The Late War Against the Book of Mormon

Author(s): Benjamin L. McGuire
Source: *Interpreter: A Journal of Mormon Scripture*, Volume 7 (2013), pp. 323-355
Published by: The Interpreter Foundation

Abstract: No abstract available.

INTERPRETER

A Journal of Mormon Scripture

Volume 7 · 2013 · Pages 323-355

The Late War Against the Book of Mormon

Benjamin L. McGuire

Offprint Series

The goal of The Interpreter Foundation is to increase understanding of scripture through careful scholarly investigation and analysis of the insights provided by a wide range of ancillary disciplines, including language, history, archaeology, literature, culture, ethnohistory, art, geography, law, politics, philosophy, etc. Interpreter will also publish articles advocating the authenticity and historicity of LDS scripture and the Restoration, along with scholarly responses to critics of the LDS faith. We hope to illuminate, by study and faith, the eternal spiritual message of the scriptures—that Jesus is the Christ.

Although the Board fully supports the goals and teachings of the Church, Interpreter Foundation is an independent entity and is neither owned, controlled by nor affiliated with The Church of Jesus Christ of Latter-day Saints, or with Brigham Young University. All research and opinions provided are the sole responsibility of their respective authors, and should not be interpreted as the opinions of the Board, nor as official statements of LDS doctrine, belief or practice.

This journal is a weekly publication. Visit us at MormonInterpreter.com

The Late War Against the Book of Mormon

Benjamin L. McGuire

Recently, the Exmormon Foundation held their annual conference in Salt Lake City.[1] A presentation by Chris and Duane Johnson proposed a new statistical model for discussing authorship of the Book of Mormon.[2] The study attempts to connect the Book of Mormon to a text published in 1816: *The Late War Between the United States and Great Britain*.[3] The latter is a history of the war of 1812 deliberately written in a

1 The conference occurred between October 18th and October 20th, 2013.

2 The presentation was titled "How the Book of Mormon Destroyed Mormonism." It was presented on Saturday, October 19th, by Chris Johnson. The study was co-authored by Chris and Duane Johnson. The presentation can be viewed here: http://buggingmos.wordpress.com/2013/10/25/chris-johnson-how-the-book-of-mormon-destroyed-mormonism/

3 The full title of the work is given as: *The Late War Between the United States and Great Britain From June, 1812, to February 1815* (G.J. Hunt: New York, 1816). Rick Grunder provides this description of the various publications of this text: "This work went through at least sixteen editions or imprints 1816-19, all but two in 1819. All were published in New York City, under a total of ten different publishers' names. First "Published and sold for the author, by David Longworth," 1816… the book was then issued as *The Historical Reader*, Containing "*The Late War…* Altered and Adapted for the Use of Schools… ," etc., promoted particularly as a textbook (Samuel A. Burtus, 1817). There was no edition in 1818, but in 1819 there appeared no fewer than six separate editions or imprints under the original title and eight more editions or imprints as The Historical Reader. All fourteen of these 1819 publications called themselves the third edition. In five instances that year, both of the titles were published by the same parties, including the author himself. Furthermore, most of the 1819 editions (irrespective of title) seem to have had the same pagination (233 pp., with possible differences in plates and ads)." (Rick Grunder. *Mormon Parallels: A Bibliographic Source*. [Lafayette, New York: Rick Grunder—Books, 2008], p. 724.)

scriptural style. A traditional (non-statistical) comparison between this text and the Book of Mormon was apparently introduced by Rick Grunder in his 2008 bibliography *Mormon Parallels*. I will discuss only the statistical model presented by the Johnsons here.[4]

The history of author attribution is nearly as long as the history of reading and writing.[5] Within the field of literary studies, author attribution has developed into a field of scholarship, complete with its own history, its discussions on methodology, and even its own tightly contested difficult questions. This development has resulted in large reference volumes like the *Dictionary of Anonymous and Pseudonymous English Literature* (based on a work first published in 1882-3, and expanded twice to the current publication's 9 volumes, with the most recent volume added in 1962).[6] Scholarly discussion of author attribution continues, but is largely unknown within Mormon Studies, whose participants rarely come from a field of literary and textual criticism. This has lent a novel feel to those engaged in statistical approaches to the authorship of the Book of Mormon, even though few of these techniques are really new. Most of the participants seem unaware of the body of scholarly work that already exists which often supports or points out critical flaws in current assumptions. These can be

4 I may at some future point deal in a more detailed fashion with the thematic parallels presented by Grunder, along with his discussion of potential Hebraisms in the text.

5 "The scholarly study of attributions made its appearance at a period when literacy had ceased to be the monopoly of small cadres of specialist scribes and reading was for the first time practiced by a substantial public, ministered to by booksellers, stationers, scribal publishers, schoolmasters and grammarians. In the Western tradition such a public seems first to have consolidated itself in the fifth and fourth centuries BCE in Athens,… One important project was to distinguish the genuine works of Homer from other works that still at that time went under his name." (Harold Love, *Attributing Authorship: An Introduction* [Cambridge: Cambridge University Press, 2002], 14-15.

6 For further references and discussion see Love, *Attributing Authorship*, pp. 14-31.

found in past and current searches for influence and author attribution. Scholars of literary studies have been engaged in this endeavor for over two centuries, critiquing and evaluating the success of various methods as they have moved along. I detail some of this history in my review essay on Grunder's bibliography.[7]

This statistical modeling approach is in many ways simply an expansion of early attempts to investigate literary works using digital archives. However, after the first round of resulting scholarship it became apparent that electronic searches engaging in source attribution were plagued by many of the same flaws as non-electronic authorship attribution efforts. As an authority in the field, Harold Love, put it:

> When Byrne wrote, the accumulation of parallels was a labour-intensive business which depended on incessant reading of the works concerned. Today a phrase can be pursued almost instantaneously through the magnificent on-line LION archive, which covers all fields of English and American drama and of authored volumes of poetry up to 1900, and in many cases beyond, and is rapidly expanding into prose.... Now that the capacity to multiply parallels — most of which will be misleading — is almost unlimited, intelligent selectivity has never been more important.[8]

Love's point is that these digital archives create an almost unlimited supply of texts, in which searches can be performed easily for an almost unlimited number of phrases. When these searches are made, long lists of parallels are inevitably

7 I responded more generally to Grunder's entire work in a two part essay found here: http://www.mormoninterpreter.com/finding-parallels-some-cautions-and-criticisms-part-one/ and http://www.mormoninterpreter.com/finding-parallels-some-cautions-and-criticisms-part-two/

8 Love, *Attributing Authorship*:, 90.

discovered. However, parallels found in this manner — stripped of context and extracted from their sources — are, for the most part, illusory. This situation is similar to the way that visual look-alikes eventually pop up — somewhere, sometime — for virtually every public figure. We marvel at the uncanny resemblance between the two people, sometimes even theorizing familial relationships, forgetting about the automatic massive-scale search for similarities that occurs whenever someone becomes a public figure. When literary parallels are the result of intensive searches of massive databases, they cannot help us identify an author (or even influences on an author), nor can they help us understand the relationships between texts. This doesn't make these searches without value. Love points out where these electronic searches are most helpful:

> Here LION, Gutenberg and similar electronic archives come into their own, since as well as providing illusory parallels they also assist mightily in shooting down those which arise from the common parlance of the time. Once we have encountered an unusual expression in the writings of three or four different authors it ceases to have any value for attribution. What we are looking for is occurrences restricted to two sources only: one the anonymous work and the other a signed one! Even that might not be final: if the two authorial corpora are both large enough, chance alone would dictate that they should contain a few exclusive parallels.[9]

This may seem counter-intuitive. Love is not arguing that parallels are only valid if they are unique. Rather, within the massive electronic search model, illusory parallels are inevitable and must be treated with caution. Hence, parallels are more likely to be valid indicators of influence if they are unique.

9 Love, 91.

Parallels can be identified with electronic searches – but must then be evaluated in more traditional ways to determine if there is evidence for borrowing or influence. I provided a tentative methodology that I use for this purpose in the second part of my review of Grunder. The Johnsons' presentation seems to be based on the premise that numerical weighting of shared phrases between texts can overcome the weaknesses inherent in using an electronic search of a massive database to study the relationship between texts. I concur with Love in disagreeing with this premise. However, I believe there are additional problems with the Johnsons' methodology and conclusions. On their website, they candidly list several potential weaknesses of their study. What follows is a discussion of the Johnsons' approach, including additional problems with their database and algorithm.

Description of the Data and the Methodology

To introduce their methodology, Duane Johnson provided this "high level pseudocode" description of the score that he and his brother devised to measure similarity between two books:[10]

> For each book, create n-gram frequency counts:
> *Clean the Text*
>
> 1) remove non-alphabetic characters including newlines; keep spaces
>
> 2) normalize the case (e.g. lower case)
>
> 3) Slice the entire text into n-grams (i.e. "n" word sequences) e.g. in our study we chose 4-grams: "i nephi having been born of" becomes ["i nephi having been", "nephi having been born", "having been born of"] etc.

10 See http://askreality.com/hidden-in-plain-sight/#phrases downloaded on 10/26/13. I have adjusted the formatting, replacing the bullets with numerical references.

4) Sort the n-grams and count their frequencies

To Make a Baseline of n-gram Frequencies:

5) Randomly select a large sample of books (arbitrarily chosen number in our study: 5,000)

6) Add up all of the n-gram frequencies

7) Discard n-grams with frequency <= 4 (due to OCR errors, it's common to get many, many erroneous n-grams. Incidentally, discarding 4-grams with freq < 4 is why our highest matches have a score of 0.25)

To Get a Score for Each Book:

8) Find common n-grams between the Book of Mormon and each book

9) Eliminate from the list of common n-grams any n-grams found in the KJV or Douay-Reihms bibles [*sic*]

10) Take the inverse baseline frequency (e.g. if the phrase "having been born of" shows up 6 times in all of the books sampled for the baseline, then the inverse baseline frequency would be 1/6 = 0.167.

11) Sum all inverse baseline frequencies for common n-grams to get a "score"

12) Finally, divide the score by the total word count of each book (since larger books will, by random chance, have more matches). The result is a score that can be used to rank books by similarity.

To Rank All Books:

13) Use the score / wordcount

14) exclude small books whose signal-to-noise ratio is low ("small" was arbitrarily defined as 15,000 words in our study).

The authors do recognize some problems in the data and methodology. The first identified problem involves dealing

properly with variations in the lengths of the texts. The second is the problem of OCR errors. OCR stands for *optical character recognition* – the process by which a book page is captured as an image or a picture and converted into an electronic text for searching. Often, depending on the quality of the image (which in turn is affected by the condition of the book and other issues), the OCR process can create errors in the text. Sometimes the result is a recognizable (but different) word, but more often than not the resulting term is unrecognizable. The third issue is the confusion associated with the inclusion or exclusion of biblical texts. I will discuss this particular issue in greater detail below. To resolve the first, they excluded shorter texts. To resolve the second, they arbitrarily removed all of the word sequences that occurred fewer than four times. To resolve the third, all of the four-word sequences that could be constructed from the KJV or Douay-Rheims Bibles were excised from the data set. (This is not insubstantial as my own use of the KJV results in just under 680,000 different four-word phrases occurring in just that volume alone.)[11]

The data I am using for my analysis comes in part from Duane Johnson's blog (see fn. 11). He does not recognize the historical problems associated with electronic searches for authorship attribution. But it is clear from statements in the blog that he believes that this study has uncovered a textual reliance of the Book of Mormon on the work by Hunt:

> Using a "Uniform Match Score" (based on a size-independent matching scale), Hunt's *The Late*

11 My own work, which I will explain later, gathers just under 680,000 four-word phrases from the King James Version. All of these phrases were simply eliminated from the data set. I have not yet parsed the Douay-Rheims Bible at this time, and given the greater significance of the KJV for this discussion, it did not seem necessary. The primary necessity of excluding it seems to stem from the early date of included sources, starting in 1500, since it was published prior to the KJV. It seems reasonable though, that between the two editions of the Bible, close to a million four-word phrases were eliminated from consideration.

> *War* transmitted textual influence to *The Book of Nullification* is highest (0.37), followed by The Book of Mormon (0.24), and finally *Chronicles of Eri* (0.08).… all of which were significantly higher than the baseline scores, indicating textual transmission, or common influence.

According to this post, the textual reliance is either direct (from Hunt to the Book of Mormon) or based on a common source (i.e., a genetic connection of some sort is claimed to exist). However, Chris Johnson clarifies this in the comments:

> After a few tests it was clear that the Book of Mormon was a product of its culture, and could not have been made before 1822 since it relied on too many phrases found only in an 1822 Koran. It also could not have been written prior to 1816 since it relies on Hunt's *The Late War*.

Chris Johnson's conclusion is much less ambiguous than Duane Johnson's. In examining this claim, I will be providing my own analysis using a different set of tools.

My own work with texts and textual locutions began nearly a decade ago. I first wrote on this topic in an Internet forum, and my comments were eventually picked up and relocated.[12] I have some limited abilities to compare larger sets – performing logical operations on these sets of locutions,[13] but, in general, my tools are a bit simpler than those described above. I have an automated process – an algorithm – that takes a text,

12 http://solomonspalding.com/SRP/parallels.htm

13 My tools allow me to combine sets, to extract only elements common to two sets, and to subtract one set from another (removing all of the common sets). Using these tools, I could in theory build up a baseline data set similar to that used by Chris Johnson, although it would potentially take several weeks of constant computation, as they note, and a huge data repository. My tools were not designed with this functionality in mind.

isolates the numbers, changes capitals to lower cases, and strips out all punctuation. This process is generally referred to as *normalization* and corresponds to the first stage of the text cleaning above.[14] At this point, a text can be sorted in multiple ways. It can either be broken up into various-sized locutions (the n-grams used above) or sorted on frequency. Sorting on frequency provides details about the size of the vocabulary (in terms of unique words), and so on. Breaking up the text into locutions of a certain size creates a list of all of the phrases found in a text of a certain length.[15] This new set of locutions can then be treated much like a normalized text. Its entries can also be sorted by frequency. With no repetition, a text can have almost as many unique phrases as it has words.[16] The Book of Mormon is a good example of a text with a great deal of repetition (think: "and it came to pass"), and this can be seen in the larger gap between the number of total words in a text and the number of unique four-word phrases.

We can take this list of four-word phrases (along with their frequencies, if we choose) and compare them to similar lists from other books. The study in question proposed creating a baseline database – a combination of these frequency lists from a series of books chosen at random from a specific date range and matching criteria. This would create a very large list of phrases, along with total frequencies (the total number of times a phrase is used through the entire set of works). This overall frequency then became the basis for a weighting value for each phrase. My tools do not include this baseline database or the

14 http://en.wikipedia.org/wiki/Text_normalization

15 Often a marker is reinserted into the text (a punctuation mark) so that the new phrase can be seen as a single word for comparison purposes. So the phrase: "I Nephi, having been born of goodly parents becomes "I-Nephi-having-been" : "Nephi-having-been-born" : "having-been-born-of" : "been-born-of-goodly", and so on. Johnson's blog post displays this feature in his data.

16 When broken into four-word phrases, it would in fact have three phrases fewer than the total number of words in the text.

weighting it produces. My own analysis will not refer to a set of baseline data.

What follows, then, is a discussion of some of the inherent flaws in this basic methodology described by Johnson that are revealed through my own analysis and expectations.

Flaw 1. Preparation of the Texts

The first major issue occurs in the texts themselves. In preparing my notes, I used an existing digital copy of the Book of Mormon in my possession that I had cleaned up in a way similar to this method – by normalizing the text as well as removing material not strictly associated with the text or with its authorship. In the Johnsons' video presentation, an awareness of this need is discussed. Hunt's volume contained a lengthy appendix, which of necessity needed to be removed – it was described as creating noise.[17] This appendix seems to have been an 18-page addition including the full text of three treaties made by the U.S. government.

In the blog post, we are provided with a list of the four-word locutions that they used to produce their weighted connection between the Book of Mormon and Hunt's *The Late War*. The list is provided in ascending order of weights. While the earliest entries then have the least impact on the overall score of the connection, some of those entries are obviously problematic. For example, the text that was used to value the Book of Mormon included a copyright statement.

The copyright statement reads as follows (I have omitted the part in the middle that was written by Joseph Smith):[18]

17 The noise can come from additional instances of phrases used in the text itself. But it also creates new phrases when truncated elements are mashed together. I did not remove these pages.

18 The part written by Joseph Smith is a description of the text made at the time the application was made in June of 1829. This was hand copied from a proof-sheet of the title page of the Book of Mormon. While this text is closely associated with the Book of Mormon, it was never claimed to be a part of the

> Northern District of New York, to wit:
>
> BE IT REMEMBERED, That on the eleventh day of June, in the fifty-third year of the Independence of the United States of America, A. D. 1829, JOSEPH SMITH, JUN. of the said District, hath deposited in this office the title of a Book, the right whereof he claims as author in the words following, to wit: …
>
> In conformity to the act of Congress of the United States, entitled, "An act for the encouragement of learning, by securing the copies of Maps, Charts, and Books, to the authors and proprietors of such copies, during the times therein mentioned;" and also the act entitled, "An act supplementary to an act, entitled, 'An act for the encouragement of learning, by the securing copies of Maps, Charts, and Books, to the authors and proprietors of such copies, during the times therein mentioned,' and extending the benefits thereof to the arts of designing, engraving, and etching historical and other prints."

This is seen in the list of parallels provided. In fact, of the 549 distinct four-word locutions given in the blog and shared between the two texts, 75 of them (13.7%)[19] come from this

---

translation of the text, and so its potential value in ascertaining authorship is likely to be limited. Further, it is likely that the similarities between the summary on the title page and the text itself are derivative of the text of the Book of Mormon. For these reasons, I have generally excluded the entire title page with its summary from my past assessments.

19 Listed alphabetically by the first word of each four word set: act-entitled-an-act, act-for-the-encouragement, act-supplementary-to-an, an-act-entitled-an, an-act-for-the, an-act-supplementary-to, and-books-to-the, and-etching-historical-and, and-extending-the-benefits, and-proprietors-of-such, arts-of-designing-engraving, authors-and-proprietors-of, be-it-remembered-that, benefits-thereof-to-the, books-to-the-authors, by-securing-the-copies, charts-and-books-to, conformity-to-the-act,

copyright statement. This may have simply been an oversight. Unlike much of the text (except for the appendix, that they had already excluded), the copyright statement was not authored as part of the Book of Mormon, and it has a recognizable history.

The copyright statement comes from the copyright application form, a preprinted document in which the applicant had to fill in the blanks. The original application is known.[20] Only part of the copyright statement is original to Joseph Smith, and those parts were produced in 1829 when the application was filed. The statement in the Book of Mormon simply duplicates this application (as was generally required). This use of a form may explain why it duplicates in such great quantity the material from Hunt's volume (which was also copyrighted in New York and used an apparently identical or nearly identical pre-printed copyright application form.) It also explains why parts appear in so many other volumes

---

copies-during-the-times, copies-of-maps-charts, deposited-in-this-office, designing-engraving-and-etching, during-the-times-therein, encouragement-of-learning-by, engraving-and-etching-historical, entitled-an-act-for, entitled-an-act-supplementary, etching-historical-and-other, extending-the-benefits-thereof, for-the-encouragement-of, he-claims-as-author, historical-and-other-prints, in-conformity-to-the, in-the-words-following, in-this-office-the, independence-of-the-united, it-remembered-that-on, language-of-the-people, learning-by-securing-the, maps-charts-and-books, mentioned-and-extending-the, of-designing-engraving-and, of-learning-by-securing, of-maps-charts-and, of-such-copies-during, of-the-independence-of, office-the-title-of, proprietors-of-such-copies, remembered-that-on-the, right-whereof-he-claims, securing-the-copies-of, such-copies-during-the, supplementary-to-an-act, the-arts-of-designing, the-authors-and-proprietors, the-benefits-thereof-to, the-copies-of-maps, the-encouragement-of-learning, the-independence-of-the, the-times-therein-mentioned, the-united-states-of, the-words, following-to, therein-mentioned-and-also, therein-mentioned-and-extending, thereof-to-the-arts, this-office-the-title, times-therein-mentioned-and, to-an-act-entitled, to-the-act-of, to-the-arts-of, to-the-authors-and, united-states-of-america, whereof-he-claims-as, words-following-to-wit, year-of-the-independence

20 For the full text of the original copyright application, see Nathaniel Hinckley Wadsworth, "Copyright Laws and the 1830 Book of Mormon." *BYU Studies* 45/3 (2006), p. 97.

(as indicated by the low weights)[21] – the copyright application quotes statements from the U.S. Constitution (from Article I, Section 8, Clause 8) and the Copyright Act of 1790. These statements (or portions of them) would appear in most works printed in the United States between 1790 and 1831. (In 1831 we had the first major update to the Copyright Law.)

Removing this text wouldn't impact the weight much (it only reduces it by a little more than a half of one percent) because of the frequency in other texts. But it does dramatically reduce the number of parallels presented.

Additionally, there is the problem of the texts as they are. Most of the archived material that is searchable is produced by scanning the books into an image format, after which OCR is used to convert the images into a searchable text format. Despite recent improvements in the technology, texts that have been produced retain significant problems. The text I used for Hunt's *The Late War*[22] had some of these issues. In various places, 'Gilbert' becomes '6ilbert', 'With' becomes '7vith', and 'account' becomes 'accouut'. Since it is the OCR software that makes these mistakes and since the same combination of letters which may be confusing in one book can also be confusing in another (there were fewer typefaces back then), OCR software often makes the same kinds of mistakes in different texts. To deal with this, the proposal above excludes phrases found less than four times across the entire studied body of works. This

21 From the chart on the blog, simple math can be performed to discover the frequency of occurrences in the baseline data set. The formula is 1/[the listed value]. So, for example, from my list in fn. 19, if we make this calculation for the first five items in that alphabetical list, we get these frequencies: *act-entitled-an-act* – 3,350, *act-for-the-encouragement* – 360, *act-supplementary-to-an* – 241, *an-act-entitled-an* – 2,279, and *an-act-for-the* – 3,608. It is safe to suggest that a copyright statement with some degree of similarity occurs in a significant number of these texts.

22 For my analysis, I downloaded the text file at this address: http://www.archive.org/stream/latewarbetween_00hunt/latewarbetween_00hunt_djvu.txt

helps ferret out many of these errors.[23] The result runs right into Harold Love's suggestion about searching for parallels: "Once we have encountered an unusual expression in the writings of three or four different authors it ceases to have any value for attribution." In an effort to deal with bad data, has this collection effectively crippled their own weighting system by removing all of the instances that Love would find of real value? I believe that it has, although part of that explanation will also come up a little later. For this system to work in the long run, it would need texts that had been checked and found to be free of error. This has already been done with popular texts that are still in print, like the Book of Mormon or the KJV. However, it is not so easily done with archived scanned images of less interesting and less read works. (It is certainly not a chore that we would look forward to doing with the 130,000 volumes or even the 5,000 volumes randomly selected for the baseline data.)

The impact of removing these phrases is to create a hole in the text where the problematic word exists. By removing the four-word phrases that include the error (and there would generally be four phrases removed if there were an error),[24] it is quite likely that there is little impact on the baseline data. If a phrase is popular, it will remain popular in other works. However, the risk isn't in the removal of the errors, it's in the removal of legitimate phrases that are relatively unique.[25]

---

23 It also reduces the size of the data accumulated and the times required to process and search the data compilations. (I am fairly confident that this wasn't the intention, it was just a beneficial side effect.)

24 The phrases will be the phrases with the error in each of the positions X-2-3-4, 1-X-3-4, 1-2-X-4, and 1-2-3-X.

25 There are several potential ways to correct this that would not be too computationally intensive. A separate database could be maintained of all of the phrases removed for a lack of frequency, and this separate database could be matched up against the text in question. The matches (which would in theory be a relatively small number if most of the removed examples are errors) could then be examined individually for significance)..

Flaw 2: Length of Texts

A second major issue comes up with regard to the length of source texts. While the word count is referenced in the final score (generally with respect to the text in question),[26] this application seems to ignore much of what makes text length (or word count) interesting to us. Two useful features when dealing with locutions (or n-grams) are the size of the vocabulary (the number of unique words) and the overall length of the text in words. Both of these factors can influence the degree to which the texts are similar. And these are somewhat related figures. Shorter texts generally have a smaller vocabulary, while larger texts correspondingly have a larger vocabulary.[27] My lengths are likely to be a little different from those given by the blog site – due in part to minor differences in the process of cleaning the texts for use, and because I potentially use different sources for both texts. Given the size of the two texts, this discrepancy probably has a small impact on the outcomes of my examination.

The Book of Mormon text used in my apparatus was 269,551 words long with a unique vocabulary of 5,638 words (compared with the text of 271,240 words used in the Johnson study). *The Great War* was 56,632 words long (compare this to 55,378 words in the blog study – a difference most likely due to the inclusion of the appendix material) with a unique vocabulary of 5749 words. Significantly, the Book of Mormon text, while being nearly five times longer, has a vocabulary of similar size. And the shared vocabulary amounts to roughly forty percent of the respective vocabularies (specifically, they have a shared vocabulary of 2,281 words). My experience is that

26 In the comments, Duane Johnson points out: "When we say 'Score / WC' in the table, we mean 'Score divided by wordcount' which is the same slope you see in the graph."

27 One of the reasons why word count studies can work with shorter texts is that they are far more interested in the common words rather than the unusual words that make up the rare phrases that the Johnsons are looking for.

the vocabulary size of *The Late War* is consistent with books of similar length, while the Book of Mormon has an unusually small vocabulary. When we calculate the number of unique four-word locutions for each text, we can see the difference in repetition. The Book of Mormon contains 202,830 unique four-word locutions compared with *The Late War* containing 51,221.[28] Why is this interesting to us? If we follow the weighted matches used by the blog, there are 549 shared four word locutions common to both texts. This means that of all the possible phrases found in *The Late War*, only 1.07% of them make it into the Book of Mormon. And within the Book of Mormon, of the potential 200,000+ unique phrases, only 0.27% could be derived from *The Late War*. This is not a high number. This ratio drops substantially when we back out the 75 parallels taken from the copyright application (with 474 parallels it becomes 0.93% and 0.23% respectively).

This sort of ratio (the size of the footprint relative to the size of the text) doesn't come out in the calculations used. One of their supporting examples was provided in the blog:

> Surprisingly, the Uniform Match Score between The Book of Mormon and *The Late War* (scoring 0.24) was more significantly correlated than *Pride and Prejudice* (1813) and its most influential book *The Officer's Daughter* (1810), scoring 0.20. This indicates that Jane Austen's work was less influenced by her literary culture than The Book of Mormon.

I took copies of these two works (due to the better OCR, I used a version of *The Officer's Daughter* published in its original four volumes and combined them). I used the much cleaner text

28 For those interested, that means that the Book of Mormon has about 25% repetition at the level of four word phrases, while *The Late War* has only about 10%.

of *Pride and Prejudice* from Project Gutenberg.[29] *The Officer's Daughter* had a total word count of 140,245 with a vocabulary of 11,308 (some of this is undoubtedly due to OCR errors), while *Pride* had a word count of 122,880 with a vocabulary of 6,323. When I compared these two texts in a non-weighted comparison, it resulted in 1,934 common four-word phrases (conservative, due to the OCR errors in *The Officer's Daughter*). Having then backed out the parallels from the KJV we end up with 1,677 shared phrases. This results in a ratio in *Pride and Prejudice* of 1.4%.[30] This result is more than five times the overlap between the Book of Mormon and *The Late War*. Of the 6,323 words used in *Pride*, 3,996 of them are also found in *The Officer's Daughter* (63%).

In other words, the 'Uniform Match Score' (a term coined by the Johnsons) focuses very narrowly on one aspect of the data that is tightly controlled. It seems to have very little to do with the actual density of the overlap in the texts. Later in the comments to the blog entry, Duane Johnson offers this:

> Certain baseline data such as the false positive rate of our tools are still lacking. For example it is difficult to answer: "How often will our algorithm turn up the wrong books?" We don't know, so we wish to test our tools on as many books as possible, especially a) mystery texts where influence or authorship is unknown b) books with known influences, so that we can determine accuracy and c) books that are translated from another culture, time, language or place so that we can see how distantly connected a real Urantia or Koran text might look.

29 You can see these texts here https://archive.org/search.php?query=officer%27s%20daughter%20AND%20mediatype%3Atexts and here http://www.gutenberg.org/ebooks/1342

30 *Pride and Prejudice* contains 119,224 unique four-word phrases.

Here, with *Pride and Prejudice*, we have a text where they suggest there is a weight that is incongruous with the text that was "its most influential book." Rather than seeing this as evidence for an obvious flaw in their 'Uniform Match Score,' we instead get the conclusion that Jane Austen was simply less influenced by her environment than was Joseph Smith.[31] Given the suspect nature of the weighting system, I am unconvinced that there is actually any influence between these two books.

Without considering the size of the texts, any sense of relative proportion is lost. Harold Love pointed out that we are likely to find some degree of coincidental overlap between any two texts of sufficient size. This is a relatively small footprint (textually) – finding only 474 parallels in more than 200,000 opportunities. It is much smaller than the connection between *Pride and Prejudice* and *The Officer's Daughter*.

Flaw 3: Issues with the Biblical Text

In the discussion on method above, there is an attempt to sort out the influence of the biblical text. This was done by removing the four-word locutions that paralleled both the KJV and the Douay-Rheims translations of the Bible. Given the date of the two texts being closely examined, I only included the KJV in my testing. I did not exclude additional four-word sets equivalent to those in the Douay-Rheims.[32] For some background details, my text of the KJV is 791,539 words long.

31 There is some irony here in the degree to which Jane Austen was entirely separate from her environment. Those four-word phrases which might be entirely unique to Austen (the phrases that could hint at the degree to which Austen was independent of the literary culture in which she wrote) would be excluded by this study – both as a potential source (in that the frequency might not be high enough to include) and in the results (with no overlap at all, it would never come up in comparison). We get a conclusion that really cannot be supported by the data collected.

32 With its earlier publication date, the Douay-Rheims Bible would have a greater impact on earlier texts used in the baseline data. Hunt's volume was patterned on the KJV, and the Book of Mormon much more closely resembles the

It contains a vocabulary of 12,574 words. And it has 679,612 unique four-word locutions.[33]

I generated a comparison between this text of the King James and both *The Late War* and the Book of Mormon. The results showed an overlap with *The Late War* of 2,341 common four word locutions. The overlap with the Book of Mormon was significantly larger, at 25,020 locutions. This means that roughly 4.57% of *The Late War* duplicates material from the KJV, contrasted with 12.33% of the Book of Mormon duplicating phrases from the KJV. In both cases, these statistics trivialize the less than one percent overlap between the two books in question presented on the blog.

There are many potential reasons for excluding the KJV and Douay-Rheims phrases from consideration. I expect that including those phrases certainly skewed the baseline data. If I compare *The Late War* and The Book of Mormon using my texts without excluding the KJV data (that is, if I include all of the four-word locutions in my results), I end up with 1,478 shared phrases.[34] Of these shared phrases, a majority (57.3%) are also in common with the KJV. This leaves, at best, 631 shared four-word phrases between the Book of Mormon and *The Late War* independent of the KJV.

---

language of the KJV than it does the Douay-Rheims. For these reasons — and to keep the discussion as simple as possible — I only worked with the KJV.

33 Like my text of the Book of Mormon, this text is relatively free of OCR errors. I note that the repetition in the KJV is between the other two texts, at 15%.

34 This figure includes all of the four-word phrases used in both the Book of Mormon and *The Late War.* This figure is significantly different from the 549 weighted phrases used by the Johnsons to score the relationship. My figure includes low frequency phrases (including potentially OCR errors). Due to differences in the cleaning process, there are some additional variations. I did not include the copyright statement in the Book of Mormon (so the phrases exclusive to the copyright statement are not in this list) and I did not strip out an end material from Hunt's volume. The larger collection of phrases is useful because it provides a picture of the total ratio of textual material in common between the two books.

Removing this data also hides something that ought to have been obvious to us. The biblical text creates language in the environment (or represents that language) in an incredible density. When *The Late War* attempts to duplicate this language, we get an exact match 4% of the time. The Book of Mormon uses this language 12% of the time. It is only in removing these kinds of statistics that we get the sense of how the method is working: Without this comparison, what is otherwise a trivial overlap between two texts is magnified.

Flaw 4: Problems with the Weighting of the Phrases

There are several issues with the weighting system. The first, Chris Johnson describes remarkably well in his presentation. Here are the comments explaining this idea from the blog:

> if you find the two-word phrase "Millennium Falcon" in a book, and another two-word phrase, "it is" in a book, the former should matter a lot more than the latter. Why? Because almost every book in the world contains the 2-gram (bigram) "it is" but only a select few have "Millennium Falcon". So, what does a "weighted" value look like? It's just the inverse of the baseline frequency, i.e. 1.0/baseline-frequency. Using the example above: if "it is" occurs 5,847,361 in a sample of 5,000 pre-1830 books (which it does in our baseline sample) then the "weighted value" of the match is 1.0/5,847,361 or 0.000000171. Let's say "Millennium Falcon," on the other hand, occurs only one time in all of our sampled pre-1830 literature. Then, it would have a score of 1.0/1.0 = 1.0. So finding a "Millennium Falcon" match between the Book of Mormon and another book would be more than 5 million times more important.

Consider this challenge with respect to the biblical text. We know that the text of the KJV played a large role in the text of the

Book of Mormon. (This is seen by the large language footprint we find using these four word locutions.) However, the sheer frequency of the phrases from the Bible in the environment make this weighting approach problematic. A large number of collectively common phrases – all coming from the same ultimate source – might have virtually no impact on the weighted score if their frequencies in the baseline data were high enough.

Clearly this occurs in the case of the copyright statement. There we have a portion of the text that is not original to the Book of Mormon. Once we see it for what it is, we can track it – both to its immediate source (the copyright application) and then to its more distant sources (the pre-printed form, the legislative acts of the federal government that serve as its sources, and so on). What is interesting is how this interacts with the electronic search. This is one part of the Book of Mormon for which we can produce a genealogy for the text. It's also a part of the text that, because of its existence in the environment, doesn't trigger significant movement on the weighted scale. The collective initial weight[35] of these 75 phrases was roughly 0.33. The weight of a single phrase with a frequency of four across the entire baseline data was 0.25. All 75 of these phrases had less weight than two examples from the other end of the spectrum. So, on the one side, influence — if it is widespread, even if it comes from an identifiable source is considered negligible by this method. This is true of the copyright statement. It would also be true for the most part of the biblical text.

This fits right in line with Harold Love's assessment. Finding the phrase in more than a couple of sources (in our electronic search) means that each individual source is unlikely to be the cause of the influence. That connection becomes illusionary. Likewise, there is zero possibility that the copyright statement

35 Calculated by the sum of the inverse of the frequencies.

in Hunt's work could have been the cause of the copyright statement in the Book of Mormon.[36]

There is another corollary: Love's assessment of electronic sources didn't talk about frequencies of the phrases themselves across a corpus of work, but rather the number of sources in which a phrase occurred. The Book of Mormon uses the phrase "it came to pass" 1,353 times. If it were the only text to use this phrase, the baseline value for it would still be .000739. If that phrase occurred in only one other work, instead of being potentially highly significant (as Love suggests) it would be completely trivial in this weighting system. While this method tracks an overall frequency of a phrase within the collective pool of phrases used across an entire body of literature, it does not provide us with one very important detail, namely, how many works (or authors) use that phrase (independent of the frequency).

The next problem we have is with the sense of actual rarity. If, as Love argues, multiple instances are truly problematic, then our goal isn't to try to create a random sampling that is uniform when compared to the larger body of literature; we want to find a sampling that is most likely to give up the bad parallels in a frequency large enough to control mis-valuing the phrases. In creating a range of texts that extends from 1500 to 1830, with no geographical limitations, we tend to dilute the texts significantly. That is, even with 5,000 texts, if we had an even distribution (and I recognize that we don't), we would see a rather limited number of texts coming from an appropriate place and time. The distribution would have been far better had it been limited to a period around (both before and after) the publication of the Book of Mormon and from a much closer geographic perspective. It may not be that coincidental that

36 It's also true that part of that statement was caused (with absolutely certainty) by the existence of the federal copyright act of 1790. This method could not point us to that connection.

the closer matches occurred in those texts written closer to the publication of the Book of Mormon than those farthest away.[37] The dilution of the baseline data may enhance the value of these texts. How can we demonstrate this?

We can, as Harold Love suggested, use an existing database to function as a negative check. To do this, I selected a few of the highest scoring examples (those that have the minimal four occurrences across the selected set of texts). The texts were selected over the interval of 1500 to 1830. To duplicate this, I will use Google Books and perform a string search for identical text across that same interval. This won't give me a frequency of occurrences within an individual source text, but it will indicate (through the number of hits) how many sources the phrase occurs in – and in doing this there is a minimal boundary for a frequency.[38] Because the list is in ascending order based on score, I start from the bottom and work my way toward the top and search for the last ten items in the list.

1. your-women-and-your: 1 hit[39]
2. year-that-the-people: 2 hits
3. year-on-the-tenth: 14 hits
4. women-and-your-children: 1 hit
5. with-his-army-against: 29 hits
6. will-hearken-unto-him: 1 hit

37 I note in passing here that the KJV comes from a much earlier period of time. We don't suppose that the extreme overlap between the two is due simply to common language in the environment. Part of this is that the KJV was the most published work in the time period leading up to (and following) the publishing of the Book of Mormon.

38 There may be some duplication in the hits due to multiple editions of a single work.

39 The general search looks like this: https://www.google.com/search?q=-%22your+women+and+your%22&biw=1467&bih=608&sa=X&ei=MgtoUvSgFsHyyAGR_YCABA&ved=0CCMQpwUoBA&source=lnt&tbs=cdr%3A1%2Ccd_min%3A1%2F1%2F1500%2Ccd_max%3A1%2F1%2F1830&tbm=bks . It is created by using quotes to designate an exact phrase, then using the search tools feature to indicate a custom date range between 1/1/1500 and 1/1/1830.

7. will-give-unto-you: 3 hits
8. wickedness-which-had-been: 1 hit
9. which-he-gave-unto: 13 hits
10. were-upon-the-waters: 1 hit

These may be typical — or not. But, looking at these, three of them see a marked reduction in value. And while this may not be typical of the entire set, if it is, the impact would likely move Hunt's source down the value list. There are clearly some phrases which are rarer than others, and they may be useful. However, the selection of texts seems problematic in this regard. If this selection process takes a phrase where we can find dozens of examples elsewhere and produces only four occurrences (just enough to keep it from being eliminated but not so many that the parallel isn't simply removed), then there is clearly a problem with the process. And this valuation process could create a cumulative impact on the data. Either the number of texts in the base data is insufficient or the selection criterion needs to be re-tuned.

Part of this issue is in the assumptions that seem to be brought to the question. The desire is to identify a text which may have most influenced the text of the Book of Mormon, but to create the baseline of language you don't simply stop with the publication date of the Book of Mormon. If the Book of Mormon is a piece of nineteenth-century literature, it is both a product of, and a contributor to that language of its environment. We might opt to test the significance of earlier books against this baseline data, but unfortunately the data itself is not robust.

When we create a random sampling for statistical use, we do so on the assumption that our random sample will correlate well with the larger population. However, the sample size of 5,000 is far too small (and no work was done to verify that this random sampling was in line with the larger population). Given the nature of the problem, though, and the desire to reduce the

impact of phrases common in the environment, there doesn't seem to be a need for a truly random sample. Instead we should hand pick those texts that are most likely to share the same language – those texts that come from a closer geographic location and a closer time frame. We should try to reduce the impact of common phrases as much as possible so that those that are really unusual can stand out appropriately. We can do this by providing books with a content of history and war (and even theology) and by using travelogues.

There is another aspect to this, however. The copyright parallels (all 75 of them) are clearly the only part of the Book of Mormon for which we can point to an exact genealogy of texts. We know the textual history of this bit. We know it isn't original to the Book of Mormon, we know which sources were used, and so on. And yet if this was all that came up in the comparison, this weighting would immediately disqualify these parallels as irrelevant. There would be no reason to take a second look and discover what any one of us can see quite easily. In this regard the weighting system fails on both ends. It inappropriately overvalues some elements, and it inappropriately undervalues others. While I can suggest ways in which to accommodate for overvaluing some elements, I am not sure such an easy corrective measure can be taken to adjust for undervaluing other elements.

Finally, when we look at the list of weighted elements, we notice that many of them have little weight or value. The first 111 entries have the same value as a single later entry with a frequency of four occurrences. We can be fairly confident in these cases that the parallel is likely more environmental than direct. If we toss out all of the phrases where there are more than fifty occurrences in the baseline data, it would mean losing 225 (including the 75 from the copyright statement). This brings the overall footprint of Hunt's text in the Book of Mormon down to an underwhelming 0.16%. The real reason

for keeping them in the long list doesn't seem to be much about the mathematical impact of these common phrases (which is virtually non-existent) but rather the psychological impact of having a large list of parallels.

Flaw 5: Textual Context

The final issue is over the challenge of context. When we take texts and reduce them to these strings, we eliminate context. We rip out punctuation. Our four-word phrases cross natural textual lines. Without a more nuanced parsing, this is the only possible outcome. But it doesn't help us understand the relationship between texts. Because I quoted it verbatim earlier, the copyright statement makes a terrific example. Here are a few lines from it:

> In conformity to the act of Congress of the United States, entitled, "An act for the encouragement of learning, by securing the copies of Maps, Charts, and Books, to the authors and proprietors of such copies, during the times therein mentioned;" and also the act entitled, "An act supplementary to an act, entitled, 'An act for the encouragement of learning

The blog identified a couple of noteworthy parallels in this short text:

> entitled-an-act-for : time-therein-mentioned-and : therein-mentioned-and-also : act-entitled-an-act : entitled-an-act-supplementary : act-entitled-an-act : entitled-an-act-for …

Each of these four-word phrases crosses a textual boundary. They move from the immediate statement to a quotation of another text. This movement is lost. These phrases cross sentences and paragraphs. They string words together that don't belong together except in the sense of an n-gram – a

computational model based on removing the markers of these divisions from the text. The relationships that can sometimes be seen in these parallels don't exist for us as readers (or as writers). These aren't phrases that occur for us (or in our environment) because they don't actually exist as phrases (as locutionary acts). These are naturally rarer – because they are created entirely by coincidental circumstance and not by design of any author. And, in using them in a way that weights rarity more heavily, we tend to emphasize a feature of the language that doesn't exist except in the computational representation of word strings that no longer correlate to real writing or to real speech. These fragments, strung together, cannot provide us indicators to the language usage in comparison because they don't represent language usage at all.

Some Additional Observations

I had some additional concerns. Duane and Chris Johnson tend to use very ambiguous language to describe the relationships between texts. Some of it is incorrect, some of it is contradictory. Consider the following statements from the blog post:

> Our results point to *The First Book of S* (1809) influencing the creation of *The Late War*… Our preliminary analysis is showing that *The Late War* likely inspired the creation of quite a few books between 1820-1830, …
>
> Using a "Uniform Match Score" (based on a size-independent matching scale), Hunt's *The Late War* transmitted textual influence to *The Book of Nullification* highest (0.37), followed by The Book of Mormon (0.24), and to a lesser extent *Chronicles of Eri* (0.08). The influence from *The First Book of Napoleon* on Hunt's *The Late War* was 0.06, all of which were

> significantly higher than the baseline scores, indicating textual transmission, or common influence.
>
> We were interested in uncovering any books besides the Bible that may have played an influential role on the 1830 Book of Mormon.
>
> This indicates that Jane Austen's work was less influenced by her literary culture than The Book of Mormon.
>
> After a few tests it was clear that the Book of Mormon was a product of its culture, and could not have been made before 1822 since it relied on too many phrases found only in an 1822 Koran. It also could not have been written prior to 1816 since it relies on Hunt's *The Late War*. Also the *Chronicles of Eri* was more distant than the Book of Mormon to its most common ancestor, while *The Book of Nullification* was more connected to its ancestors than the Book of Mormon. I also tried tracing Solomon Spalding's *Manuscript Found*, and *The First Book of Napoleon*, but couldn't find a close source of textual transmission, meaning they were more out of place, and less explainable than the Book of Mormon.

These paragraphs present a confusing image. What does "influence" actually mean? Is it synonymous with reliance? By influence do the Johnsons mean that the Book of Mormon would not exist in the form it is today without the earlier book having been published? One thing that stands out to me is the statement about Jane Austen's *Pride and Prejudice*. On the basis of their weighting system they connect this book to a relatively unknown work from 1810: *The Officer's Daughter*. Jane Austen

is considered one of the most influential novelists of the modern era. This would be the first time that this connection has been offered, and it's being offered on the basis of an electronic search engine!

There is no evidence that this work was ever read by Jane Austen. In fact, just this year, Cambridge University Press released The Cambridge Companion to '*Pride and Prejudice*', in which we get details about the text, its narrative and characters, its philosophy, its composition and publication, even its historical background and literary context. Nowhere in that volume will we find a reference to Miss Walsh's *The Officer's Daughter*. For an author who wasn't very "influenced by her literary culture," an awful lot has been written about that culture and its influence. We actually know a great deal about Jane Austen and her literary influences. Part of this is due to the fact that literary scholars and historians have been discussing and detailing her achievements in terms of the relationship she had with prior literature since the mid-twentieth century (really beginning with the work of F. W. Bradbrook and Jocelyn Harris). For Austen, this interaction was often very deliberate – we know this not just from her books, but from the many letters that she wrote which detailed her own reading and re-reading. She tells us who her favorite authors were and why. And this is why we might be a bit startled to find out how this book, which she apparently never read, was in fact the most significant influence on her own writing.

Clearly something is off in this analysis. Yes, it's possible, that through any of a number of ways, this text was the most influential to her writing. Perhaps her best friend read it and shared the details over and over with her until it became ingrained in her subconscious. It's possible. It's just not very likely. Similarly, when we get to Hunt's book, there is this emphasis on the nature of the book as a school text. Actually,

we don't have any record of it being used in schools. There were, Rick Grunder points out:

> at least sixteen editions or imprints 1816-19, all but two in 1819. All were published in New York City, under a total of ten different publishers' names.… There was no edition in 1818, but in 1819 there appeared no fewer than six separate editions or imprints under the original title and eight more editions or imprints as The Historical Reader. All fourteen of these 1819 publications called themselves the third edition. In five instances that year, both of the titles were published by the same parties, including the author himself. Furthermore, most of the 1819 editions (irrespective of title) seem to have had the same pagination (233 pp., with possible differences in plates and ads).… A comparison of the Daniel D. Smith 1819 edition of *The Late War* (considered in this entry) and another in my possession under the same title, "Printed & Published by G. J. Hunt. Corner of Varick and Vandam streets," 1819, reveals what appears to be the identical typesetting (including page 41 mis-numbered, "31") except for the different publishers' names on the title pages, and their own ads filling their respective final page of the book. G[ilbert]. J. Hunt's ads at the end of his edition… provide some suggestion of his business and personality. Since the author appears to have been affiliated with both printing and a bookstore, I wonder if he printed these books himself (or had them printed), but then went around town soliciting orders from other booksellers or publishers, promising their own names on the title pages as publishers (as opposed to their appearing merely as distributors). In such a possible situation, we might be less surprised when we

> notice that after 1819, no further editions of this *wildly* published textbook appeared.[40]

The author appears to have marketed the book to booksellers (and not to schools) in an attempt to get this volume into the public view. There is no indication that it was ever actually used in a school as a school text. This is further suggested by the fact that after his wild marketing scheme ended in 1819, the book was never re-published (or even reprinted). A great deal of inappropriate emphasis is placed on the book's own description of its purpose as a way of suggesting that it be used and this potential connection to Joseph – that he likely encountered it in school as a "textbook used in the 1820s."

Finally, on the blog we notice the collection of works to which this is being compared (for which we have the composite score presented). It is obvious that these works could not have been included within the baseline data. There are at least eight different copies of Hunt's book, ranging from the highest (with an adjusted score of 4.2 to the twenty-third spot with an adjusted score of 2.3. That's a significant range. Given the shift, we have to ask: exactly which version was Joseph supposed to have come into contact with? The first edition (assigned the highest score) was not (apparently) marketed for school children. That comes with the second edition in 1817, and in the many different copies published in 1819. Yet, from the list of scored texts provided by Johnson, it is the 1816 edition which has the highest score. Of the other seven copies scored by Johnson, the second highest (a copy of the 1819 third edition) comes in with a weighted score reduced by 25%. Is this gap caused by OCR errors or is it due to textual differences?

From the same list we also have several versions of the Koran, ranking from number eight to number two hundred

40 Rick Grunder, *Mormon Parallels: A Bibliographic Source,* pp. 724-5.

and thirty. An explanation for the significance of the one version of the Koran was hinted at in this statement:

> since it [the Book of Mormon] relied on too many phrases found only in an 1822 Koran.

This isn't a claim of some sort of influence, or shared language caused by the environment. This is the claim that Joseph must have read this particular edition of the Koran (and not some other edition), and used it by incorporating it into his text of the Book of Mormon (along with the other imagined sources). This stretches credulity (although perhaps not as much as the claims about Jane Austen).

Conclusions

It isn't a particularly difficult feat to reconstruct the Book of Mormon using phrases found from many different sources. In the 1960s, Julia Kristeva coined the term *intertextuality* to describe this feature of all texts. They were, as she described them, a 'mosaic of quotations' all coming from other sources. Some of this is certainly due to textual influence and reliance. There is no doubt that the Book of Mormon owes a great deal of its contents to the King James text. But, as Harold Love points out, given a large enough body of literature, you can also find these phrases caused by coincidence. In the long run we note that there are some real similarities that can be found in the texts of these two books. But, most of these similarities are not discovered by creating a list of these four-word phrases – because these phrases are not themselves meaningful. Does this process attempt to reduce the significance of the Book of Mormon to a few hundred four-word phrases, stripped of punctuation and context? That seems to be the outcome. Hunt wanted to create a text that read like scripture as a marketing tool. In this way we get a lot of biblical sounding text. The Book of Mormon, on the other hand, doesn't just use biblical

language, it engages biblical issues – it asks questions about morality, about agency, about creation. It ponders the meaning of writing and reading. It describes religious experience.

At this point, this preliminary work of statisti cally mining electronic databases does not deal with Love's concerns or rehabilitate the practice. Perhaps future refinements will help. I do see uses for these kinds of approaches to the text. They can help us see where to start looking for real potential overlap. Substantial phrasing that does not occur commonly will encourage us to return to the text and evaluate it in a more traditional fashion. Once we do this, we may find a copyright statement with an identifiable textual history, Or we may discover that the parallels tell us absolutely nothing because they are most likely due to coincidence.

Special thanks to Bruce Schaalje for his criticism and suggestions.

Benjamin L. McGuire is a technologist in the field of healthcare in northern Michigan, where he lives with his wife and three children. He has special interest in the field of literary theory and its application to the Book of Mormon and early LDS literature. He has previously published with the Maxwell Institute.