", "green"))(n = 299)
col_breaks = c(seq(-1,0,length=100),   # for red
               seq(0,0.8,length=100),  # for yellow
               seq(0.81,1,length=100)) # for green

heatmap.2(mat_data,
          cellnote = mat_data,  # same data set for cell labels
          main = "Correlation", # heat map title
          notecol="black",      # change font color of cell labels
```

```
          density.info="none",  # turns off density plot in legend
          trace="none",         # turns off trace lines inside heat map
          margins =c(12,9),     # widens margins around plot
          col=my_palette,       # use on color palette defined earlier
          dendrogram="row",     # only draw a row dendrogram
          Colv="NA")            # turn off column clustering

# ------------------------ End Program ------------------------------
```

Bibliography

Works Cited

Albalate, Amparo, and Wolfgang Minker. Semi-Supervised and Unsupervised Machine Learning: Novel Strategies. Hoboken, NJ, USA: 2011 Print.

Anlezark, Daniel. *Old Testament Narratives*. Cambridge, Mass.: Harvard University Press, 2011. Print.

Baker, Peter S. *Introduction to Old English (3)*. Hoboken: 2011. Print.

Bjork, Robert E. *Cynewulf*. New York: 1996. Print.

Bliss, Alan. "Auxiliary and Verbal in Beowulf." *Anglo-Saxon England; Anglo-Saxon England* 9 (1980): 157-82. Print.

Bragg, Lois. *The Lyric Speakers of Old English Poetry*. Rutherford, NJ: 1991. Print.

Bredehoft, Thomas A. "The Date of Composition of Beowulf and the Evidence of Metrical Evolution." The Dating of Beowulf: A Reassessment (2014).

Breeze, A. "Old English Laerig 'Shield Rim' in "Exodus" and "Maldon": Welsh "Iloring" in "Culhwch and Olwen"." Zeitschrift für Celtische Philologie 51 (1999): 170. Print.

Calder, Daniel Gillmore, University of California, Los Angeles Center for Medieval and Renaissance Studies, and University of California, Los Angeles Department of English. *Old English Poetry: Essays on Style*. Berkeley: 1979. Print.

Clark, George. "Scandals in Toronto: Kaluza's law and transliteration errors." The dating of Beowulf: A reassessment (2014): 219-234.

Cynewulf. *The Old English Poems of Cynewulf*. Ed. Robert E. Bjork. 2013.

"Data Mining Algorithms In R/Clustering/K-Means." *Wikibooks, The Free Textbook Project*. 23 Feb 2016, 15:48 UTC. 28 May 2017, 00:26

Doane, A. N. (A. *Genesis A: A New Edition*. Madison: University of Wisconsin Press, 1978. Print.

Donoghue, Daniel. *Style in Old English Poetry: The Test of the Auxiliary*. New Haven: Yale University Press, 1987. Print.

Foley, John Miles. "How Genres Leak in Traditional Verse." *Unlocking the wordhord: Anglo-Saxon studies in memory of Edward B. Irving, Jr* (2003): 76-108. Print.

Fulk, R. D. *A History of Old English Meter*. Philadelphia: University of Pennsylvania Press, 1992. Print.

Hothorn, Torsten. *A Handbook of Statistical Analyses using R / Torsten Hothorn, Brian S. Everitt*. Ed. Brian Everitt. Third edition. ed. Boca Raton, FL: 2014. Print.

Kabacoff, Robert. *R in Action, Second Edition: Data Analysis and Graphics with R*. 2nd ed., 2015. Print.

"K-means clustering." *Wikipedia*. Wikimedia Foundation, 25 May 2017. Web. 28 May 2017.

Mardia, K. V. *Multivariate Analysis*. Eds. John T. Kent, et al. London; New York: 1979. Print.

Neidorf, Leonard. *The Dating of Beowulf: A Reassessment*. Cambridge: 2014. Print.

Niles, John D. *Old English Enigmatic Poems and the Play of the Texts*. Turnhout Belgium]: 2006. Print.

O'Grady, William D. (William Delaney). *Contemporary Linguistics: An Introduction*. 4th ed. ed. Boston: 2001. Print.

Ohmann, Richard. "Generative Grammars and the Concept of Literary Style." *WORD* 20.3 (1964): 423-39. Print.

Orchard, Andy, Catherine E. Karkov, and George Hardin Brown. "Both Style and Substance: The Case for Cynewulf." *Anglo-Saxon Styles* (2003): 271-305. Print.

Poulson, Barton. "R Statistics Essential Training." Lynda.com - from LinkedIn. Lynda.com, 26 Sept. 2013. Web. 29 May 2017.

"A short tutorial for decent heat maps in R." Sebastian Raschka's Website. N.p., 08 Dec. 2013. Web. 01 June 2017.

Shippey, T. A. *Old English Verse,* London: 1972. Print.

Stodnick, Jacqueline. "Cynewulf as Author: Medieval Reality Or Modern Myth?" *Bulletin of the John Rylands University Library of Manchester* 79.3 (1997): 25. Print.

Terasawa, Jun. *Old English Metre: An Introduction*. Toronto: University of Toronto Press, 2011. Print.

Wikipedia contributors. "K-means clustering." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 25 May. 2017. Web. 28 May. 2017.

Works Consulted

Battles, Paul. "Toward a Theory of Old English Poetic Genres: Epic, Elegy, Wisdom Poetry, and the" Traditional Opening"." *Studies in Philology* 111.1 (2014): 1-33. Print.

Bessinger, Jess B. *A Concordance to the Anglo-Saxon Poetic Records*. Ed. Philip H. Smith. Ithaca, N.Y.: Cornell University Press, 1978. Print.

Birkett, Tom. "Runes and Revelatio: Cynewulf's Signatures Reconsidered." *The Review of English Studies* 65.272 (2014): 771. Print.

Brooks, Kenneth R., Andrew, and Cynewulf. *Andreas, and the Fates of the Apostles*. Oxford: Clarendon Press, 1961. Print.

Everitt, Brian S., et al. *Hierarchical Clustering*. Chichester, UK: 2011. Print.

Greenfield, Stanley B. *The Interpretation of Old English Poems,* London:, 1972. Print.

Jain, A. K., M. N. Murty, and P. J. Flynn. "Data Clustering: A Review." *ACM Computing Surveys* 31.3 (1999): 264-323. Print.

Jolliffe, I. T. *Principal Component Analysis*. 2nd ed. New York: 2002. Print.

Kears, Carl. "Old English Mægen: A Note on the Relationship between Exodus and Daniel in MS Junius 11." *English Studies* 95.8 (2014): 1-24. Print.

Larose, Daniel T., and Chantal D. Larose. *Hierarchical and k - Means Clustering*. Hoboken, NJ, USA: 2014. Print.

Lucas, Peter J. *Exodus*. Rev. ed. Exeter: University of Exeter Press, 1994. Print.

Orchard, Andy. "The Word made Flesh: Christianity and Oral Culture in Anglo-Saxon Verse." *Oral Tradition* 24.2 (2009) Print.

Peters, Leonard J. "The Relationship of the Old English Andreas to Beowulf." *PMLA* 66.5 (1951): 844-63. Print.

Wu, Junjie. *Advances in K-Means Clustering: A Data Mining Thinking*. Berlin, Heidelberg: 2012. Print.

Zacher, Samantha, and Andy Orchard. *New Readings in the Vercelli Book*. Toronto; Buffalo, NY: 2009. Print.