EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 25–34
CILC2016. 8th International Conference on Corpus Linguistics

How Many People in the Making of Sloane 770? a Corpus-Based Approach

Jessica Carmona-Cejudo
University of Málaga, Málaga, Spain
jessicacc@uma.es

Abstract

During the transcription of a late Middle English manuscript on Medicine (London, British Library, MS Sloane 770), a series of orthographic variations appeared, several of which seemed to be arranged following a predictable pattern. Should this prove correct, it may be a clue to posit the existence of two or more scribes who were involved in the copying of the codex, or else of the dialect of the MS being an example of Mischsprache that combines the dialects of the exemplar MS and that of the scribe. To ascertain whether the MS was written by more than one copyist or whether it is an example of the coexistence of two different dialects, a morphologically lemmatized corpus was built. This paper will present the results obtained after studying that corpus in order to verify either the original hypothesis is linguistically and scientifically based or not.

1 Introduction

The edition of scientific prose written in Middle English was not in vogue up to the second half of the twentieth century. Other fields such as literature, religion (Taavitsainen and Pahta 3) or history concentrated all the attention, and it is not strange: information from those codices could be hurled as weapons at the service of political or social agendas, whereas scientific texts could not possibly serve for those aims (Moreno Olalla 388). Some of those utilitarian treatises received a little attention in the last few decades, mainly for the construction of corpora (e.g. *Corpus of Middle English Medical Texts* (*MEMT*)), but there are plenty of scientific MSS yet to be studied.

One of those scientific MSS that went unnoticed is London, British Library, Sloane 770, a medical text. It is a 15th-century codex written "part in paper, part in vellum" (Ayscough 561) in a hybrid script showing Anglicana and Secretary traces, as proved by letter <w> (see *Fig.* 1). This hybrid script may also be described as *cursiva* since it is not very polished in its execution. See for example the reversed <e> or the union between minims in opposition to *formata* scripts, in which for the writing of each of the strokes the pen was lifted from the page.

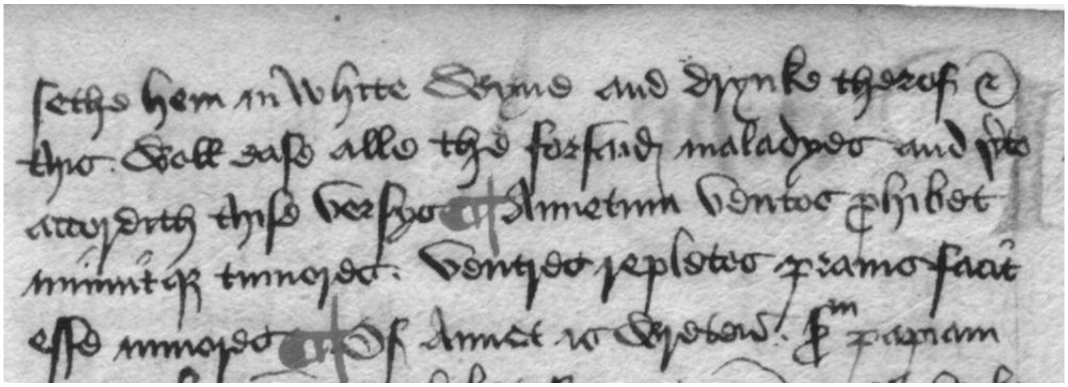

Figure 1: Excerpt of Sloane 770, showing traces of cursiva anglicana and secretary scripts.

The MS is composed of six different parts, to wit: 1. notes on some diseases in a hand dating from the 16th century (1r); 2. "rules for calculating the moon's rising" (Scott 29), written in Latin (1v); 3. a *tabula de contentis* also in Latin (2r-5v); 4. the main medical treatise (6v-43v); 5. a secondary medical treatise, yet incomplete (44r-45r, containing some entries of Platearius' *Circa Instans*); 6. part of Platearius' *Circa Instans* (45v-48v), also incomplete.

According to an inscription in the MS by a later hand, the main treatise was supposedly written by Gilbert Kymer (Ayscough 561), physician to Humphrey, Duke of Gloucester and youngest brother of King Henry V, although this may be a wrong attribution (Griffiths and Pearsall 401) that needs to be confirmed. Regarding the dialect of the text, it may have been written in a bordering area between the East and the West Midlands.

During the transcription of the text, a series of specific alternates in key words were detected that seemed to be presented following a pattern, or the simple appearance of certain utterances. This may indicate either that the text was copied by several scribes, which surely was, or maybe that the MSS is an example of *Mischsprache.* This means the dialect of the archetype (supposedly written by Kymer) and that of the scribe(s) of the MS could be detectable. For this purpose, the application of corpus and computer linguistics tools could be useful.

2 Analysis

For this study, a corpus was built. It was morphologically tagged and lemmatized. Each word in the MS was assigned a cell. In addition to this, several pieces of morphological and non morphological information were included, such as:

1. a code obtained from the electronic *Middle English Dictionary* (*eMED*)*,* to crossreference the word in the MS with that of the standard lexical work in Middle English, which is particularly useful if the corpus even becomes online as these codes can easily be turned into links, so that the user may click on any word of the text and the definition in MED could pop out;
2. a normalized lemma, also supplied by the *eMED*. One must bear in mind that given the heterogeneous nature of spelling in ME, it is interesting to provide an standardisation of any given word. This is also useful to calculate lexical richness;
3. meaning of each word in contemporary English;
4. word-class (verb, noun, pronoun, adjective, preposition, conjunction...);
5. verbal tense or gradation (comparative, superlative in adjectives and adverbs);

6. number (singular, plural. With verbs, nouns, adjectives and pronouns);
7. person (1st, 2nd, 3rd. Verbs and pronouns);
8. case (nominative, genitive or object. Pronouns);
9. gender (pronouns);
10. subclass (personal, demonstrative... For pronouns, determiners);
11. notes about interesting spellings or other useful information for the analysis of the MS. Rather than part of the tagging processs, this was designed as a sort of notepad so that any ideas on a word can be quickly annotated and at the tagger's fingertips.

The advantages of a corpus like the one presented here are that it contains several pieces of information for each word. It makes it easy, also, to compare the different spellings by reorganising the information by lemma, meaning, tense, number... One can see, for example, the different verbal endings of all the 3rd person plural forms of verbs or the complete pronominal system in a very fast way, so the dialect can be discerned all the more easily. Among the drawbacks of this system, the main one is that the corpus lacks syntactical information. A corpus with this type of data will have to wait for future studies, as it needs an extended period of time to be constructed.

To learn how many people was involved in the copying of Sloane 770, the text was divided into different sections once the corpus was created. In this type of studies it is usual to normalise the text into sections of 1,000, 2,000 or 2,500 words (Boyd et al. 21), for example. Nevertheless, a division such as this one may not be as accurate as it should, establishing blurry boundaries. In Sloane 770, the change of scribe seems to take place, usually, at the beginning of a new letter judging from the changes in the script. Thus, dividing the text into pieces of a so inflexible nature may merge the writings of two scriveners together, as each letter does not contain the same number of words. After observing the MS, the decision was reached that the best option would be, to make the divisions in the following way: first, series of sections that do not belong to the main text were discarded, as they were clearly written by other scribes. Those parts were the 16th-century notes, the astrological records, the *tabula de contentis* and the two secondary, yet incomplete, treatises (44r-48v). Then, the main treatise was divided into seventeen different groups. The text is arranged as a kind of dictionary. In it, a series of plants appear in alphabetical order, and there (usually) is a section per letter. New groups were created whenever noticeable changes in the script were observed. Thus, the sections were created as follows: eleven groups, each for a letter (*a, b, c, d, e, f, g, h, j, k, l*). G. 12 corresponds with the most part of letter *m*; g. 13 contains part of letter *m* and letters *n, o* ; g. 14 contains part of letter *p*, group 15 contains the second part of letter *p*, letter *r* and part of letter *s*; g. 16 contains the second part of letter *s* and group 17 contains letters *t* and *w.* There are no instances of plant-names in the treatise beginning with *q*, *x* or *z*. As for *i* and *y*, they naturally included within letter *j*, just like *u* and *v* are part of *w*.

As noted before the MS was studied through the observation of a series of key words and their respective alternates in each of the groups. Those alternates or either the use of certain words will be noted as *marks,* and each of them may give information not only on the idiosyncrasy of the scribes, but also on the dialect of each of the parts. The marks taken into consideration were twenty-seven, which can be associated and set together:

- 1. and 2., the apparition (or not) or the first person singular and plural pronouns. About their usage (or the lack thereof), it is true that in scientific prose (not only in the ME period, but also today) first person pronouns are sparse. Nevertheless, its use here seems to be not only a question of merest chance. In authorship and attribution studies or methods such as Burrows' *Delta* one of the main points to take into consideration are "articles and personal pronouns" (Burrows 28). In the case of Sloane 770, they seem to be consigned to very delimited areas of the MS. Thus, the passages in which *I* or *we* do appear, may indicate a change of scrivener.
- 3. and 4., denoting the 3rd person plural pronoun in object case as *them,* typical of the northern areas of the island (Mossé 65) or as *hem*, typical of southern areas (65);

- 5., 6. and 7., verb *bẹ̄n* (PDE "to be"), 3rd person plural, present indicative as *ar*(*e*), *ben* or *be.* In any case, the use of those three forms may indicate that the text was possibly composed in the East Midlands (Mossé 105);
- 8. and 9. 3rd person plural present indicative verbal forms in -*en* or -*ø*. Those verbal endings are also typical of the Midlands (Mossé 93);
- 10. and 11., ME *thurgh* as "thurgh" or"through" (metathesis and introduction of a parasitic vowel (Jordan 175));
- 12. and 13., words with or without <-i-> as a diacritic to indicate the lengthening of /ʃ/ in three different words: *flē̆sh*, *washen* and *frē̆sh*;
- 14., 15., 16. and 17.,words ending with unstressed -*ur*; -*er*, -*ir* or -*re* in words such as *wāter*, *pŏudre*, *after*, *finger*, *cū̆cŏmer*, *buter(e*, *canker* or *fēver.* It is worth noting that, Sloane 770 being a medieval MS, there were plenty of abbreviations in the text. Many words ended with an abbreviation rather than with the full letters *-ur, -er, -ir,* or *-re*. The best option here was not to include the shortened forms in the study as using a particular expansion would have skewed data: we cannot totally assume that the scribe who was copying the text meant one ending or the other when expanded forms such as "water" and "watur" were combined in the text.
- 18. and 19.,"destroy" and "d[i,y]stroy for ME *dē̆stroien* because of a change in the pre-tonic vowel (Jordan 133);
- 20. and 21., ME *jūs* written with or without a diacritic <-[i,y]->, which served to indicate that the previous vowel was long (Mossé 12);
- 22. and 23., ME *hir*(*e* written as <hir> or <her>;[1]
- 24. and 25., OE ǣ$_1$ as <a> or <e>;[2]
- 25. and 26., OE ǣ$_2$ as <a> or <e>;

The corpus was then rearranged so as to have a clear vision of the different orthographies set together. Then, a matrix was created. The groups were included as axis *x*, while axis *y* recorded the different marks. The matrix was then fed with 3 different values: <1>, in case the mark could be seen in the group; <0> if the mark does not appear and <?> if it is not possible to know whether a certain mark may have appeared or not due to the brevity of the group. The use of <?> is, obviously, limited. The different marks are related. Let us see marks 5, 6 and 7, which conform one of those "sets": they do not appear in group 10, as this is a short one. Thus, we cannot infer if the scribe of that part usually expresses the 3rd person plural forms of verb *bẹ̄n* **as** <be>, <ben> or <ar(e)>, so it was denoted with <?> to try to assign <1> or <0> by means of statistical analysis later on. The same happens, for example, with marks 10 and 11 in several groups (e.g. 5, 6, 7, 9).

The following step was to *clean* the data. Those marks that were unknown in more than three groups were removed. This was done because a mark that is lacking so many times cannot be a good indicator to discern the number of scribes. The marks whose value was always <1> have also been removed. See mark number 9: the 3rd person plural present indicative verbal forms without suffix. Its value is always <1> except in groups 11 and 17, where it was <?>. Statistically, the value in those groups would have been <1> as well. As this mark is constant throughout the text, it cannot serve to differentiate between scribes, it does not create any significant difference between any part. In addition, some of the *marks* were correlated mathematically with each other. Mark 5 (3rd person plural form, present indicative) of ME *bẹ̄n* as <ben>) and mark 15 (unstressed ending -*er*) shared a result of correlation of 0.8704. This means that marks 5 and 15 occur together the 87.04% of the time. Their

[1] *c.f.* Mossé §§ 135, 150, 151.

[2] To learn more about the development of OE ǣ$_1$ and ǣ$_2$, see Jordan's *Handbook of Middle English Grammar: Phonology*. §47-50.

rate of correlation served to try to figure out the remaining seven unknown values (<?>), creating 128 different hypotheses on the number of scribes.

For each hypothesis, a hierarchical agglomerative cluster3 was created. This means that, for each conjecture, the groups were ordered in a hierarchical way. The groups are united by nodes, in tree-like fashion. The lower a node appears uniting groups, the more related those groups are. When the representations were created, a great amount of them were repetitions. Thus, it was time to see how many different "trees" (dendrograms) were actually different. There were only eight (see *Figs.* 2, 3, 4, 5, 6, 7, 8, 9).

At this point, it was necessary to explore visually the results, as the cluster analysis could not help any more without the help of an expert on the area. The problem is *where* to look at the divisions. At a first glance, if one concentrates his or her attention on the lower nodes of the trees, it seems that the text was possibly written by five to eight scribes. There are groups that are clearly related and, therefore, supposedly copied by the same scribe. They are, for example, groups 7 and 8, groups 10 and 17, groups 1, 3, 5 and 6. It remains unsolved whether groups 11 and 13 are related to group 12. But five to eight scribes means a very high rate for such a small text: this probably indicates that we must take into consideration the upper nodes of the trees rather than the lower ones. This would mean that the main treatise was written only by two or three people. To this number, it is necessary to add 4-5 more scribes who copied the other parts of the MS and whose hands make clear that they were not part of the copying of the main piece.

3 Conclusions

Although this type of study throws some light on authorships issue of this MS, it is necessary to complement this study with further researches on the application of hierarchical clustering to corpus linguistics in the analysis of MSS. This is already being done with the help of a data mining expert. It seems necessary to implement the application of corpus linguistics and cluster analysis by a thorough study on the palaeographical aspects of the MS, as it may shed light on the obscure question of how many people were involved in the making of Sloane 770. The mystery of the MS being an example of *Mischsprache* is hopefully to be solved as well.

4 Appendix: figures

[3] For further information on the issue see Rokach, Lior, and Oded Maimon, "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005. 321-352

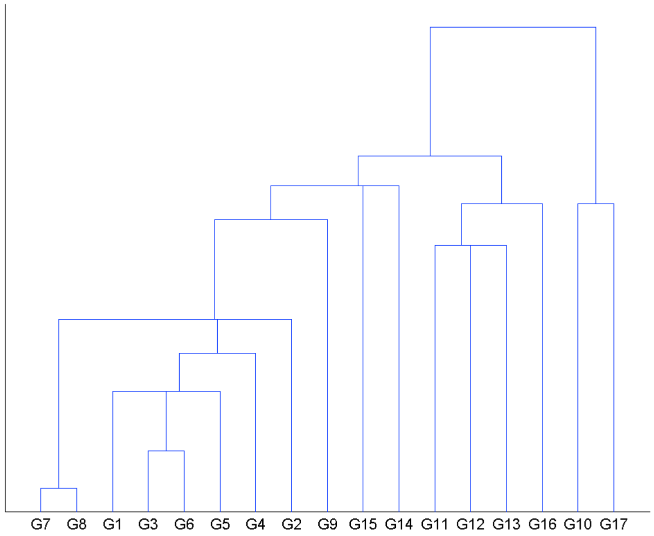

Figure 2: Dendrogram 1

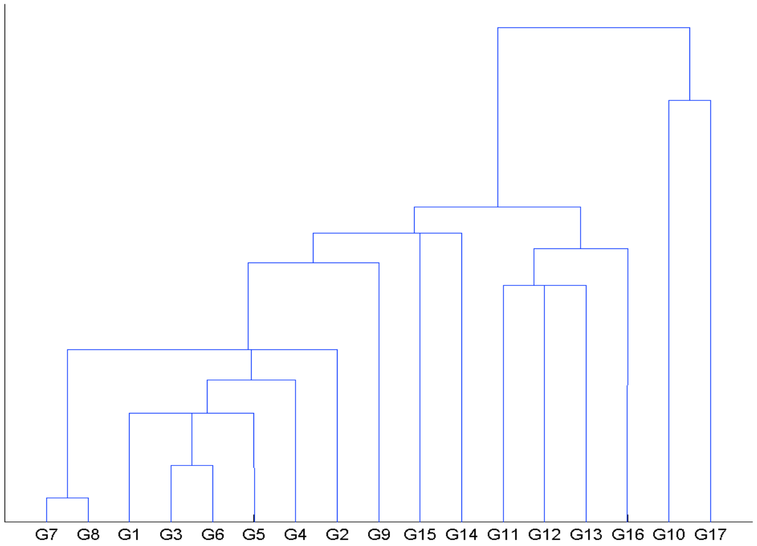

Figure 3: Dendrogram 2

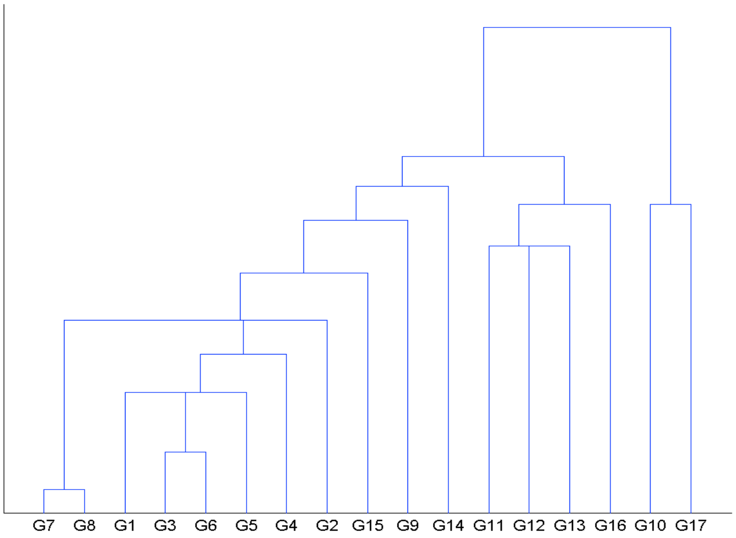

Figure 4: Dendrogram 3

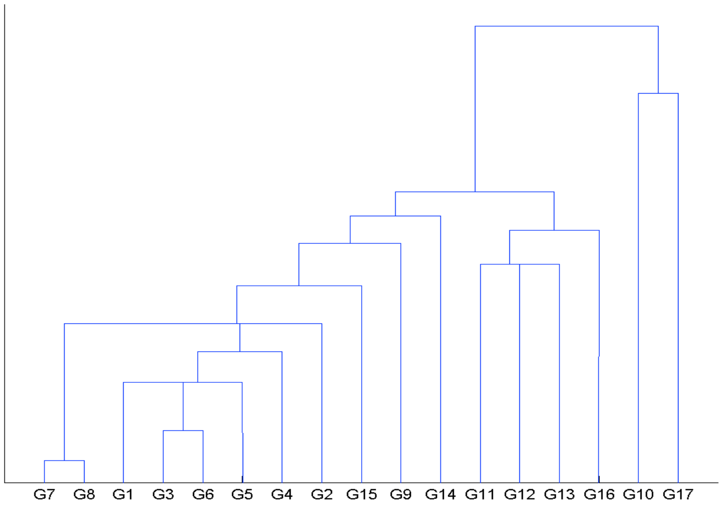

Figure 5: Dendrogram 4

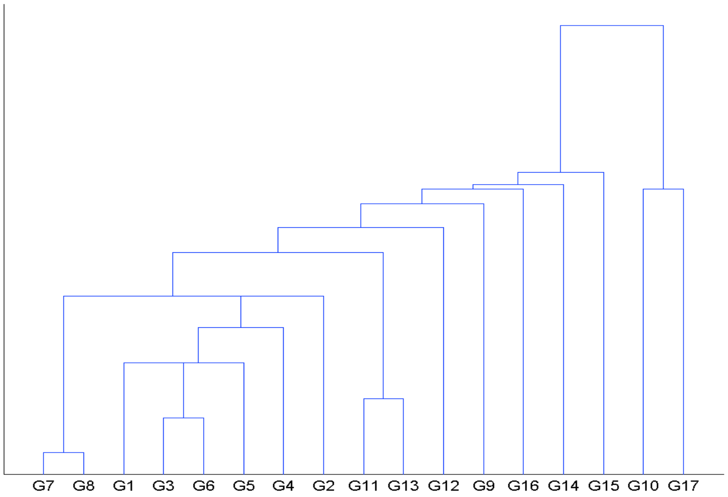

Figure 6: Dendrogram 5

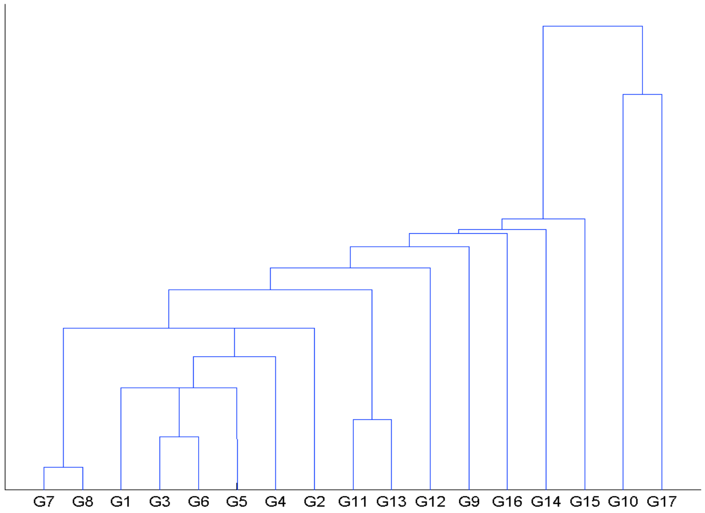

Figure 7: Dendrogram 6

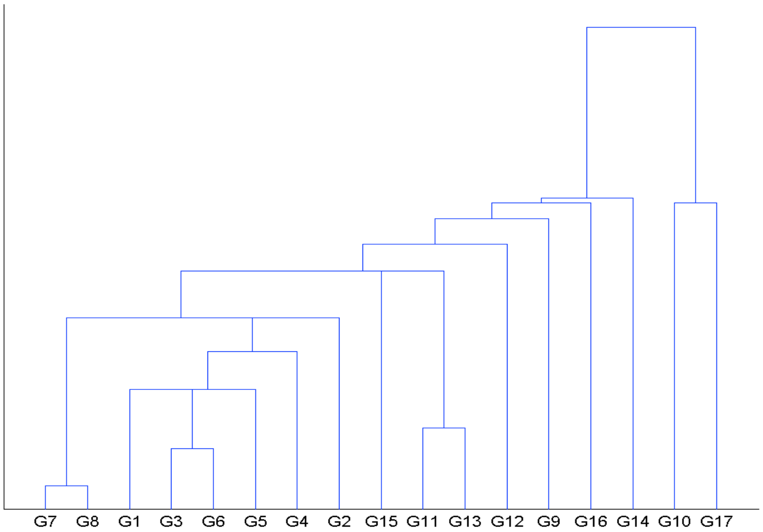

Figure 8: Dendrogram 7

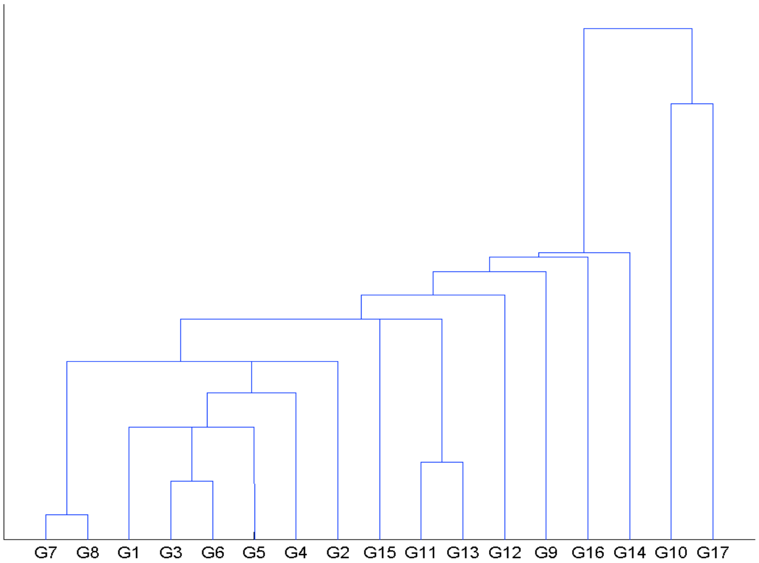

Figure 9: Dendrogram 8

References

Ayscough, Samuel. *A Catalogue of the Manuscripts Preserved in the British Museum Hitherto Undescribed: Consisting of Five Thousand Volumes; Including the Collections of Sir Hans Sloane, Bart. the Rev. Thomas Birch, D. D. and about Five Hundred Volumes Bequeathed, Prefented, or Purchafed at Various Times.* II. London : John Rivington, 1782. Print.

Boyd, Phoebe et al. "Lexomic Analysis of Anglo-Saxon Prose: Establishing Controls with the Old English Penitential and the Old English Translation of Orosius." *Journal of the Spanish Society for Mediaeval English Language and Literature (SELIM)* 19 (2012): 7–58. Print.

Burrows, John. "Questions of Authorship: Attribution and Beyond." *Computers and the Humanities* 37 (2003): 5–32. Print.

Griffiths, Jeremy, and Derek Pearsall. *Book Production and Publishing in Britain 1375-1475*. Cambridge University Press, 1989. Print.

Jordan, Richard. *Handbook of Middle English Grammar: Phonology*. Trans. Joseph Crook. Paris: Mouton, 1974. Print.

Mossé, Fernand. *Handbook of Middle English*. Trans. James A. Walker. Fifth. Baltimore, USA: Johns Hopkins Press, 1968. Print.

Moreno Olalla, David. "A Plea for Middle English Botanical Synonyma." *Probable Truth*. Ed. Gillespie, Vincet and Hudson, Anne. Vol. 5. N.p., 2013. 387–404. *MetaPress*. Web. 16 Mar. 2014.

Scott, Edward J. L. *Index to the Sloane Manuscripts in the British Museum*. London: Williams Clowks and sons, limited, 1902. Print.

Taavitsainen, Irma, and Päivi Pahta, eds. *Medical and Scientific Writing in Late Medieval English*. First edition. Cambridge: Cambridge University Press, 2009. Print.