The Geometric Structure of Scribal Variation among Manuscripts of Langland's *Piers Plowman*

Roger Bilisoly
Department of Mathematical Sciences, Central Connecticut State University
1615 Stanley St., P. O. Box 4010, New Britain, CT 06050 USA
BilisolyR@CCSU.edu

Abstract

In preparation of his edition of the 14th century alliterative poem *Piers Plowman*, the 19th century philologist, Walter Skeat, was able to find forty-five manuscripts. These were used in two different ways. First, he studied these with respect to their dialects, which led to his identification of three versions of the poem, denoted as texts A, B, and C. Second, he published about ten lines from each extract to display the variability of spelling of Middle English. Today grouping texts can be thought of as a clustering problem, which requires measuring distances between pairs of strings. This is a well studied problem from information retrieval, and many methods such as the Levenshtein edit distance exist. This paper visualizes the distance matrix obtained from applying edit distance to the forty-five extracts that Skeat published by using multidimensional scaling, which reveals the geometric structure of the variability among his manuscripts.

Introduction

Extensive work has been done connecting mathematics and the visual arts at the Bridges conferences and elsewhere, but investigating the mathematical structures underlying the literary arts is less common. Nonetheless, scholars have thought about this for at least 150 years. For example, in 1870 the mathematician J. J. Sylvester published *The Laws of Verse* [6] that interspersed his translations of poetry into English with his principles of versification. For instance, he gives his translation of "An Ode of Horace," followed by a discussion of its particular structure along with generalizations such as his law of phonetic syzygy (pages 25–49).

This paper investigates the geometric structure underlying a study of Middle English spelling variability performed by the philologist Walter Skeat in the 1860s. At that time he edited an edition of William Langland's 14th century alliterative poem, *Piers Plowman* [4, 5], where he examined forty-five manuscripts and concluded that these reflected three underlying versions, texts A, B, and C. This paper shows that combining Levenshtein edit distance with dimensional reduction techniques gives excellent agreement with Skeat's classification. Although the geometric underpinnings of certain visual artists such as M. C. Escher are obvious to all, it is intriguing that distances can be defined between texts, the geometry of which can be explored using a computer.

Skeat's Text Clustering Task

Skeat is famous for his work on philology, etymology, and place-names. He helped Frederick Furnivall found the Early English Text Society (EETS) in 1864, the goal of which has been to edit and publish Old and Middle English handwritten manuscripts. He edited many volumes for EETS, including the 14th century alliterative poem, *Piers Plowman* [3, 4, 5], which are the sources for the analysis below.

In preparation for his edition of *Piers Plowman*, Skeat requested that anyone with knowledge of a manuscript contact him so that he could make a complete list [3]. Eventually forty-five manuscripts came

to his attention, and from analyzing these he concluded that there were three versions, which he called A, B, and C. The A text was already recognized in his day because it is much shorter (only about 2500 lines of poetry) than the rest (about 7300 lines), but distinguishing the B and C texts was his discovery.

To show the variability contained in these texts, Skeat published roughly ten lines from each of the forty-five manuscripts in [3], which will be the input data for the analysis below. However, he had access to many of the manuscripts and was an expert in Middle English vocabulary and dialects, so his conclusions are taken as the gold standard here, and the success of modern algorithms used below is judged by how consistent they are with his results.

Finally, Skeat's ultimate goal was to construct an authoritative version of the A, B, and C texts, which is much more challenging than clustering. He wanted to create a synthesis that best reflected Langland's own dialect, believed today to be from the west Midlands, but this is difficult because scribes did not copy without error. In fact, sometimes they introduced words from their own dialect. An extreme case of this is manuscript XII (using Skeat's numbering here and below) which is entirely written in the Northumbrian dialect as noted on page 8 of [3].

Distances between Texts

Clustering algorithms require distances between pairs of objects. We need distances between strings, and many metrics have been developed to do this. Here we use Levenshtein's edit distance (see Chapter 6 of [2]), which is defined to be the minimum number of edits required to transform one string into another, where inserting, deleting, and substituting one character are the only allowed changes. For instance, the distance between *pepil* and *puple* (two Middle English forms of the Modern English word, *people*) is three because the following three edits cannot be reduced to two: *pepil* $\rightarrow$ *pupil* $\rightarrow$ *pupl* $\rightarrow$ *puple*. That is, change the *e* to a *u*, delete *i*, and insert *e*. This distance is a metric in the mathematical sense: for example, it satisfies the triangle inequality.

After the Norman victory at the Battle of Hastings in 1066, French and Latin became the language for business, the aristocracy, the courts, and religious writings. At the time Langland wrote, the late 14th century, English was being revived for official uses, but there were many dialects, and it is believed that spelling reflected the way each writer pronounced the word. Consequently, while today orthography is established by dictionaries, style manuals, editors, and so forth, in Langland's time using several spellings of a word in a single document was common. For example, the sixth and seventh lines from Skeat's extract from manuscript II (page 7 of [3]) have the forms *peple* and *pepul*.

We make use of this spelling variability below to cluster the Skeat's forty-five extracts. Because he had access to much more than the ten or so lines per manuscript published in [3], and because he was an expert in Middle English dialects, we assume that his clustering into the A, B, and C texts is correct. Consequently, the goal below is to see how consistent the results using edit distances along with dimensional reduction of the resulting distance matrix are to his.

Distances between Skeat's Extracts

Recall that the A texts are much shorter than the other two, so the challenge Skeat faced was distinguishing the B and C texts. However, the lines Skeat used, which were suggested to him by Furnivall, are quite different in these latter two versions as seen below. The first is manuscript XIII, which Skeat thought was the most accurate of the B texts, and the second is manuscript XXIX, the best of the C texts.

Meires and maceres · that menes ben bitwene
þe kynge and þe comune · to kepe the lawes
To punyschen on pillories · and pynynge stoles
Brewesteres and bakesteres · bocheres and cokes
For þise aren men on þis molde · þat moste harme worcheth
To the pore peple · þat parcel mele buggen
For they poysoun þe peple · priueliche and oft
Thei rychen þorw regraterye · and rentes hem buggen
With þat þe pore people · shulde put in here wombe
For toke þei on trewly · þei tymbred nouȝt so heiȝe
Ne bouȝte non burgages · be ȝe ful certeyne

The above B version has eleven lines, but the C version below only has nine lines, so the lines cannot all match up. In fact, only the third and fourth lines as well as the last four lines correspond well to each other. Because edit distance easily detects differences in length, a more challenging task is to cluster the forty-five extracts using only these six lines, which is done below.

ȝut mede myldeliche · þe meyre hure bysouhte
Bothe shereues and seriauns · and suche as kepeþ lawes
To punyshen on pillories · and on pynyng stoles
As bakers and brewers · bouchers and cokes
For þees men doþ most harme · to þe mene puple
Richen þorw regratrye · and rentes hem byggen
Whit þat þe poure puple · sholde putten in hure womben
For toke þey on triweliche · they tymbrid nat so heye
Noþer bouhten hem burgages · be ȝe ful certayn

The Calculation and Results

Figure 1 was produced by the following steps. First, the forty-five extracts were reduced to lines 3 and 4 plus the last four lines. Punctuation was removed (including the caesura of each long line), and all text was put into lower case. Second, a 45-by-45 distance matrix, D, was constructed using edit distance. Third, by using multidimensional scaling (introduced in Chapter 5 of [1]), we construct X_3, a 45-by-3 matrix representing 45 points in $\mathbb{R}^3$, the Euclidean distance matrix of which is the best approximation of D by 3-dimensional points. Last, an interactive 3-dimensional plot of X_3 was colored according to Skeat's classification and then rotated by the author in hopes of finding a view that separates his clusters. As seen below, that turned out to be possible.

In conclusion, here are three final points. This analysis uses about a quarter of one percent of text A and one third as much of texts B or C. Moreover, the edit distance is unweighted and makes no use of known properties of Middle English. For example, the letters *u* and *v* can both be used as either a consonant or vowel. Hence *vntrewly* from manuscript XLIII is a form of *untruly*. Consequently, given these minuscule samples and complete lack of a priori knowledge of Middle English, the results seem excellent.

Second, edit distance penalizes changing the order of words. For example, line 4 sometimes starts with "bakers and brewers" but other times with "brewers and bakers." The distance between these two Modern English phrases is 6, but that might be considered too high since the only difference is swapping two words. However, other string distances exist, and one of these could be used above if it seemed more appropriate. However, scribes copy a manuscript letter-by-letter, so the operations of insertion, deletion, and substitution are plausible in this context.

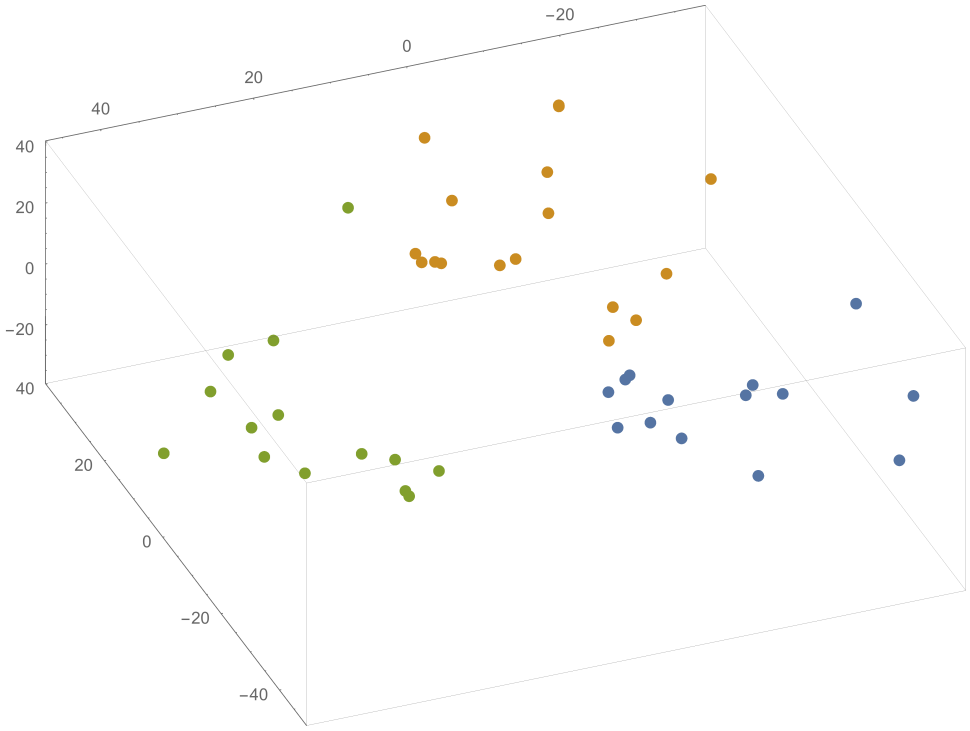

Figure 1 : *Skeat's forty-five manuscripts plotted on the three dimensions found by the multidimensional scaling algorithm. Blue represents the A texts, yellow the B texts, and green the C texts.*

Finally, Skeat died in 1912, and so it is impossible to say what he would have thought of this approach. Although he is known today for his expertise in the history of English, he was well acquainted with mathematics. As an undergraduate at Christ's College, Cambridge, he read theology and mathematics, and in 1864 his first academic job was a lectureship in mathematics, returning to his old college. In this case, the results of multidimensional scaling do detect his classification of the A, B, and C texts.

References

[1] B. S. Everitt and G. Dunn, *Applied Multivariate Data Analysis*, Wiley.

[2] M. A. Russell, *Mining the Social Web* (2011), O'Reilly Media.

[3] W. W. Skeat, Parallel Extracts from Forty-Five Manuscripts of Piers Plowman, Second Edition (1885), Truebner, `https://books.google.com/books?id=W6wxAQAAMAAJ` (as of Feb. 27, 2015).

[4] W. W. Skeat, The Vision of William Concerning Piers the Plowman in Three Parallel Texts, Volume I (1886), Oxford, `http://books.google.com/books?id=N_8qAAAAIAAJ` (as of Feb. 27, 2015).

[5] W. W. Skeat, *The Vision of William Concerning Piers the Plowman in Three Parallel Texts, Volume II* (1886), Oxford, `http://books.google.com/books?id=OP8qAAAAIAAJ` (as of Feb. 27, 2015).

[6] J. J. Sylvester, *The Laws of Verse or Princliples of Versification Exemplified in Metrical Translations* (1870), Longmans, Green, and Co., `https://ia700406.us.archive.org/18/items/lawsofverseorpri00sylviala/lawsofverseorpri00sylviala.pdf`, as of April 16, 2015.