Linguistica Silesiana nr 42, 2021
ISSN 0208-4228
DOI: 10.24425/linsi.2021.137234

JERZY WÓJCIK
The John Paul II Catholic University of Lublin
jerzy.wojcik@kul.pl

MEASURING INTERNAL SPELLING VARIATION OF AN EARLY MODERN ENGLISH TEXT

The history of English spelling is an eventful one, from Old English with an almost one-to-one sound-to-spelling relationship, to Modern English, notorious for its sound-to-spelling unpredictability. In between lies a vast period characterised by immense spelling variability, reflecting the cumulative effect of dialectal variation and lack of uniformity, additionally compounded by the mode of text transmission in the manuscript culture, whose characteristics were adopted in a wholesale fashion into the culture of early print. In effect, early printed books present a rich kaleidoscope of spelling variants, which – not infrequently – co-occur on the same page or even in the same line of a printed text. This paper addresses the issue of this variability with a view to measuring in mathematical terms the degree of internal spelling variation within a text and showing that much of the spelling variation is associated with compositors as agents in the printing process. The analysis of internal spelling variation is based on George Joye's 1534 English translation of the Psalms printed in Antwerp and aims at identifying parts of the text which are similar or different in terms of spellings by applying cosine similarity measurements performed on individual quires of the publication.

Keywords: *spelling variation, early print, cosine distance, compositor*

1. Introduction

The goal of this paper is to measure the degree of internal spelling variation within the text of George Joye's 1534 English translation of the Psalms with a view to identifying parts of the text which are similar or different in terms of spellings. The presence of identifiable spelling similarities or differences between parts of the text can be taken as indicating that different people were responsible for the creation of different parts of the same text since the observed

spelling variation can be seen as resulting from different orthographic idiolects being used in parts of the same text. Following Shute (2017), I shall use cosine similarity measurements performed on different quires of George Joye's Psalter to measure the degree of similarity between spellings used in different parts of the text.

A considerable degree of spelling variation is a typical feature of Middle English and early Modern English texts produced before the end of the seventeenth century, when a regular and generally accepted spelling system for English emerged (Scragg 1974: 68). A typical example of this variation in Joye's 1534 Psalms would be the interchangeability of <i/y> in words like *kinge/kynge* 'king' and presence or absence of final <e> in words like *with/withe* 'with',[1] which can also be realised as *whith*, *wyth* and *wythe* with different variants often appearing within the same line of text. Generally, there is little scholarly attention to this particular feature of early Modern English spelling as it tends to be dismissed as non-distinctive and hence regarded as unimportant (Scragg 1974: 21, Fisher 1996: 50). Consequently, research into English historical spelling has concentrated on identifying spelling differences as markers of dialect provenance (e.g. Samuels 1963, McIntosh *et al.* 1986) or has looked into the rise of standard spellings and the development of particular spelling features in an attempt to trace the regularisation of spellings (Blake 1965, Aronoff 1989, Horobin 2001, Howard-Hill 2006, Rutkowska 2013). In what follows, I shall demonstrate that internal spelling variation, taken for granted in previous accounts of early Modern English spelling, can be used as an interesting source of information about the number of people who were involved in creating the book.

The paper is organised as follows: Section 2 will concentrate on the presentation of the data used in the analysis, Section 3 will discuss cosine similarity - one of the methods used in calculating similarity between texts (Wang and Dong 2020) and will show how to extend the application of cosine similarity to the analysis of spelling data. Cosine similarity scores between all quires of Joye's 1534 Psalms will be provided. Finally, Section 4 will offer some conclusions.

2. The data

George Joye's 1534 English translation of the Psalms was printed in Antwerp by Martin Emperowr. It is one of the four printings of Joye's Psalms which were published within a relatively short period of time between 1530 and

[1] Of course, the reasons for the variation are associated with phonological developments in the history of English having to do with the 10th c. Old English /i/ – /y/ merger in the case of <i/y> variation, and the 14th c. loss of final /ə/ in the case of the presence or absence of the final <e> (Lass 1992: 54, 79).

1541 or 1544.[2] Joye's Psalms were printed in 1530 in Antwerp (this was in fact the first time the Book of Psalms appeared in print in English), in 1534 and 1541/4? in London by two different printers Tomas Godfray and Edward Whitchurch at the time when printing and distributing translations of biblical texts in English was no longer a capital offence (Hotchkiss and Robinson 2008: 13). These three printings were all prepared on the basis of Joye's original 1530 translation from Latin.[3] The Psalms analysed here are another translation prepared by Joye but this time on the basis of a different Latin text of the Psalms by Huldrych Zwingli.[4] As mentioned above, this translation was printed in Antwerp, which was at the time a place of refuge for a growing community of English Protestants and an important centre of book trade which was involved in printing books for the English market (Juhász 2014: 19). At the same time, the cheap labour of English immigrants was readily exploited by publishers, who employed English protestant refugees as translators or proofreaders (Juhász 2014: 20). Joye himself worked, for example, as a proof reader for a 1534 edition of William Tyndale's New Testament (Juhász 2014: 24).

In order to understand possible reasons for internal spelling variation within the text of a book printed in the 16th century we first need to take a closer look at the process of book production. As observed by Hellinga (1999: 80), manuscripts and printed books were produced in codex form, which means that sheets of paper (or vellum) were folded together to form quires and these were combined and bound together to form volumes. The printer had to calculate precisely where in the text each page had to begin because typesetting and printing were executed in the order deviating from reading order (Hellinga 1999: 81). What is more, texts for printing were frequently divided between multiple compositors for financial reasons as splitting the work between a few men reduced the time needed to print the book (Gaskell 1972). Each quire in Joye's Psalms consists of 16 pages, i.e. four sheets of paper folded to make 16 pages. Joye's Book of Psalms contains 28 quires marked by letters of the alphabet A-Z^8 Aa-Ee^8.

Shute (2017: 3–6) provides a detailed discussion of possible reasons for spelling variation in early printed books. She argues that there are two ways in

[2] The dating of the fourth Psalter, i.e. the one printed by Edward Whitchurch in London, is uncertain (Wójcik 2019).

[3] Joye's English translation is a rendering of Martin Bucer's Latin translation from Hebrew issued in 1529. Bucer was one of the prominent leaders of the Strasbourg Reformation (Juhász 2014: 208).

[4] Zwingli's Psalter was a posthumously published Latin translation of the Hebrew Psalter, which appeared in 1532 under the title *Enchiridion Psalmorum*. It contained the text of the Psalm and a commentary. Butterworth and Chester (1962: 129) state that the exact reasons why Joye was inclined to do another translation only four years after preparing the 1530 version is not known, but they suggest that upon learning about Zwingli's Latin Psalter from 1532 he may have felt that Zwingli's version was superior to that of Bucer.

which the spellings of different people can be introduced into the texts and hence be responsible for spelling variation: first, the layers of spellings can be introduced into the text through successive copying of a copy text used in book production, and second, the discrete sections of different spellings can be introduced into the text through more than one person (compositor) working on typesetting a text at the same time. In the case of Joye's Psalms it is rather unlikely that the text translated by Joye possessed numerous successive copies since, as mentioned above, the 1534 translation of the Psalms was a new translation prepared on the basis of a different Latin text than his original 1530 translation so, quite simply, the copy text used for printing could not be the product of successive copying. It follows, then, that whatever differences in spelling variation are found between different parts (quires) of the book, the differences had to be most likely introduced by multiple compositors typesetting different quires of the Psalms.

The text of George Joye's 1534 Psalms analysed here was made available as part of the *Early English Books Online* Text Creation Partnership (*EEBO*-TCP). For the spelling data to be used in the analysis they had to be specially prepared since *EEBO*-TCP transcriptions use a set of conventions for representing early printed texts which have to be taken into account. For example, *EEBO*-TCP texts use the vertical bar symbol | to denote the division of a word over two lines marked by the double oblique hyphen in the original ⸗. The symbol ¦ is used to mark words which were divided between two lines without any hyphenation in the original. These marks had to be first removed from the text or otherwise the software which was used to count the number of variant occurrences would not be able to correctly analyse a word represented for example as *kin|ge* as the intended *kinge*. On the other hand, *EEBO*-TCP texts retain all the abbreviations used in the original text and use a tilde, as in *ād*, where it stands for <*n*>. Similarly, typical abbreviations for *the* and *that*, i.e. y^e and y^t are retained in *EEBO*-TCP texts.[5] I retained all the abbreviations used in the text and counted all variant spellings containing abbreviations since their use may serve as an important indication of idiolectal spelling variation in the analysed text. Finally, 171 words from the text could not be properly read due to fragments of the text being illegible. *EEBO*-TCP uses special marking to highlight illegible words – these words were not taken into account in the analysis.

In the next step the text of the Psalms was processed using the VARD software (Baron and Rayson 2008), which was designed to assist users of historical corpora in dealing with spelling variation, particularly in early MnE texts. The most important feature of this software is the ability to detect different spelling variants present in the text and to provide frequencies of their

[5] The use of abbreviations in early printed books is just one example of the many continuities between early print and manuscript culture (Hotchkiss and Robinson 2008: 47).

use.[6] Each quire of the text was processed individually so that all variant spellings within the 28 quires of Joye's Psalms were gathered together and the frequency of the occurrence of each variant has been calculated with the VARD software. The whole text of the Psalms contains 47661 words representing 5472 distinct spellings. Some words were not taken into account in the analysis as they either represented words which were always spelt in the same way (e.g. pronouns *I*, *he* or *she*), or the difference in spelling was phonological in nature (e.g. an indefinite article spelt *a* before a consonant and *an* before a vowel). After the elimination of all such items irrelevant for the purposes of this study, the final data set was reduced to 24410 occurrences representing 1128 distinct spelling variants. The analysis also took into account the frequency of the use of abbreviations. In particular, I focused on the frequency of abbreviations using a tilde over a vowel letter for <n> or <m>, as well as abbreviations for *the* and *that*, i.e. y^e and y^t. In the final step of the data preparation process, the data were fed into Excel spreadsheet to be sorted and converted into a format which can be read by the computer software R used to perform the similarity calculations. Table 1 provides a sample of the dataset where 8 spelling variants of *about*, *almighty* and *are* in all 28 quires are presented. The entire dataset contains 1128 columns representing 1128 spelling variants which were subjected to analysis.

Table 1. Frequencies of 8 spelling variants in 28 quires of Joye's 1534 Psalms

	about	**aboute**	**almighty**	**almightye**	**almyghty**	**almyghtye**	**ar**	**are**
Quire A	1	0	0	0	0	0	0	8
Quire B	0	0	0	0	0	0	0	7
Quire C	1	0	0	1	0	0	1	4
Quire D	1	1	0	1	0	0	2	4
Quire E	0	1	1	0	2	0	3	4
Quire F	2	2	0	0	1	0	0	8
Quire G	0	0	0	0	0	0	2	9
Quire H	0	0	0	0	0	0	1	4
Quire I	0	2	0	0	0	0	3	4
Quire K	1	0	0	0	0	0	2	5

[6] The VARD software cannot distinguish between homographs. For example, a high-frequency word like <*the*> can stand both for a definite article *the* and a 2nd-person pronoun *thee*. Consequently, the frequencies of <*the*> provided by VARD had to be manually checked and each instance needed to be individually categorised on the basis of the context. Other instances of homographs were not taken into account in the analysis. The omission of these words has a negligible impact on the overall result of the analysis.

	about	aboute	almighty	almightye	almyghty	almyghtye	ar	are
Quire L	1	3	0	0	0	0	0	7
Quire M	0	0	2	0	0	1	1	6
Quire N	0	0	0	0	1	0	2	5
Quire O	0	1	0	0	0	0	0	6
Quire P	0	3	0	0	1	0	2	2
Quire Q	1	0	0	0	0	0	0	1
Quire R	0	0	0	0	0	0	1	14
Quire S	1	2	0	0	0	1	5	7
Quire T	0	3	1	1	0	0	3	6
Quire V	0	0	0	0	0	0	3	7
Quire X	0	0	0	0	1	0	0	6
Quire Y	0	1	0	0	0	0	0	5
Quire Z	0	2	0	0	0	0	6	3
Quire Aa	0	0	0	0	0	0	9	3
Quire Bb	0	1	0	0	0	0	2	10
Quire Cc	0	0	0	0	0	0	2	1
Quire Dd	0	1	0	0	0	0	6	0
Quire Ee	0	0	1	0	0	0	4	1
Total	**9**	**23**	**5**	**3**	**6**	**2**	**60**	**147**

3. Cosine similarity between quires of Joye's Psalms

Since the goal of this paper is to measure the degree of similarity between spellings used in different quires of Joye's Psalms it is necessary to find a way of measuring the extent to which the same combinations of spellings are used in different quires of the text. As such the task is conceptually similar to the task of measuring the degree of similarity between different texts. As observed by Wang and Dong (2020), text similarity measurements are the basis of natural language processing tasks, which play an important role in information retrieval, automatic question answering, machine translation, dialogue systems, and document matching.

One of the methods of measuring text similarity is based on calculating the distance between two texts, which traditionally has been assessed by measuring length distance, using the numerical characteristics of text to calculate the distance length of vector text (Wang and Dong 2020: 421). As observed by

Welbers *et al.* (2017: 246), such distance measurements use bag-of-words text analysis models, meaning that only the frequencies of words per text are used and word positions are ignored. One of the most common formats for representing a text in a bag-of-words format is a document term matrix (DTM), which is a matrix in which rows are documents, columns are terms, and cells indicate how often each term occurred in each document. The advantage of this representation is that it allows the data to be analysed with vector and matrix algebra, effectively moving from text to numbers (Welbers *et al.* 2017: 252).

To use this model for the analysis of spelling variation between different quires of Joye's Psalms, the text has to be turned into a document term matrix in which rows are quires and columns are spelling variants with cells indicating the frequency of occurrence of each spelling variant in every quire, i.e. it has to be represented as in Table 1 above. To perform similarity calculations using vector algebra, each quire has to be represented by what is called a term-frequency vector, where frequencies of occurrence of each spelling variant are components of a vector and each spelling variant adds a new dimension to a term-frequency vector. For example, in Table 1 quire A is represented by an 8-dimensional term-frequency vector (1, 0, 0, 0, 0, 0, 0, 8), while quire B is represented by (0, 0, 0, 0, 0, 0, 0, 7)[7] and so on. Han *et al.* (2012: 77) make a crucial observation concerning term-frequency vectors and note that they are typically very long and sparse (i.e. they have many 0 values). They further note that traditional distance measures do not work well for such sparse numeric data because two term-frequency vectors may have many 0 values in common, meaning that the corresponding quires do not share many spelling variants, but this does not make them similar. Han *et al.* (2012: 77) emphasise that when dealing with sparse term-frequency vectors what is needed is a measure of similarity that will focus on the words and their frequencies (spelling variants and their frequencies) that the documents (quires) have in common, i.e. a measure for numeric data that ignores zero-matches. Their proposal for measuring similarity between sparse term-frequency vectors is to use a measure of similarity known as cosine similarity, which computes the cosine of the angle between vectors. A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no match. The closer the cosine value to 1, the smaller the angle and the greater the match (similarity) between vectors (Han *et al.* 2012: 78). The mathematical formula for calculating cosine similarity is given below.

$$similarity\ (x,\ y) = \cos(\theta) = \frac{x \cdot y}{||x||\ ||y||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

[7] Recall that Table 1 represents a sample of the data which consist of 1128 spelling variants corresponding to 1128 dimensions of vectors representing quires of the text.

where x and y are n-dimensional vectors, θ is an angle between vectors, while A_i B_i are components of vectors x and y. Table 2 below provides a sample of cosine similarity scores with regard to spelling variation in Joye's 1534 Psalms. The table presents the similarities[8] for quires A, B, C, D, E, F, and G. Due to space limitations the raw data for all the quires will not be presented.

Table 2. Cosine similarity scores for quires A-G

	Quire A	**Quire B**	**Quire C**	**Quire D**	**Quire E**	**Quire F**	**Quire G**
Quire A	1.0000	0.9578	0.9571	0.9568	0.9711	0.9614	0.9602
Quire B	0.9578	1.0000	0.9632	0.9564	0.9682	0.9632	0.9595
Quire C	0.9571	0.9632	1.0000	0.9550	0.9647	0.9601	0.9586
Quire D	0.9568	0.9564	0.9550	1.0000	0.9655	0.9613	0.9496
Quire E	0.9711	0.9682	0.9647	0.9655	1.0000	0.9701	0.9637
Quire F	0.9614	0.9632	0.9601	0.9613	0.9701	1.0000	0.9535
Quire G	0.9602	0.9595	0.9586	0.9496	0.9637	0.9535	1.0000
Quire H	0.9162	0.8923	0.9073	0.9203	0.9234	0.9304	0.9133
Quire I	0.9361	0.9414	0.9338	0.9444	0.9496	0.9291	0.9204
Quire K	0.9243	0.9120	0.9238	0.9394	0.9299	0.9342	0.9348
Quire L	0.9390	0.9099	0.9174	0.9182	0.9256	0.9367	0.9281
Quire M	0.9406	0.9333	0.9407	0.9462	0.9453	0.9610	0.9243
Quire N	0.9369	0.9238	0.9412	0.9437	0.9435	0.9469	0.9431
Quire O	0.9266	0.9006	0.9137	0.9230	0.9208	0.9320	0.9225
Quire P	0.9293	0.9207	0.9347	0.9276	0.9346	0.9268	0.9429
Quire Q	0.7942	0.7986	0.8300	0.8110	0.7986	0.8204	0.8030
Quire R	0.8096	0.7953	0.8144	0.8306	0.8118	0.8280	0.7901
Quire S	0.8194	0.8179	0.8265	0.8477	0.8290	0.8488	0.8335
Quire T	0.9221	0.9297	0.9149	0.9155	0.9275	0.9227	0.9257
Quire V	0.8275	0.8351	0.8464	0.8206	0.8412	0.8516	0.8397
Quire X	0.7654	0.7766	0.8087	0.7671	0.7712	0.8027	0.8002
Quire Y	0.9122	0.9177	0.9302	0.9142	0.9224	0.9357	0.9373
Quire Z	0.9359	0.9512	0.9431	0.9275	0.9590	0.9483	0.9384

[8] All calculations were performed using the software R and the *lsa* package (Wild 2015). Heat map plots presented in Tables 9 and 11 were generated with *ggplot2* package (Wickham 2016).

Quire Aa	0.5816	0.5707	0.5950	0.6114	0.5969	0.5747	0.5816
Quire Bb	0.8407	0.8313	0.8490	0.8589	0.8529	0.8356	0.8499
Quire Cc	0.9193	0.9338	0.9358	0.9135	0.9349	0.9270	0.9185
Quire Dd	0.8405	0.8273	0.8515	0.8496	0.8590	0.8548	0.8281
Quire Ee	0.7810	0.8193	0.8278	0.7903	0.8137	0.8323	0.7858

The most interesting aspect of the spelling similarity measurements between different quires of Joye's Psalms is finding out which quires are most similar in terms of spelling and which ones are most distinct. By way of illustration, let us take a look at similarity scores between quire A and all the other quires. The scores are sorted from the most similar (score 1, i.e. complete identity, which is the result of measuring similarity between a given quire and itself) to the least similar (score 0.5816).

Table 3. Similarity scores for quire A sorted from most similar to least similar

Quire A	**Quire A**	**Quire E**	**Quire F**	**Quire G**	**Quire B**	**Quire C**	**Quire D**
	1.0000	0.9711	0.9614	0.9602	0.9578	0.9571	0.9568
	Quire M	**Quire L**	**Quire N**	**Quire I**	**Quire Z**	**Quire P**	**Quire O**
	0.9406	0.9390	0.9369	0.9361	0.9359	0.9293	0.9266
	Quire K	**Quire T**	**Quire Cc**	**Quire H**	**Quire Y**	**Quire Bb**	**Quire Dd**
	0.9243	0.9221	0.9193	0.9162	0.9122	**0.8407**	**0.8405**
	Quire V	**Quire S**	**Quire R**	**Quire Q**	**Quire Ee**	**Quire X**	**Quire Aa**
	0.8275	**0.8194**	**0.8096**	**0.7942**	**0.7810**	**0.7654**	**0.5816**

An inspection of the data in Table 3 reveals that quire A is most similar to quire E (score 0.9711) in terms of spellings used and least similar to quire Aa (score 0.5816). Following Shute (2017) and her results obtained on the basis of a quantitative analysis of spellings in Caxton's texts, it will be assumed that similarity score will be around 0.9, if the text was typeset by one compositor and there is not a change in the copy text.[9] When independent evidence exists revealing

[9] Shute's (2017) results were obtained by measuring similarities between different quires of books which are known to have been typeset by a single compositor. The results she obtained for different texts were 0.90. 0.89 0.90. 0.92, hence her assumption that the similarity result around 0.9 indicates that different parts of the book were made by a single compositor. In effect, the similarity measure around 0.9 can be taken as a measure of inherent spelling variation in an early Modern English text.

that Caxton's texts were typeset by two compositors the similarity score reported by Shute (2017: 165) is 0.81 or below, depending on the text analysed. Applying Shute's results as baseline measures to the similarities between quire A and the remaining quires suggests that 9 quires (Bb, Dd, V, S, R, Q, Ee, X, and Aa) were typeset by a different compositor than quire A since similarity scores for all these quires are well below the baseline 0.9. This is indicated by grey shading in the table. Tables 4–8 below provide sorted similarity scores for Quires B, C, D, E, and F. In all cases, similarities below 0.9 are shaded in grey.

Table 4. Similarity scores for quire B sorted from most similar to least similar

Quire B	**Quire B**	**Quire E**	**Quire F**	**Quire C**	**Quire G**	**Quire A**	**Quire D**
	1.0000	0.9682	0.9632	0.9632	0.9595	0.9578	0.9564
	Quire Z	**Quire I**	**Quire Cc**	**Quire M**	**Quire T**	**Quire N**	**Quire P**
	0.9512	0.9414	0.9338	0.9333	0.9297	0.9238	0.9207
	Quire Y	**Quire K**	**Quire L**	**Quire O**	**Quire H**	**Quire V**	**Quire Bb**
	0.9177	0.9120	0.9099	0.9006	0.8923	**0.8351**	**0.8313**
	Quire Dd	**Quire Ee**	**Quire S**	**Quire Q**	**Quire R**	**Quire X**	**Quire Aa**
	0.8273	**0.8193**	**0.8179**	**0.7986**	**0.7953**	**0.7766**	**0.5707**

Quires V, Bb, Dd, Ee, S, Q, R, X, and Aa once again have similarity scores well below 0.9. The case of quire H is less straightforward as its score is 0.8923, which places it below 0.9 but by a very narrow margin.

Table 5. Similarity scores for quire C sorted from most similar to least similar

Quire C	**Quire C**	**Quire E**	**Quire B**	**Quire F**	**Quire G**	**Quire A**	**Quire D**
	1.0000	0.9647	0.9632	0.9601	0.9586	0.9571	0.9550
	Quire Z	**Quire N**	**Quire M**	**Quire Cc**	**Quire P**	**Quire I**	**Quire Y**
	0.9431	0.9412	0.9407	0.9358	0.9347	0.9338	0.9302
	Quire K	**Quire L**	**Quire T**	**Quire O**	**Quire H**	**Quire Dd**	**Quire Bb**
	0.9238	0.9174	0.9149	0.9137	0.9073	**0.8515**	**0.8490**
	Quire V	**Quire Q**	**Quire Ee**	**Quire S**	**Quire R**	**Quire X**	**Quire Aa**
	0.8464	**0.8300**	**0.8278**	**0.8265**	**0.8144**	**0.8087**	**0.5950**

Table 6. Similarity scores for quire D sorted from most similar to least similar

Quire D	**Quire D**	**Quire E**	**Quire F**	**Quire A**	**Quire B**	**Quire C**	**Quire G**
	1.0000	0.9655	0.9613	0.9568	0.9564	0.9550	0.9496
	Quire M	**Quire I**	**Quire N**	**Quire K**	**Quire P**	**Quire Z**	**Quire O**
	0.9462	0.9444	0.9437	0.9394	0.9276	0.9275	0.9230
	Quire H	**Quire L**	**Quire T**	**Quire Y**	**Quire Cc**	**Quire Bb**	**Quire Dd**
	0.9203	0.9182	0.9155	0.9142	0.9135	**0.8589**	**0.8496**
	Quire S	**Quire R**	**Quire V**	**Quire Q**	**Quire Ee**	**Quire X**	**Quire Aa**
	0.8477	**0.8306**	**0.8206**	**0.8110**	**0.7903**	**0.7671**	**0.6114**

Table 7. Similarity scores for quire E sorted from most similar to least similar

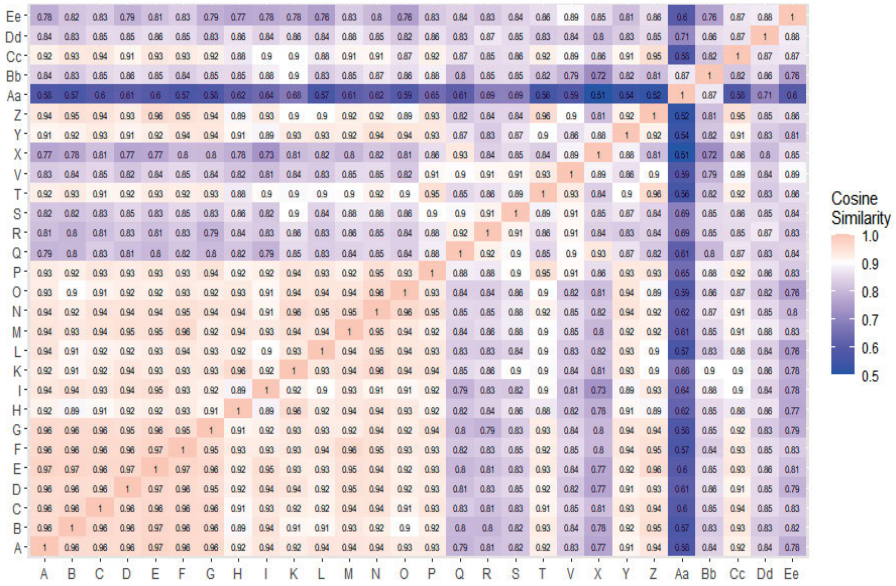
Quire E	**Quire E**	**Quire A**	**Quire F**	**Quire B**	**Quire D**	**Quire C**	**Quire G**
	1.0000	0.9711	0.9701	0.9682	0.9655	0.9647	0.9637
	Quire Z	**Quire I**	**Quire M**	**Quire N**	**Quire Cc**	**Quire P**	**Quire K**
	0.9590	0.9496	0.9453	0.9435	0.9349	0.9346	0.9299
	Quire T	**Quire L**	**Quire H**	**Quire Y**	**Quire O**	**Quire Dd**	**Quire Bb**
	0.9275	0.9256	0.9234	0.9224	0.9208	**0.8590**	**0.8529**
	Quire V	**Quire S**	**Quire Ee**	**Quire R**	**Quire Q**	**Quire X**	**Quire Aa**
	0.8412	**0.8290**	**0.8137**	**0.8118**	**0.7986**	**0.7712**	**0.5969**

Table 8. Similarity scores for quire F sorted from most similar to least similar

Quire F	**Quire F**	**Quire E**	**Quire B**	**Quire A**	**Quire D**	**Quire M**	**Quire C**
	1.0000	0.9701	0.9632	0.9614	0.9613	0.9610	0.9601
	Quire G	**Quire Z**	**Quire N**	**Quire L**	**Quire Y**	**Quire K**	**Quire O**
	0.9535	0.9483	0.9469	0.9367	0.9357	0.9342	0.9320
	Quire H	**Quire I**	**Quire Cc**	**Quire P**	**Quire T**	**Quire Dd**	**Quire V**
	0.9304	0.9291	0.9270	0.9268	0.9227	0.8548	0.8516
	Quire S	**Quire Bb**	**Quire Ee**	**Quire R**	**Quire Q**	**Quire X**	**Quire Aa**
	0.8488	0.8356	0.8323	0.8280	0.8204	0.8027	0.5747

The emerging pattern is quite clear: quires Q, R, S, V, X, Aa, Bb, Dd, and Ee are the 9 quires which consistently score well below 0.9 when compared with the first 6 quires of Joye's Psalms, i.e. quires A, B, C, D, E, and F. Additionally, it can be observed that quire Aa's score is always the lowest, its spellings differ from those found in other quires by the biggest margin. Whether this tendency continues for the rest of the quires can be assessed by inspecting a cosine similarity heat map presented in Table 9 below.

Table 9. Cosine similarity heat map for all quires



The heat map uses colour coding to display information about cosine similarities between all 28 quires. As the lowest similarity score (0.5051) was obtained between quires X and Aa, the scale in the heat map starts with 0.5, which is marked in dark blue. All results falling between 0.5 and the baseline result 0.9 are marked in shades of blue and the decreasing intensity of the blue colour in the plot indicates more similar scores between the compared quires. The baseline score, i.e. 0.9 is marked in white, while results higher than 0.9 are marked by shades of pink, where the growing intensity of the colour indicates rising similarity. It is important to remember that the baseline result, i.e. the similarity score of 0.9 is the assumed cut point differentiating between quires which were typeset by different compositors. This means that all results which are marked in Table 9 in different shades of blue indicate that the compared quires were typeset by different compositors than those marked in white or

shades of pink, i.e. scores of 0.9 or higher. It can be noticed that the tendency we observed for the first 6 quires, namely that their spellings differed consistently from those in quires Q, R, S, V, X, Aa, Bb, Dd, Ee continues also for quires G, H, I, K, L, M, N, O, P, T, Y, Z, and Cc. We can thus distinguish two clear groups of quires with respect to the overall similarity of spelling variants used in their texts – quires whose mutual similarity scores are higher than 0.9, i.e. quires A, B, C, D, E, F, G, H, I, K, L, M, N, O, P, T, Y, Z, and Cc; and quires Q, R, S, V, X, Aa, Bb, Dd, Ee, whose similarity scores below 0.9 indicate that they were typeset by a different compositor or perhaps different compositors – the point which I shall return to below.

An inspection of the similarity heat map in Table 9 allows us to make further interesting observations. Note, for example, that within the first group of quires it is possible to distinguish two sub-groups. Quires A, B, C, D, E, F, and G are particularly stable with respect to spelling variants used and consistently display similarity scores of 0.95 and higher, while the second sub-group comprising quires H, I, K, L, M, N, O, P, T, Y, Z, and Cc generally shows similarity scores between 0.9 and 0.94 with only three scores falling outside this range by a small margin: between Cc and H/L (0.88); and Cc and O (0.87). While the results for all quires in this group are above the assumed baseline score of 0.9 indicating a single compositor, the observed differences perhaps suggest that the copy text used by compositors contained layers of spelling variation which are detected by similarity scores – a possibility which was dismissed at the outset perhaps too hastily, assuming that the copy text of Joye's new translation of the Psalms could not have been affected by significant spelling variation.

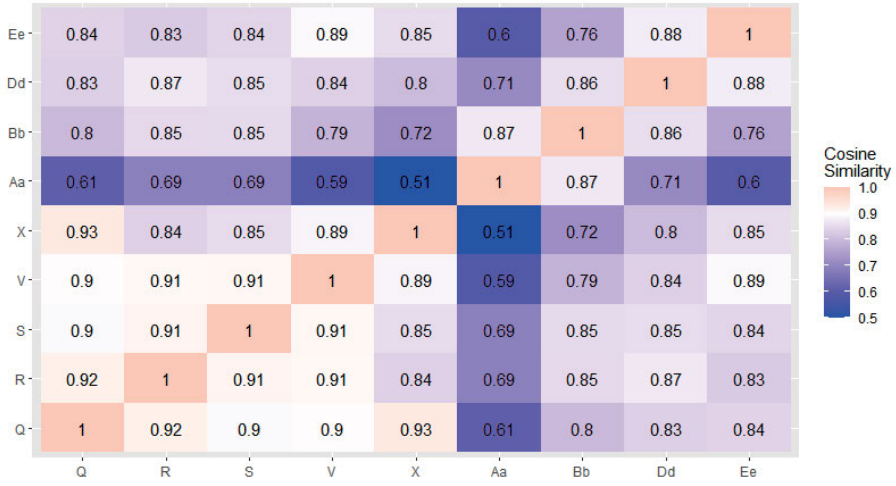
Another interesting observation which can be made concerns the similarity scores for quire Aa. These are given in Table 10.

Table 10. Similarity scores for quire Aa sorted from most similar to least similar

Quire Aa	**Quire Aa**	**Quire Bb**	**Quire Dd**	**Quire S**	**Quire R**	**Quire K**	**Quire P**
	1,0000	0.8725	0.7120	0.6912	0.6909	0.6778	0.6457
	Quire I	**Quire N**	**Quire H**	**Quire M**	**Quire Q**	**Quire D**	**Quire Ee**
	0.6403	0.6243	0.6226	0.6145	0.6123	0.6114	0.5990
	Quire E	**Quire C**	**Quire O**	**Quire V**	**Quire Cc**	**Quire G**	**Quire A**
	0.5969	0.5950	0.5943	0.5887	0.5828	0.5816	0.5816
	Quire F	**Quire B**	**Quire L**	**Quire T**	**Quire Y**	**Quire Z**	**Quire X**
	0.5747	0.5707	0.5659	0.5573	0.5358	0.5195	0.5051

As can be seen, similarity scores between quire Aa and the majority of other quires are particularly low. Only one quire (Bb) displays a similarity score in the vicinity of the baseline 0.9 score. The average similarity score for quire Aa is 0.6237 with the median value of 0.5979. The similarity heat map tells the same story, as the intensity of the heat map colour is the greatest for quire Aa, which clearly stands out from the rest of the quires depicted in Table 9. The results suggest that quire Aa was typeset by a compositor involved only in making this single quire of Joye's Psalms. Quire Aa clearly stands out as an outlier and the similarity scores indicate that its spellings are different than the ones used in all other quires. Consider Table 11 below, where similarities between 9 quires, i.e. Q, R, S, V, X, Aa, Bb, Dd, Ee are presented once again but this time the comparison is made only between these nine quires, which declutters the picture and allows for a better comparison of tendencies displayed by this group of quires.

Table 11. Cosine similarity heat map for quires Q, R, S, V, X, Aa, Bb, Dd, Ee



Note that on closer inspection the nine quires we singled out initially because their similarity scores were well below 0.9 do not form a homogeneous group. On the one hand, similarities for quires Q, R, S, and V are placed within 0.9 – 1.0 range expected for fragments of texts created by one compositor, but quires X, Aa, Bb, Dd, and Ee almost always score below 0.9 when compared to all other quires. Quire X shows some affinity with quires Q (score 0.93) and V (0.89) but its spellings display similarities below 0.9 when compared with quires R, S, Aa, Bb, and Dd. Quire Aa, as noted above, is an obvious outlier as its spellings are significantly different than those used in all other quires. In the case of quires Bb,

Dd, and Ee this difference is not as striking as in the case of quire Aa, but their similarity scores are also consistently different from those identified in other quires. The emerging picture is that of more than one compositor involved in typesetting quires Q, R, S, V, X, Aa, Bb, Dd, and Ee. In all probability quires Q, R, S, V and possibly X were typeset by one compositor (different from the one who was responsible for quires A, B, C, D, E, F, G, H, I, K, L, M, N, O, P, T, Y, Z, and Cc), while the remaining quires, i.e. Aa, Bb, Dd, and Ee show the amount of spelling variation suggesting the work of four different individuals responsible for typesetting their text.

Conclusions

The goal of this paper was to find out whether different sections of the text of George Joye's Psalms printed in 1534 in Antwerp include similar spellings and to measure the degree of this similarity in a mathematically rigorous way. The analysis presented here involved the examination of 1128 spelling variants and the frequencies of their use in all 28 quires of Joye's Psalms by means of turning quires of the text into vectors and measuring cosine similarity between them. Working on the assumption that the differences in the range of spelling variation between individual quires of the text result from different compositors typesetting a given quire, the paper revealed the presence of distinct groups of quires whose different spelling patterns can be ascribed to a number of compositors working on individual quires of the text and leaving the trace of their orthographic idiolect. Differences between spelling variants used in parts of the text reveal quires in Joye's 1534 Psalms can be divided into the following groups: quires A, B, C, D, E, F, G, H, I, K, L, M, N, O, P, T, Y, Z, and Cc which were typeset by one compositor; quires Q, R, S, V and possibly X, which were typeset by a different compositor with quires Aa, Bb, Dd, Ee clearly standing out from the former two groups and additionally not showing significant degrees of similarity to one another.

The paper shows the applicability of cosine similarity measurements to examining internal spelling variation in early printed texts by allowing to systematise observations which inevitably spring to mind in the process of text examination and to objectivise the effects of these observations. The achieved results show that reliance on this method enhances traditional philological analyses in a significant way and can therefore contribute to our better understanding of historical phenomena.

References

Primary Source:

Joye, George. 1534. *Dauids Psalter, Diligently and Faithfully Tra[n]slated by George Ioye, with Breif Arguments before Euery Psalme, Declaringe the Effecte Therof*. Antwerp: Martyne Emperowr. Early English Books Online Text Creation Partnership. 2011. http://name.umdl.umich.edu/a13409.0001.001, accessed 29 November 2020

Secondary Sources:

Aronoff, M. 1989. The orthographic system of an Early English printer: Wynkyn de Worde. *Folia Linguistica Historica* 8(1–2): 65–97.

Baron, A., and P. Rayson 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Aston University, Birmingham, UK, 22 May 2008.

Blake, N. 1965. English versions of Reynard the Fox in the fifteenth and sixteenth centuries. *Studies in Philology* 62: 63–77.

Butterworth, Ch.C., and A.G. Chester 1962. *George Joye (1495?-1553). A Chapter in the History of the English Bible and the English Reformation*. Philadelphia, PA: University of Pennsylvania Press.

Fisher, J.H. 1996. *The Emergence of a Standard English*. Lexington, KY: University Press of Kentucky.

Gaskell, P. 1972. *A New Introduction to Bibliography*. Oxford: Oxford University Press.

Han, J., M. Kamber, and J. Pei 2012. *Data Mining: Concepts and Techniques*. 3rd ed. Waltham, MA.: Morgan Kaufmann Publishers.

Hellinga, L. 1999. Printing. In L. Hellinga and J. Trapp (eds.), *The Cambridge History of the Book in Britain*, 65–108. Cambridge: Cambridge University Press.

Horobin, S. 2001. The language of the fifteenth-century printed editions of *The Canterbury Tales. Anglia* 119(2): 249–258.

Hotchkiss, V., and F.C. Robinson 2008. *English in Print from Caxton to Shakespeare to Milton*. Urbana, IL: University of Illinois Press.

Howard-Hill, T.H. 2006. Early modern printers and the standardization of English spelling. *The Modern Language Review* 101(1): 16–29.

Juhász, G.M. 2014. *Translating Resurrection. The Debate between William Tyndale and George Joye in Its Historical and Theological Context*. Leiden and Boston, MA: Brill.

Lass, R. 1992. Phonology and morphology. In N. Blake (ed.), *The Cambridge History of the English Language*, 23–155. Cambridge: Cambridge University Press.

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

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Rutkowska, H. 2013. *Orthographic Systems in Thirteen Editions of the Kalender of Sheperdes (1506–1656)*. Frankfurt am Main: Peter Lang.

Samuels, M. 1963. Some applications of Middle English dialectology. *English Studies* 44: 81–94.

Scragg, D.G. 1974. *A History of English Spelling*. Manchester: Manchester University Press.

Shute, R. 2017. A Quantitative Study of Spelling Variation in William Caxton's Printed Texts. Doctoral dissertation, University of Sheffield.

Wang, J., and Y. Dong 2020. Measurement of text similarity: A survey. *Information* 11: 421. https://doi.org/10.3390/info11090421

Welbers K., W. Van Atteveldt, and K. Benoit 2017. Text analysis in R. *Communication Methods and Measures* 11(4): 245–265.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag. https://ggplot2.tidyverse.org

Wild, F. 2015. *lsa: Latent Semantic Analysis* (R package version 0.73.1) [Computer software]. https://CRAN.R-project.org/package=lsa

Wójcik, J. 2019. The first English printed psalters – George Joye's translations and their editions. *Roczniki Humanistyczne* 67(5): 143–154.