ROCZNIKI HUMANISTYCZNE
Tom LXXII, zeszyt 5 – 2024

ZESZYT SPECJALNY / SPECIAL ISSUE
DOI: https://doi.org/10.18290/rh24725.9s

JERZY WÓJCIK

ON MEASURING PSALM SIMILARITY: A CASE FOR WORD-LEVEL N-GRAMS

INTRODUCTION

In a number of recent papers, different versions of historical texts have been compared and their mutual dependencies studied using cosine similarity measurements. For example, Charzyńska-Wójcik (2021), Charzyńska-Wójcik and Wójcik (2022) and Wójcik (2023) demonstrated the efficacy of cosine similarity in tracing textual affinities between different versions of the Psalms translated into English, Lis and Wójcik (2023) employed cosine similarity to analyse the complex relationships between the French texts of the *Laws of Oléron* and their multiple eMnE translations, while Hordyjewicz (2023) used it in examining English translations of the *Book of Hours*. What these studies have demonstrated is that digital humanities methods not only complement but often surpass traditional philological analysis when it comes to identifying similarities between multiple historical versions of the same underlying (source) text. A common feature of these studies is the use of cosine similarity for gauging the difference between the objects of study, i.e. the compared texts, represented as term-frequency vectors, which makes it possible to express the similarities between the studied texts in objective terms. This has to be considered as a significant development since, as observed by Charzyńska-Wójcik and Wójcik (2022, p. 204), scholarly literature on Psalm translations all too often classifies different versions of the Psalms as "revisions, deep revisions, or … practically new translations", i.e.

Jerzy Wójcik, PhD, Assistant Professor at the Department of English-Polish Contrastive Studies, Institute of Linguistics, John Paul II Catholic University of Lublin; correspondence address: Instytut Językoznawstwa KUL, Al. Racławickie 14, 20-950 Lublin, Poland; e-mail: jwojcik@kul.pl; ORCID: https://orcid.org/0000-0001-5283-9017.

researchers resort to inherently subjective labels to express the relationships between the studied texts. At the same time, these compared historical texts are typically collections of translations of the same underlying original (e.g. different translations of the Psalms into English from Latin or Hebrew originals) so it should not come as a surprise that they share quite a number of features, which makes them notoriously difficult to compare without using some objective measures of similarity.

Although the cosine measure is one of the most common measures used for computing similarity between documents (Steinbach et al., 2000, p. 5), it is but one of the many methods that can be employed for gauging text similarity and the rapid development of various digital humanities tools and techniques we have been witnessing in the recent years makes it possible to tackle the problem of tracing textual dependencies by applying a variety of methods. In this context, one interesting approach comes from quantitative intertextual studies that concentrate on text-reuse detection that largely boil down to identifying chunks of text shared among compared texts. A number of projects such as PAIR (Olsen & Horton, 2009), Tesserae (Coffee et al., 2012), and TRACER (Buchler, 2016) have developed methods for the identification of text reuse and as such seem well-suited for finding similarities between different versions of the same text.

The aim of this paper is to assess the applicability of one of the above-mentioned text-reuse detection tools, namely Tesserae and compare it with cosine similarity measurements, for the study of textual affinities between different translations of the same source. The data for the analysis will be drawn from the variant texts of Psalm 6 found in 16th-century English manuals of devotion. It will be argued that although both Tesserae and cosine similarity are capable of identifying similarities between compared texts, Tesserae suffers from the lack of appropriate similarity measure, while cosine similarity measurements may be significantly improved by being applied to texts represented as vectors consisting of word-level n-grams rather than words alone. To the best of my knowledge, no other study exists that employs n-gram text representation for the analysis of historical text affinities.

1. TESSERAE: TEXT-REUSE DETECTION TOOL APPLIED FOR MEASURING TEXTUAL SIMILARITY

As indicated above, Tesserae (Coffee et al., 2012) is one of a number of projects developed in the recent years for the purpose of identifying cases of text reuse. In what follows

I will apply the tools developed by the Tesserae Project research group,[1] in particular their basic search and text-reuse detection functionality which was rewritten in the R programming language and made available as supplemental material to chapter 3 of Forstall and Scheirer (2019).[2] Tesserae is employed there to identify text reuse for detecting the presence of allusions between the source and target texts (Forstall & Scheirer, 2019, pp. 55–78).

In order to show how Tesserae[3] works, I will use the R code provided by Forstall and Scheirer (2019) and apply it to detect similarities between 20 early Modern English versions of Psalm 6 found in publications printed between 1530 and 1557.[4] The list of compared psalm versions is provided in (1).

(1)[5]

01 *Ortulus anime* from 1530 (STC 13828.4)
02 George Joye's English Psalter translated from the Latin text of Martin Bucer; first published in 1530 (STC 2370)[6]
03 George Joye's English Psalter translated from the Latin text of Huldrych Zwingli; first published in 1534 (STC 2372)
04 Marshall's primer from 1534 (STC 15986)
05 Godfray's primer from 1535 (STC 15988a)
06 Psalms from Coverdale's first complete Bible issued in 1535 (STC 2063)
07 Rouen primer from 1536 (STC 15993)
08 Redman's primer from 1537 (STC 15997)
09 *Manual of prayers* from 1539 (STC 16009)
10 Psalms from Coverdale's second complete Bible, known as the Great Bible; first issued in 1539 (STC 2068)
11 Psalms from Richard Taverner's Bible issued in 1539 (STC 2067)
12 1539 edition of Coverdale's Psalter translated from the Latin of Johannes Campensis (first printed in 1535) (STC 2372.6)

[1] The Tesserae Project search engine has been originally designed and developed at the University of Buffalo to find parallels between Latin poems. The project website offers tools for intertextual detection and analysis and is available at http://tesserae.caset.buffalo.edu/.3

[2] R code from chapter 3 of Forstall and Scheirer (2019) can be accessed here: https://github.com/wjs3/quantitative-intertextuality/tree/master/chapter3/src

[3] From now on, when referring to Tesserae what I have in mind is the R code provided in Forstall and Scheirer (2019) rather than the original Tesserae project website.

[4] The historical background of the versions of Psalm 6 in (1) is discussed in more detail in Wójcik (2023), where an analysis of these textual data using hierarchical clustering is provided.

[5] In what follows, the texts from (1) will be referred to by their numbers, i.e. 01, 02, 03, and so on.

[6] Joye's 1530 Psalter was the first complete English Psalter to be printed.

13 Coverdale's Psalter translated from the Vulgate; issued in 1540 (STC 2368)
14 Henry VIII's primer from 1545 (STC 16034)
15 *Book of Common Prayer* from 1549 (STC 16270a)
16 Primer from 1552 (STC 16057)
17 *Book of Common Prayer* from 1552 (STC 16288)
18 Caly's primer from 1555 (STC 16062)
19 Wayland's primer from 1555 (STC 16063)
20 Wayland's primer from 1557 (STC 16080)

The texts of Psalm 6 from (1) appeared in 13 manuals of devotion printed between 1530 and 1557, i.e. during the turbulent times of English Reformation. The publications selected for analysis include seven manuals printed during Henry VIII's rule, i.e. *Ortulus anime* from 1530[7] (STC 13828.4), Marshall's primer from 1534 (STC 15986), Godfray's primer from 1535 (STC 15988a), Rouen primer from 1536 (STC 15993), Redman's primer from 1537 (STC 15997),[8] *Manual of prayers* from 1539 (STC 16009), and Henry VIII's primer from 1545 (STC 16034). Three texts originate from the time of Edward VI's rule, i.e. the *Book of Common Prayer* from 1549 (STC 16270a), primer from 1552 (STC 16057), and *Book of Common Prayer* from 1552 (STC16288). Finally, there are three primers printed during Mary I's reign, i.e. Caly's Primer from 1555 (STC 16062), Wayland's Primers from 1555 (STC 16063) and 1557 (STC 16080). The text of Psalm 6 contained in these 13 manuals was juxtaposed for comparison with the text of this psalm appearing in seven prose translations of the Psalter which were in circulation at the time when the manuals were published either as new translations of the Psalter or as part of complete translations of the Bible. These seven translations of the Psalms used in the examination were: George Joye's English Psalter translated from the Latin text of Martin Bucer first published in 1530 (STC 2370), George Joye's English Psalter translated from the Latin text of Huldrych Zwingli first published in 1534 (STC 2372), Psalms from Coverdale's first complete Bible issued in 1535 (STC 2063), Psalms from Coverdale's second complete Bible (generally referred to as the Great Bible) first issued in 1539 (STC 2068), Psalms from Richard Taverner's Bible issued in 1539 (STC 2067), 1539 edition of Coverdale's Psalter translated from the Latin of Johannes Campensis (first printed in 1535) (STC 2372.6), and Coverdale's Psalter translated from the Vulgate in 1540 (STC 2368).

The documents loaded into memory by Tesserae's text loading function are stored as R data structures known as environments. These contain all the information associated

[7] This is a revised 1530 edition of Joye's lost *Primer* from 1529 (Butterworth & Chester, 1962, p. 52).
[8] The original 1536 edition is in the possession of the Bibliothèque Nationale in Paris. Here, I rely on a 1537 edition.

with a given ingested text. Table 1 shows the representation of the first 10 and the last 10 words (tokens) of Psalm 6 from 1530.

Table 1. A data table representation of the text ingested into R

display	***form***	***type***	***unitid***	***tokenid***
Oh	oh	W	1	1
Lord	lord	W	1	2
/		P	1	3
rebuke	rebuke	W	1	4
me	me	W	1	5
not	not	W	1	6
in	in	W	1	7
thy	thy	W	1	8
wrath	wrath	W	1	9
:		P	1	10
they	they	W	10	181
shall	shall	W	10	182
be	be	W	10	183
put	put	W	10	184
to	to	W	10	185
flight	flight	W	10	186
and	and	W	10	187
confounded	confounded	W	10	188
suddenly	suddenly	W	10	189
.		P	10	190

As can be seen, the *display* column contains all the elements of the original text (including punctuation marks), while *form* holds all words presented in lowercase. The *type* column distinguishes words (W) from punctuation marks (P). Columns marked as *unitid* and *tokenid* provide indexing information for each element of the text. All the elements associated with a given psalm verse will have the same *unitid*: the first 10 words coming from verse 1 are all marked 1, while the last 10 words are located in verse 10, which is also indicated by the *unitid*. In contrast, each text element has its own *tokenid*.

Using Tesserae it is also possible to extract n-grams[9] from the ingested text. Table 2 below shows the first five bigrams and trigrams together with their *tokenid*, which correspond to the *tokenid* of the first element of the n-gram. This n-gram information can be added to a data structure by using Tesserae's add.col.ngrams() function.

Table 2. First five bigrams and trigrams extracted from the text of Psalm 6 from text 01

ramiform	***tokenid***	***ngram_form***	***tokenid***
oh-lord	1	oh-lord-rebuke	1
lord-rebuke	2	lord-rebuke-me	2
rebuke-me	4	rebuke-me-not	4
me-not	5	me-not-in	5
not-in	6	not-in-thy	6

Once all the documents under examination are stored in R as environments containing all the word and/or n-gram and indexing information, it is possible to perform the next step in identifying text-reuse between compared documents, i.e. linking. Linking means that indices from two texts of interest are compared for each word or n-gram which occurs in both texts, and every such location in one text is linked to every location from the second text. Consider Table 3, where texts of Psalm 6 from 1530 (text 01) and 1534 (text 03) are compared and the *tokenids* of identical words from both texts are linked together. The table shows the result for the first five words, i.e. "oh lord rebuke me not" (*s.tokenid* = 1,2,4,5,6), of Psalm 6 from text 01. In Table 3, *s.tokenid* stands for *tokenid* in the source text (Psalm 6 from text 01), while *t.tokenid* means *tokenid* in the target text (Psalm 6 from text 03). Likewise, *s.unitid* and *t.unitid* refer to source and target *unitid*, i.e. psalm verse numbers. The total number of cases where a word from the source text was found in the target text was 412.

[9] I postpone a more detailed discussion of n-grams until section 2.2, where they will be employed in cosine similarity calculations.

Table 3. Words occurring both in the source and target texts

s.tokenid	*t.tokenid*	*feature*	*s.unitid*	*t.unitid*
1	147	oh	1	8
2	1	lord	1	1
2	25	lord	1	2
2	35	lord	1	2
2	62	lord	1	4
2	156	lord	1	8
2	167	lord	1	9
2	175	lord	1	9
4	3	rebuke	1	1
5	4	me	1	1
5	15	me	1	1
5	24	me	1	2
5	33	me	1	2
5	70	me	1	4
5	145	me	1	8
5	170	me	1	9
6	5	not	1	1
6	16	not	1	1

The next step in Tesserae's procedure is gathering, i.e. collecting all cases of shared elements by *s.unitid* and *t.unitid* (i.e. verse numbers in our case). This effectively transforms Table 3 and results in a list of all instances where a word is shared between the two psalm versions at the level of a single verse. This is illustrated in Table 4, where the column named *shared* lists all instances from column *feature* in Table 3 which share a given *s.unitid* and *t.unitid*. Table 4 shows the result for the first two verses of the source text (*s.unitid* = 1 and 2). The complete table contains 65 such *s.unitid* and *t.unitid* combinations. The identified words shared between psalm verses compared are displayed in the *shared* column in Table 4 in alphabetical order.[10]

[10] It has to be remembered that in the previous step, i.e. linking, every location housing a word in one text has been linked to every location with the same word from the second text. This will result in some words appearing multiple times in the 'shared' column. For example, if a word 'in' is used twice in the first verse of both the source and target text, it will appear as many as four times in the 'shared' column because both occurrences in the source text are shared with two occurrences in the target text.

Table 4. All instances of shared words gathered together

s.unitid	*t.unitid*	*shared*
1	1	anger in in in in lord me me me me neither not not rebuke thy thy thy thy wrath
1	8	lord me me oh
1	4	lord me me thy thy
1	2	in in lord lord me me me me
1	9	lord lord me me
2	10	all are my sore
2	2	all am am for for for heal i i lord lord lord lord me me me me
2	7	am i my my with
2	8	all but for for lord lord me me my oh
2	3	but for for lord lord my
2	4	for for lord lord me me my
2	1	i lord lord me me me me
2	9	lord lord lord lord me me my
2	6	i i my my my with

There are a couple of interesting observations that can be made in connection with Table 4 in the context of using Tesserae as a tool for text comparison. Note, first of all, that the majority of the detected similarities involve the use of the same word in different verses of the compared psalm. This obviously follows from Tesserae's design and logic as its focus is on detecting text-reuse and, hence, on finding any shared features across the two texts compared, making it clearly unsuitable for measuring the level of similarity between texts as it produces a lot of irrelevant matches when applied to measuring the degree of similarity between variant translations of the same text. Clearly, if we want to find out how (dis)similar two versions of a psalm are, we are not interested in the fact that the same word(s) happen to be used in different verses of the compared psalms, since their presence in different parts of the text is coincidental. One way to overcome this problem would be to filter out the results obtained in Table 4 and display only those matches that indicate the use of the same item if *s.unitid* and *t.unitid* are the same. That means that we are concentrating on similarities between the same verses of the psalm. The result of filtering the detected similarities, so that only words shared between the same verses of the two compared psalms are taken into account, is demonstrated in Table 5.

Table 5. All instances of shared words with the same *s.unitid* and *t.unitid*

s.unitid	***t.unitid***	***shared***
1	1	anger in in in in lord me me me me neither not not rebuke thy thy thy thy wrath
2	2	all am am for for for for heal i i lord lord lord lord me me me me
3	3	but how long lord my soul
4	4	and deliver for lord me my save soul thee thy turn
5	5	for in in praise that that thee thee thee thee think who
6	6	bed every i i i i in my my my my my my night tears with with
7	7	enemies is my my my my up with with
8	8	avoid for from has heard lord me my of the the ye ye
9	9	has has has has heard lord lord lord lord my my received the the the the
10	10	all and and and and be be be be confounded enemies my shall shall shall shall shamed suddenly they they

The final step in Tesserae's procedure of identifying identical chunks of text in the compared documents is scoring. This is an important step in all text-reuse discovery techniques, which serves to rank the identified identical parts of text according to their significance. As explained by Forstall and Scheirer (2019, p. 66), "[t]he goal of the fourth step, scoring and filtering, is to rank these [results] in some way that raises meaningful intertexts to the top of the list and filters out the background noise." This background noise clearly stems from the fact that a lot of identical words identified by Tesserae in the two compared documents will be very common function words. To counter this, the Tesserae's scoring system attaches more importance to less common words and additionally takes into account the distance (in words) between the two most infrequent words from the detected list of identical words in the source and target documents.[11] Table 6 gives the results from Table 5 with the Tesserae's scores attached.

[11] Tesserae uses the following scoring formula: $score = ln\left(\frac{\sum\frac{1}{f(t_i)} + \sum\frac{1}{f(s_i)}}{d_t + d_s}\right)$, where $f(t_i)$ is frequency of the ith matching word in the target document, $f(s_i)$ is the frequency of the ith matching word in the source document, d_t is the inclusive distance in words between the two rarest matching words in the target phrase, d_s is the analogous distance in the source (Forstall & Scheirer 2019, p. 66). It is clear then that chunks of text involving rare words (where frequency is counted for each compared document separately) will score higher. Also, the scoring formula will attach higher scores to chunks of text whose elements are next to one another (i.e. their distance is smaller), which makes it more sensitive to identifying cases of text-reuse, i.e. identifying the same combinations of words in different places in the source and target documents.

Table 6. Tesserae scores for verse by verse comparison between Psalms from 1530 and 1534

s.unitid	***t.unitid***	***shared***	***score***
1	1	anger in in in in lord me me me me neither not not rebuke thy thy thy thy wrath	5.18
2	2	all am am for for for for heal i i lord lord lord lord me me me me	4.30
3	3	but how long lord my soul	5.61
4	4	and deliver for lord me my save soul thee thy turn	5.22
5	5	for in in praise that that thee thee thee thee think who	4.64
6	6	bed every i i i i in my my my my my my night tears with with	6.12
7	7	enemies is my my my my up with with	4.43
8	8	avoid for from has heard lord me my of the the ye ye	6.19
9	9	has has has has heard lord lord lord lord my my received the the the the	4.18
10	10	all and and and and be be be be confounded enemies my shall shall shall shall shamed suddenly they they	4.45

Having presented the details of text-reuse identification procedure used by Tesserae, we can now assess its applicability to the task of measuring similarity between the compared documents, i.e. different versions of Psalm 6. It has to be observed, first of all, that some modification needs to be introduced to Tesserae's calculations to adjust the method to serve as a tool for measuring text similarity. One way of doing this, for example, is to calculate the total sum of scores from Table 6 and treat the result as a measurement of pairwise document similarity. In the example in Table 6 above, the total sum of individual scores in the table is 50.32, and this could be regarded as a measure of text similarity between texts 01 and 03. When these total score values for all pairs of Psalm 6 versions listed in (1) are calculated the result is a 20 by 20 table with score results expressing the level of similarity between the analysed texts. For clarity of presentation, Table 7 shows the results for the first 10 versions of Psalm 6 from the list in (1).

Table 7. Tesserae scores expressing the level of similarity between texts[12]

	01	02	03	04	05	06	07	08	09	10
01	60.86	60.86	50.32	60.69	60.79	49.69	47.73	48.26	48.23	50.16
02	60.86	60.86	50.32	60.69	60.79	49.69	47.73	48.26	48.23	50.16
03	**50.32**	50.32	55.20	50.32	50.32	50.01	50.84	50.53	50.47	46.46
04	60.69	60.69	50.32	59.69	60.69	49.69	47.73	48.26	48.23	50.16
05	60.79	60.79	50.32	60.69	60.86	49.69	47.73	48.26	48.23	50.16
06	49.69	49.69	50.01	49.69	49.69	56.9	47.74	47.94	47.56	51.30
07	47.73	47.73	50.84	47.73	47.73	47.74	54.39	54.67	54.66	48.97
08	48.26	48.26	50.53	48.26	48.26	47.94	54.67	54.37	54.38	49.13
09	48.23	48.23	50.47	48.23	48.23	47.56	54.66	54.38	54.39	49.14
10	50.16	50.16	46.46	50.16	50.16	51.3	48.97	49.13	49.14	54.75

A quick look at the results is enough to spot a major problem. The measurements obtained give different score values when similarity between identical texts is measured (the grey cells in Table 7), e.g. the similarity of text 03 with itself is 55.20 while text 04 when measured with itself scores 59.69. In other words, the scoring system does not provide a uniform measure of sameness. What is more, some texts show higher scores when compared with other texts (04 with 05—score 60.69) than when they are compared with themselves (04 with 04—score 59.69). Consequently, it has to be concluded that the scoring formula used by Tesserae, or more precisely, the adaptation of Tesserae proposed here, is not useful for assessing the level of similarity between texts. This is not surprising in view of the fact that the scoring formula is sensitive to the frequency of items in a given text and the distance between the compared words as it was devised for identifying examples of text reuse.[13]

[12] Note the level of redundancy in Table 7 as, quite clearly, the score following from measuring similarity between text (i) and text (j) will be the same as the result of measuring similarity of text (j) and text (i). Consequently, parts of the table above and below the grey cells in Table 7 are their own mirror images. The grey cells themselves provide the score result of the level of similarity between a given psalm text and itself.

[13] It could be argued that perhaps the problem stems from the way in which Tesserae scored cases of text-reuse in the compared psalms, i.e. by looking at individual verses rather than the psalm as a whole. Note, however, that even if the basic chunk of text for comparison is modified and whole psalms are compared, the scoring formula will continue to be sensitive to the same parameters that will produce results as the ones documented in Table 7. The only way out here would be to modify the scoring formula, which only underscores the fact that the problem boils down to finding an appropriate measure of similarity between compared texts.

In what follows I will argue that cosine similarity which, as mentioned in the introductory section, has been successfully applied for the measurement of historical text similarity, may be significantly improved by being applied to texts represented as vectors consisting of word-level n-grams rather than words alone.

2. CALCULATING COSINE SIMILARITY BETWEEN PSALM 6 TEXTS

2.1 Calculating cosine similarity using words as features

As demonstrated in Charzyńska-Wójcik (2021), Charzyńska-Wójcik and Wójcik (2022), Wójcik (2023) and Lis and Wójcik (2023), the degree of similarity between various versions of historical texts can be successfully measured by means of cosine similarity. Performing cosine similarity calculations typically involves a number of steps, such as text preparation, text representation and, finally, performing the actual calculations.

In the context of working with historical texts, the first step,[14] i.e. text preparation, encompasses spelling normalisation,[15] which removes irrelevant differences between the compared texts caused by the lack of spelling standardisation. In the next step, the compared texts are represented as term-frequency vectors. This step is necessary since similarity calculations cannot be done on text documents directly (Feldman & Sanger, 2007, p. 68), hence each document is first represented as a term-frequency vector.[16] In this approach, a text document is regarded as a bag of words, that is, is represented as a set of words (i.e. terms) appearing in the document, completely disregarding word-order or word semantics (Sidorov, 2019, p. 14). Consequently, each word (term) in a document corresponds to a dimension in the resulting data space and each document then becomes a vector consisting of non-negative values (the number of occurrences of each term in a document) on each dimension (Huang, 2008, p. 50). The relevant formula is given in (2).

[14] Assuming, of course, that the text in question is already available in electronic form. If this is not the case, the steps described here have to be preceded by transcribing the text into electronic form.

[15] Normalisation here refers to *spelling* normalisation, i.e. eliminating textual differences following from spelling inconsistences typical of historical texts. This should not be confused with *data* normalisation (see e.g. Han et al., 2012, p. 83), whereby data are scaled to fall within a smaller range as part of the data-cleaning step in different data-mining techniques.

[16] Note that the psalm texts that were used for finding similarities using Tesserae's R code above were, even if somewhat indirectly, represented as term-frequency vectors. The example of the text ingested into R by Tesserae shown in Table 1 contains all the information necessary to obtain such a vector representing a given text.

(2) Document represented as an m-dimensional vector $\overrightarrow{t_d}$ (Huang, 2008, p. 50)

$\overrightarrow{t_d}$ = (tf(d, t_1), ……, tf(d, t_m)),

where D = {d_1,……, d_n} is a set of documents, T = {t_1, ……, t_m} a set of distinct terms occurring in D, and tf(d, t) denotes the frequency of term t ∈ T in document d ∈ D.

As indicated above, the formula in (2) depicts text representation in which all words appearing in a compared set of documents are distinct terms. Consequently, the number of *m* dimensions of each term-frequency vector $\overrightarrow{t_d}$ is equal to the number of different words in all of the compared documents.

The final step is calculating the cosine similarity[17] between the compared documents represented as vectors. The formula for cosine similarity is given in (3).

(3) Cosine similarity (Han et al., 2012, p. 78)

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

where *x* and *y* are n-dimensional vectors, is an angle between vectors *x* and *y*, while A_i B_i are components of vectors *x* and *y*

In effect, cosine similarity is the cosine of the angle between two vectors and is expressed by a value between 0 and 1, since the compared vector's dimensions can only be expressed by non-negative numbers as these correspond to term frequencies. Cosine similarity value of 1 indicates complete identity of the two texts, while the minimal possible value of 0 results from the two texts sharing no features in common.[18] As noted by Huang (2008, p. 50), an important property of cosine similarity is its independence of document length. This follows from the fact that documents with the same composition

[17] Of course, cosine similarity is not the only measure of document similarity/distance that could be applied here. However, as noted by Han et al. (2012, p. 77) alternative similarity/distance measures such as, Euclidean distance, Manhattan distance, or Jaccard coefficient do not perform well for sparse numeric data, i.e. data such as term-frequency vectors with a lot of dimensions represented as 0 values. These 0 values correspond to words attested in (at least one document within the compared document set D) and not present in a given document. Unlike traditional measures of similarity, cosine similarity is not sensitive to two objects having a lot of dimensions expressed as 0's, i.e. it does not make documents that do not have a lot of words in common seem similar.

[18] Note that, unlike Tesserae's scoring formula, cosine similarity provides a uniform framework for text comparison, where identical texts will receive the same score of 1, and all texts not sharing any features will uniformly score 0.

(the same words are present in both documents) but different totals (if, for example, the number of times all words are used is doubled) will be treated as identical because the angle between vectors representing texts will remain the same.[19]

To perform the cosine similarity calculations for all the versions of Psalm 6 listed in (1), I used the Tesserae's R code and its text loading function to ingest all the 20 texts into R. As discussed above, Tesserae stores texts in R as data structures known as environments which were used to create an R data frame with two columns—'doc_id' displaying text name and column 'text' containing psalm text in lowercase and without punctuation. Table 8 shows two rows (out of 20) of this data table containing data for text 01, i.e. *Ortulus anime* from 1530 and text 12, i.e. 1539 edition of Coverdale's Psalter translated from the Latin of Johannes Campensis.

Table 8. R data frame containing Psalm 6 from texts 01 and 12

doc_id	text
01 *Ortulus anime* from 1530 (STC 13828.4)	oh lord rebuke me not in thy wrath neither chasten me in thy anger but deal favourably with me oh lord for full sore broken am i heal me lord for my bones are all toshaken my soul trembles sore but lord how long turn thee lord and deliver my soul save me for thy mercies sake for they verily that are in this deadly anguish cannot think upon thee in their helly pains who may praise thee i am weary with sighing i shall water my bed every night with my tears so that it shall swim in them my face is wrinkled and dried up with care and anger my enemies have made it full thin with trouble avoid from me ye workers of wickedness for the lord has heard my complaints poured out with weepings the lord has heard my deep desire the lord has received my petition all my enemies shall be shamed and stunned they shall be put to flight and confounded suddenly

[19] For this reason, in many practical applications frequency-term vectors are normalised to the unit length 1, so that the similarity is calculated between normalised vectors and, consequently, the results are not biased by the magnitude of vectors. Normalisation changes the magnitude of vectors, not their direction. The calculations of cosine similarity performed on the psalm texts from (1) were conducted on vectors which were not normalised to the unit length 1.

12 1539_Coverdale's Psalter trans. from Latin of Campensis (first printed in 1535) (STC 2372.6)	oh lord chasten me not after the wrath that thou have taken against me though i be worthy of it neither punish me according to the displeasure that i have provoked thee unto but rather have mercy upon me oh lord for i am sick and have more need of thee to be my physician for all my whole body is disquieted and my mind is much more vexed but in the mean season oh lord when will thou at the last consider my wretchedness oh lord i beseech thee return to the kindness that thou were wont to have and deliver my soul from evil and restore me to my old health not regarding my sins so much as thy infinite goodness for how can a dead man think upon thy name or how can they which are drowned in hell show a thankful heart for the benefits that they have received of thee among them which are living i am weary with sorrowful mourning every night have i washed my bed and watered my couch with my tears my eye is darkened by reason of overmuch sorrow the white of it is waxed dim for fear of this great multitude of my enemies which couyte to destroy me go from me all ye that intend me evil for the lord is appleased with my sorrowful and careful complaint the lord has heard my prayer right well and accepted my desire let them be ashamed and sore vexed that owe me evil will let them be driven backward and suddenly brought to confusion

The psalm data from Table 8 were subsequently used to perform the cosine similarity calculations with the help of the R *quanteda* package (Benoit et al., 2018). Table A in the Appendix shows the results for the 20 versions of Psalm 6 from (1). As can be noticed, the range of the cosine similarity results falls between 0.734 and 1.0. The lowest scores (0.734 and 0.735) identify the level of similarity between the least similar texts, i.e. 1539 edition of Coverdale's Psalter translated from the Latin of Johannes Campensis (text 12) and texts (01, 02, 04, 05). The highest score recorded is 1.00 (which is obviously the result of the similarity between a text and itself, as well as the value of similarity between texts 01 and 02, i.e. *Ortulus anime* from 1530 and George Joye's English Psalter from 1530, which clearly appear to be the same George Joye's text published separately). Also, it can be noticed that the compared psalm versions seem to form a few groups: one group of highly similar texts comprising texts 01, 02, 04, 05 (similarity range between texts in this group 0.996-1.00); another one with texts 07, 08, 09 (0.993–0.998) and 18, 19, 20 (0.994–0.998); texts 10, 15, 17 (identical texts with 1.00 scores) and closest to them texts 06, 11 (similarity 0.977); and, finally, similar texts 14, 16 (0.996). In some cases, the results make it possible to easily identify the textual affinities between the compared texts. In the case of texts 01, 04, 05 the text of Psalm 6 contained in them is clearly based on the text of Joye's 1530 Psalter, i.e. text 02. The texts of Psalm 6 in primers 15, 17 are taken from 10 (Coverdale's Great Bible), and not surprisingly show a lot of similarity to Coverdale's earlier complete Bible from 1535 (06) as well as Taverner's Bible from 1539 (11). The remaining translations of Psalm 6 that were used for comparison show the following levels of similarity: text 12 is the least similar to any of the remaining versions of Psalm 6. This should not come as

a surprise as it originates from Coverdale's 1539 paraphrase translation based on the Latin of Campensis (unlike any other versions). This is evidenced by the fact that the text of this psalm in Table 8 above diverges quite significantly from other translations with the similarity scores ranging between 0.733 and 0.799.[20] Text 13 (from Coverdale's Psalter translated from the Vulgate) is closest to texts 06, 10, 11, 17, and 15, with scores between 0.917 and 0.930. Finally, text 03 (Joye's Psalter from 1534) shows closest affinity to texts 07, 14, 16, scoring between 0.902 and 0.905.

2.2 Calculating cosine similarity using word-level n-grams

As discussed above, cosine similarity calculations were performed on texts represented as term-frequency vectors, where the terms were all the words from the corpus of Psalm 6 texts. In this section I want to introduce an important modification in the representation of the compared texts and instead of words employ word-level n-grams as features of texts in term-frequency vectors. The concept of n-grams was first introduced in Shannon (1948) in a paper devoted to Information Theory, where an n-gram analysis of written English was undertaken. Vinson et al. (2016) note that the resurgence of n-gram models in language analysis came about in the mid 1970s and 1980s, following their successful use in speech recognition systems. As defined by Vinson et al. (2016, p. 910), an n-gram is a sequence of *n* items from a given text. In the case of n = 1, we are dealing with a 1-gram (unigram); if n = 2, a 2-gram (or bigram) is created; if n = 3, we have a trigram, and so on. N-grams are typically[21] classified into two categories: character-based, i.e. when individual letters are used to build an n-gram and word-based, i.e. when words are used in the construction of an n-gram (Mohan et al., 2010, p. 2).

It can be thus stated that all the cosine calculations performed above used unigrams for text representation. Importantly, however, Russel and Norvig (2021) note that n-gram models (where n > 1) maintain word-order information. Consequently, the comparison between texts represented as n-grams should now be sensitive to word order and should, therefore, result in an increased sensitivity of the method. Table 9 shows the representation of Psalm 6 from text 01, where the text is converted into a sequence of bigrams. The bigrams were extracted from texts using Tesserae's text handling functions described above.

[20] Space limitations make it impossible to demonstrate the text of all the psalms compared here but the text of Psalm 6 from Joye's *Ortulus anime* from 1530 can serve as a good illustration of the range of differences between Coverdale's paraphrase from 1539 (text 12) and all the other texts compared.

[21] See Sidorov (2019, p. 14) for a discussion of different types of n-grams whose elements are formed from part of speech tags (POS tags) or grammatical tags.

Table 9. Psalm 6 from text 01 represented as a sequence of bigrams

doc_id	text
01_Ortulus anime_1530	oh-lord lord-rebuke rebuke-me me-not not-in in-thy thy-wrath wrath-neither neither-chasten chasten-me me-in in-thy thy-anger anger-but but-deal deal-favourably favourably-with with-me me-oh oh-lord lord-for for-full full-sore sore-broken broken-am am-i i-heal heal-me me-lord lord-for for-my my-bones bones-are are-all all-toshaken toshaken-my my-soul soul-trembles trembles-sore sore-but but-lord lord-how how-long long-turn turn-thee thee-lord lord-and and-deliver deliver-my my-soul soul-save save-me me-for for-thy thy-mercies mercies-sake sake-for for-they they-verily verily-that that-are are-in in-this this-deadly deadly-anguish anguish-cannot cannot-think think-upon upon-thee thee-in in-their their-helly helly-pains pains-who who-may may-praise praise-thee thee-i i-am am-weary weary-with with-sighing sighing-i i-shall shall-water water-my my-bed bed-every every-night night-with with-my my-tears tears-so so-that that-it it-shall shall-swim swim-in in-them them-my my-face face-is is-wrinkled wrinkled-and and-dried dried-up up-with with-care care-and and-anger anger-my my-enemies enemies-have have-made made-it it-full full-thin thin-with with-trouble trouble-avoid avoid-from from-me me-ye ye-workers workers-of of-wickedness wickedness-for for-the the-lord lord-has has-heard heard-my my-complaints complaints-poured poured-out out-with with-weepings weepings-the the-lord lord-has has-heard heard-my my-deep deep-desire desire-the the-lord lord-has has-received received-my my-petition petition-all all-my my-enemies enemies-shall shall-be be-shamed shamed-and and-stunned stunned-they they-shall shall-be be-put put-to to-flight flight-and and-confounded confounded-suddenly

Each text was subsequently turned into a term-frequency vector (with terms corresponding to bigrams) and cosine similarities were calculated using the R *quanteda* package (Benoit et al., 2018). The results of the calculations are presented in Table B in the Appendix.

Even a cursory look at Table B reveals important differences between the two sets of cosine similarity results in Tables A and B in the Appendix. First of all, notice a completely different range of the obtained results. Cosine similarities calculated for texts represented as bigrams range from 0.176 (the similarity score between texts 03 and 12) and 1.00 (for the obvious case of text identity). Clearly, taking into consideration word ordering of a text produces more accurate results. Consider for example, the case of text 12, i.e. Coverdale's paraphrase from 1539. As could be observed in Table 8, this text, being a paraphrase,

diverges significantly from all the other texts compared here,[22] yet the similarity results (Appendix: Table A) calculated for unigrams ranged from 0.734 to 0.799. While it is true that these are the lowest scores obtained, it can be intuitively felt that the difference between complete identity (score 1.00) and (score 0.734), i.e. texts as different as texts 12 and 04 (or any other text) is simply not big enough. The situation is different in Table B: on a scale of possible values from 0 to 1, the cosine similarity value between texts 12 and 04 is 0.189, which much better captures our perception of the scale of textual difference involved. The lower score clearly results from factoring in word-order differences. Even more importantly perhaps, the sensitivity of the similarity calculations is much better with bigrams. For example, texts 10, 15, and 17 were measured to be identical in Table A, while the similarity score between texts 15 and 17 is 0.985 in Table B, as texts 10 and 15 are indeed identical and slightly different than 17. Another case in point is demonstrated by the changes in the relationships between the scores. As was observed above, text 03 (Joye's Psalter from 1534) shows highest levels of similarity with texts 07, 14, and 16. The results in Table A suggest that it is slightly closer to text 07 (score 0.905) than to either text 14 or 16 (0.902). However, the scores from Table B indicate that the opposite is true—the score between 03 and 07 is 0.449 and it is lower than the score between 03 and 14 (0.468) and between 03 and 16 (0.465). This is an important consequence of performing text comparison based on bigrams preserving word-order information.

Consequently, the conclusion to be drawn from the application of bigram representation of the examined texts is that, since the enhanced bigram representations capture the linguistic reality of the texts in a much better way, the scores in Table B represented a significant improvement as far as the textual affinities between the variant translations of Psalm 6 are concerned.

3. CONCLUSION

The paper discussed and assessed two ways of measuring similarity between historical texts, using 20 early Modern English versions of Psalm 6 found in publications printed between 1530 and 1557 as the source of data. First of all, an attempt was made to employ Tesserae (Coffee et al., 2012) for measuring the level of similarity between different versions of Psalm 6. It was demonstrated that Tesserae is of limited applicability for this purpose, mainly due to the way in which its scoring mechanism is designed to work. The second

[22] Due to space limitations, Table 8 provided the comparison between text 12 and text 01 only. Note, however, that the differences observed between these two texts are generally indicative (and illustrative) of the divide between text 12 (a paraphrase) and the remaining texts.

method discussed above involves the application of cosine similarity measurements to gauge the level of affinity between texts. As indicated above, the method has been successfully used in a number of recent studies where different texts were compared and their level of similarity measured by cosine similarity. The paper proposes a significant modification of the cosine similarity measurements and argues that the texts which are compared with this method should be represented as n-grams. This effectively introduces word-order information into the picture and results in a much-improved sensitivity of the method.

REFERENCES

Sources

[1539 edition of Coverdale's Psalter] = *A paraphrasis vpon all the Psalmes of Dauid, made by Iohannes Campensis, reader of the Hebrue lecture in the vniuersite of Louane, and translated out of Latine into Englysshe*. (1539). STC (2nd ed.) / 2372.6. London: Prynted in the house of Thomas Gybson.

[*Book of Common Prayer*, 1549] = *The booke of the common prayer and administracion of the sacramentes, and other rites and ceremonies of the Churche: after the vse of the Churche of England*. (1549). STC (2nd ed.) / 16270a. London: in officina Edouardi Whitchurche [and Nicholas Hill] Cum priuilegio ad imprimendum solum.

[*Book of Common Prayer*, 1552] = *The booke of common prayer and adminystracion of the sacramentes, and other rytes and ceremonies in the Churche of Englande*. (1552). STC (2nd ed.) / 16288. London: by in officina Edovardi whitchurche [sic].

[Caly's Primer] = [*The primer in English and Latin, after Salisburie vse, set out at length with manie praiers and goodly pictures, newly imprinted this present yeare*]. (1555). STC (2nd ed.) / 16062. London: In æibus Roberti Caly.

[Coverdale's Psalter, 1540] = *The Psalter or boke of Psalmes both in Latyn and Englyshe. wyth a kalender, & a table the more eassyer and lyghtlyer to fynde the psalmes contayned therin*. (1540). STC (2nd ed.) / 2368. London: Ricardus grafton excudebat. Cum priuilegio ad imprimendum solum.

[George Joye's English Psalter translated from the Latin text of Huldrych Zwingli] = *Dauids Psalter, diligently and faithfully tra[n]slated by George Ioye, with breif arguments before euery Psalme, declaringe the effecte therof*. (1534). STC (2nd ed.) / 2372. Antwerp: [Maryne Emperowr].

[George Joye's English Psalter translated from the Latin text of Martin Bucer] = *The Psalter of Dauid in Englishe purely a[n]d faithfully tra[n]slated aftir the texte of Feline: euery Psalme hauynge his argument before, declarynge brefly thentente [and] substance of the wholl Psalme*. (1530). STC (2nd ed.) / 2370. Antwerp: In the yeare of oure lorde 1530. the. 16. daye of Ianuary by me Francis foxe [i.e. Martin de Keyser].

[Godfray's primer] = *A primer in Englysshe with dyuers prayers & godly meditations. The contentes* […] (1535). Cum priuilegio regali. STC (2nd ed.) / 15988a. London: By Thomas Godfray.

[Henry VIII's primer] = *The primer, set foorth by the Kynges maiestie and his clergie, to be taught lerned, [and] read: and none other to be vsed throughout all his dominions*. (1545). STC (2nd ed.) / 16034.

London: VVithin the precinct of the late dissolued house of the gray Friers, by Richard Grafton printer to the Princes grace.

[*Manual of prayers*] = *The manual of prayers or the prymer in Englysh & Laten set out at length, whose contentes the reader by y[e] prologe next after the kale[n]der, shal sone perceaue, and there in shall se brefly the order of the whole boke. / Set forth by Ihon by Goddes grace, at the Kynges callyng, Byshoppe of Rochester at the comaun demente [sic] of the ryghte honorable lorde Thomas Crumwell, lorde priuie seale, vicegerent to the Kynges hyghnes*. (1539). STC (2nd ed.) / 16009. London: by me John Wayland in saynt Du[n]stones parysh at the signe of the blewe Garland next to the Temple bare.

[Marshall's primer] = *A prymer in Englyshe with certeyn prayers [et] godly meditations, very necessary for all people that vnderstonde not the Latyne tongue*. (1534). Cum priuilegio regali. STC (2nd ed.) / 15986. London: In Fletestrete by Johan Byddell. Dwellyng next to Flete Brydge at the signe of our Lady of pytye. for Wyllyam Marshall.

[*Ortulus anime*] = *Ortulus anime the garden of the soule* […] (1530). STC (2nd ed.) / 13828.4. Antwerp: by me Francis Foxe [i.e. M. de Keyser].

[Primer, 1552] = *The primer, and cathechisme, sette furthe by the kynges highnes and his clergie, to be taught, learned, and redde, of all his louing subiectes al other set apart corrected accordyng to the statute, made in the thirde and iiii. yere, of our souereigne Lordes the kynges maiestie reigne*. (1552). STC (2nd ed.) / 16057. London: by Richard Grafton, printer to the Kynges Maiestie.

[Psalms from Coverdale's first complete Bible] = *Biblia the Bible, that is, the holy Scripture of the Olde and New Testament, faithfully and truly translated out of Douche and Latyn in to Englishe*. (1535). STC (2nd ed.) / 2063. Cologne: Printed by E. Cervicornus and J. Soter?].

[Psalms from Coverdale's second complete Bible, known as the Great Bible] = *The Byble in Englyshe that is to saye the content of all the holy scrypture, both of ye olde and newe testament, truly translated after the veryte of the Hebrue and Greke textes, by ye dylygent studye of dyuerse excellent learned men, expert in the forsayde tonges*. (1539). STC (2nd ed.) / 2068. Paris: Prynted by [Francis Regnault, and in London by] Rychard Grafton [and] Edward Whitchurch. Cum priuilegio ad imprimendum solum.

[Psalms from Richard Taverner's Bible] = *The most sacred Bible, whiche is the Holy Scripture conteyning the Old and New Testament / translated into English, and newly recognised with great diligence after most faythful exemplars, by Rychard Taverner*. (1539). STC (2nd ed.) / 2067. London: Prynted at London in Fletestrete at the sygne of the Sonne by John Byddell, for Thomas Barthlet.

[Redman's primer] = [*This prymer in Englyshe and in Laten* …] (1537). STC (2nd ed.) / 15997. London: printed by R. Redman.

[Rouen primer] = [*This prymer in Englyshe and in Laten is newly tra[n]slatyd after the Laten texte*.] (1536). STC (2nd ed.) / 15993. Rouen: [by N. le Roux?].

[Wayland's primer, 1555] = [*The primer in Englishe (after the vse of Sarum)*]. (1555). STC (2nd ed.) / 16063. London: J. Wailande.

[Wayland's Primer, 1557] = *The prymer in Englishe and Latine after Salisbury vse: set out at length wyth many prayers and goodlye pyctures*. (1557). STC (2nd ed.) / 16080. London: By the assygnes of Ihon Wayland, forbyddynge all other to prynt thys or any other prymer.

Works cited

Buchler, M. (2016). *TRACER: Text reuse detection machine*. http://www.etrap.eu/research/tracer

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

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

Coffee, N., Koenig, J. P., Poornima, S., Forstall, C., Ossewaarde, R., & Jacobson, S. (2012). The tesserae project: Intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, *28*, 221–228.

Charzyńska-Wójcik, M. (2021). Familiarity and favour: Towards assessing psalm translations. *Linguistica Silesiana*, *42*, 43–77. https://doi.org/10.24425/linsi.2021.137231

Charzyńska-Wójcik, M., & Wójcik, J. (2022). Similarity measurements in tracing textual affinities. A study of psalm 129 in 16th-century devotional manuals. *Token*, *14*, 191–220.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook*. Cambridge University Press.

Forstall, C. W., & Scheirer, W. J. (2019). *Quantitative intertextuality*. Springer.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers.

Hordyjewicz, M. (2023). Scriptural content of the English medieval Book of Hours: Tracing textual traditions of nine lessons from the Book of Job. *Polish Journal of English Studies*, *9*(1), 82–96.

Huang, A. (2008). Similarity measures for text document clustering. *New Zealand Computer Science Research Student Conference*, *8*, 49–56.

Lis, K., & Wójcik, J. (2023). French and English texts of the *Laws of Oléron* – Assessing proximity between copies and editions by means of cosine similarity. *Bulletin of the John Rylands Library*, *99*(2), 103–126. https://manchesteruniversitypress.co.uk/9781526178503

Mohan, A., Baggili, I. M., & Rogers, M. K. (2010). Authorship attribution of SMS messages using an n-grams approach. *Proceedings of CERIAS Tech Report 2010-11*, 1–12. Center for Education and Research Information Assurance and Security Purdue University.

Olsen, M., & Horton, R. (2009). *PAIR: Pairwise alignment for intertextual relations*. https://code.google.com/archive/p/text-pairR Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach. Global edition*. Pearson Higher Ed.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(4), 623–656.

Sidorov, G. (2019). *Syntactic n-grams in computational linguistics*. Springer.

Vinson, D. W., Davis, J. K., Sindi, S., & Dale, R. (2016). Efficient N-gram analysis in R with Cmscu. *Behavior Research Methods*, *3*, 909–921. https://doi.org/10.3758/s13428-016-0766-5

Wójcik, J. (2021). Measuring internal spelling variation of an Early Modern English text. *Linguistica Silesiana*, *42*, 107–123. https://doi.org/10.24425/linsi.2021.137234

Wójcik, J. (2023). Cluster analysis in tracing textual dependencies – A case of psalm 6 in 16th-century English devotional manuals. *Digital Humanities Quarterly*, *17*(3), 1–16. http://www.digitalhumanities.org/dhq/vol/17/3/000694/000694.html

ON MEASURING PSALM SIMILARITY: A CASE FOR WORD-LEVEL N-GRAMS

Summary

The article offers a comparison between Tesserae (a text-reuse detection tool) and cosine similarity (used here as a measure of similarity between texts) and assesses their applicability to tracking textual affinities of different versions of historical texts on the basis of Early Modern English versions of Psalm 6 found in publications printed between 1530 and 1557. It is shown that cosine similarity is a better tool for the task of identifying and measuring the level of similarity between texts. At the same time, the article argues that cosine similarity measurements should be performed on texts represented as feature vectors consisting of n-grams.

Keywords: digital humanities; cosine similarity; n-grams; Psalm translations; R; Tesserae

O POTRZEBIE STOSOWANIA N-GRAMÓW WYRAZOWYCH W BADANIU PODOBIEŃSTWA PSALMÓW

Streszczenie

Niniejszy artykuł zawiera porównanie pomiędzy Tesserae (narzędziem do wykrywania ponownego użycia tekstu, tj. *text-reuse*), a podobieństwem kosinusowym (używanym tutaj jako miara podobieństwa między tekstami) i ocenia ich przydatność do analizy powiązań tekstowych pomiędzy różnymi wersjami tekstów historycznych, w oparciu o wczesno-nowoangielskie wersje Psalmu 6 opublikowane pomiędzy 1530 a 1557 rokiem. Dowiedziono, że podobieństwo kosinusowe jest lepszym narzędziem do identyfikacji i pomiaru poziomu podobieństwa między tekstami. Jednocześnie w artykule wykazano, że pomiary podobieństwa kosinusowego powinny być wykonywane na tekstach reprezentowanych jako wektory cech składające się z n-gramów wyrazowych.

Słowa kluczowe: humanistyka cyfrowa; podobieństwo kosinusowe; n-gramy; tłumaczenia psalmów; R; Tesserae

Author's bio

Jerzy Wójcik is Assistant Professor at the Institute of Linguistics, John Paul II Catholic University of Lublin, Poland. His current research interests include employing digital humanities tools in analysing early English texts.

Table A. Cosine similarity scores for texts represented as words (unigrams)

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
01	1.000	1.000	0.856	0.996	0.998	0.847	0.846	0.846	0.843	0.854	0.865	0.735	0.818	0.847	0.854	0.852	0.854	0.837	0.839	0.837
02	1.000	1.000	0.856	0.996	0.998	0.847	0.846	0.846	0.843	0.854	0.865	0.735	0.818	0.847	0.854	0.852	0.854	0.837	0.839	0.837
03	0.856	0.856	1.000	0.855	0.856	0.861	0.905	0.902	0.898	0.854	0.887	0.779	0.866	0.902	0.854	0.902	0.854	0.886	0.884	0.886
04	0.996	0.996	0.855	1.000	0.996	0.846	0.844	0.845	0.841	0.853	0.863	0.734	0.817	0.846	0.853	0.850	0.853	0.835	0.837	0.835
05	0.998	0.998	0.856	0.996	1.000	0.847	0.846	0.846	0.843	0.854	0.865	0.735	0.818	0.847	0.854	0.852	0.854	0.837	0.839	0.837
06	0.847	0.847	0.861	0.846	0.847	1.000	0.858	0.865	0.864	0.951	0.977	0.779	0.917	0.883	0.951	0.893	0.951	0.846	0.848	0.846
07	0.846	0.846	0.905	0.844	0.846	0.858	1.000	0.994	0.993	0.868	0.895	0.780	0.908	0.966	0.868	0.964	0.868	0.985	0.981	0.983
08	0.846	0.846	0.902	0.845	0.846	0.865	0.994	1.000	0.998	0.876	0.903	0.783	0.914	0.966	0.876	0.964	0.876	0.981	0.981	0.979
09	0.843	0.843	0.898	0.841	0.843	0.864	0.993	0.998	1.000	0.874	0.901	0.791	0.912	0.963	0.874	0.961	0.874	0.979	0.979	0.977
10	0.854	0.854	0.854	0.853	0.854	0.951	0.868	0.876	0.874	1.000	0.947	0.799	0.929	0.896	1.000	0.903	1.000	0.857	0.859	0.857
11	0.865	0.865	0.887	0.863	0.865	0.977	0.895	0.903	0.901	0.947	1.000	0.786	0.930	0.912	0.947	0.913	0.947	0.882	0.884	0.882
12	0.735	0.735	0.779	0.734	0.735	0.779	0.780	0.783	0.791	0.799	0.786	1.000	0.774	0.770	0.799	0.775	0.799	0.754	0.755	0.754
13	0.818	0.818	0.866	0.817	0.818	0.917	0.908	0.914	0.912	0.929	0.930	0.774	1.000	0.922	0.929	0.923	0.929	0.901	0.903	0.901
14	0.847	0.847	0.902	0.846	0.847	0.883	0.966	0.966	0.963	0.896	0.912	0.770	0.922	1.000	0.896	0.996	0.896	0.957	0.953	0.957
15	0.854	0.854	0.854	0.853	0.854	0.951	0.868	0.876	0.874	1.000	0.947	0.799	0.929	0.896	1.000	0.903	1.000	0.857	0.859	0.857
16	0.852	0.852	0.902	0.850	0.852	0.893	0.964	0.964	0.961	0.903	0.913	0.775	0.923	0.996	0.903	1.000	0.903	0.955	0.951	0.955
17	0.854	0.854	0.854	0.853	0.854	0.951	0.868	0.876	0.874	1.000	0.947	0.799	0.929	0.896	1.000	0.903	1.000	0.857	0.859	0.857
18	0.837	0.837	0.886	0.835	0.837	0.846	0.985	0.981	0.979	0.857	0.882	0.754	0.901	0.957	0.857	0.955	0.857	1.000	0.996	0.998
19	0.839	0.839	0.884	0.837	0.839	0.848	0.981	0.981	0.979	0.859	0.884	0.755	0.903	0.953	0.859	0.951	0.859	0.996	1.000	0.994
20	0.837	0.837	0.886	0.835	0.837	0.846	0.983	0.979	0.977	0.857	0.882	0.754	0.901	0.957	0.857	0.955	0.857	0.998	0.994	1.000

1.00	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.00

Table B. Cosine similarity scores calculated for texts represented as word-level bigrams

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
01	1.000	1.000	0.408	0.980	0.990	0.457	0.411	0.416	0.416	0.453	0.460	0.190	0.413	0.441	0.453	0.454	0.453	0.360	0.365	0.360
02	1.000	1.000	0.408	0.980	0.990	0.457	0.411	0.416	0.416	0.453	0.460	0.190	0.413	0.441	0.453	0.454	0.453	0.360	0.365	0.360
03	0.408	0.408	1.000	0.406	0.408	0.405	0.449	0.444	0.439	0.390	0.431	0.176	0.363	0.468	0.390	0.465	0.395	0.399	0.394	0.399
04	0.980	0.980	0.406	1.000	0.980	0.454	0.409	0.414	0.414	0.451	0.457	0.189	0.411	0.439	0.451	0.451	0.451	0.358	0.363	0.358
05	0.990	0.990	0.408	0.980	1.000	0.457	0.411	0.416	0.416	0.453	0.460	0.190	0.413	0.441	0.453	0.454	0.453	0.360	0.365	0.360
06	0.457	0.457	0.405	0.454	0.457	1.000	0.398	0.403	0.403	0.783	0.946	0.280	0.595	0.477	0.783	0.499	0.774	0.348	0.348	0.348
07	0.411	0.411	0.449	0.409	0.411	0.398	1.000	0.969	0.959	0.423	0.429	0.232	0.523	0.802	0.423	0.791	0.429	0.918	0.897	0.907
08	0.416	0.416	0.444	0.414	0.416	0.403	0.969	1.000	0.990	0.434	0.435	0.232	0.537	0.812	0.434	0.801	0.439	0.928	0.928	0.918
09	0.416	0.416	0.439	0.414	0.416	0.403	0.959	0.990	1.000	0.434	0.435	0.232	0.537	0.802	0.434	0.791	0.439	0.918	0.918	0.907
10	0.453	0.453	0.390	0.451	0.453	0.783	0.423	0.434	0.434	1.000	0.746	0.291	0.586	0.499	1.000	0.522	0.985	0.372	0.378	0.372
11	0.460	0.460	0.431	0.457	0.460	0.946	0.429	0.435	0.435	0.746	1.000	0.244	0.581	0.486	0.746	0.493	0.736	0.377	0.377	0.377
12	0.190	0.190	0.176	0.189	0.190	0.280	0.232	0.232	0.232	0.291	0.244	1.000	0.269	0.252	0.291	0.270	0.291	0.199	0.199	0.199
13	0.413	0.413	0.363	0.411	0.413	0.595	0.523	0.537	0.537	0.586	0.581	0.269	1.000	0.546	0.586	0.559	0.576	0.473	0.483	0.473
14	0.441	0.441	0.468	0.439	0.441	0.477	0.802	0.812	0.802	0.499	0.486	0.252	0.546	1.000	0.499	0.974	0.504	0.751	0.740	0.751
15	0.453	0.453	0.390	0.451	0.453	0.783	0.423	0.434	0.434	1.000	0.746	0.291	0.586	0.499	1.000	0.522	0.985	0.372	0.378	0.372
16	0.454	0.454	0.465	0.451	0.454	0.499	0.791	0.801	0.791	0.522	0.493	0.270	0.559	0.974	0.522	1.000	0.527	0.739	0.729	0.739
17	0.453	0.453	0.395	0.451	0.453	0.774	0.429	0.439	0.439	0.985	0.736	0.291	0.576	0.504	0.985	0.527	1.000	0.378	0.383	0.378
18	0.360	0.360	0.399	0.358	0.360	0.348	0.918	0.928	0.918	0.372	0.377	0.199	0.473	0.751	0.372	0.739	0.378	1.000	0.979	0.990
19	0.365	0.365	0.394	0.363	0.365	0.348	0.897	0.928	0.918	0.378	0.377	0.199	0.483	0.740	0.378	0.729	0.383	0.979	1.000	0.969
20	0.360	0.360	0.399	0.358	0.360	0.348	0.907	0.918	0.907	0.372	0.377	0.199	0.473	0.751	0.372	0.739	0.378	0.990	0.969	1.000