Latent Semantic Analysis Applied to Authorship Questions in Textual Analysis

Caleb Z. Marshall[1] & Erin M. Buchanan[1]

[1] Missouri State University

Author Note

Caleb Z. Marshall is an undergraduate student in mathematics and psychology at Missouri State University. Erin M. Buchanan is an Associate Professor of Quantitative Psychology at Missouri State University.

Correspondence concerning this article should be addressed to Caleb Z. Marshall, 901 S. National Ave, Springfield, MO 65897. E-mail: Marshall628@live.missouristate.edu

Abstract

This study used a multi-dimensional text modeling technique, latent semantic analysis (LSA), to examine questions of authorship within the biblical book Isaiah. The Deutero-Isaiah hypothesis, which cites significant lexical and semantic differences within Isaiah as evidence for tripartite authorship, is well supported among biblical scholars. This quantitative textual analysis explored authorship and contextual semantics of Isaiah through LSA by examining the cosine vector relatedness between and across chapters and proposed authors. Because of the general semantic asymmetry across Isaiah, it is reasonable to conclude that a portion of Isaiah's semantic change is the result of multiple authorship. Further, our analysis helps demonstrate how statistically focused psycholinguistic work may be used to answer seemingly philosophical or subjective questions in other research fields.

Keywords: applied research, latent semantic analysis, semantics

Latent Semantic Analysis Applied to Authorship Questions in Textual Analysis

From a linguistic standpoint, perhaps nothing is as central to the function of language as an individual word's meaning (Jones, Willits, & Dennis, 2015). Yet meaning as a cognitive action presents a challenge to contemporary methods of empirical research within psychology, as semantics is a core base to understanding psychological phenomenon. While this study is focused specifically on computational linguistics, it intrinsically depends on related theories of memory and cognition. Within psychology, the conceptual understanding of a specified item or process is known as semantic memory. Cognitively, semantic memory is the individual's ability to abstract and store usable information from personal experience, or episodic memory. More generally, semantics is often considered our knowledge for facts and world knowledge (Tulving, 1972), and the semantic system stores both conceptual and propositional knowledge. For example, the semantic memory of *dog* would contain conceptual information about dogs (*has a tail*, *has fur*), which are built in a propositional network wherein links can be evaluated as true or false (Collins & Loftus, 1975; Collins & Quillian, 1969).

Newer models have incorporated associative or thematic information about concepts, such as *are friendly*; *are found in parks* (Lund & Burgess, 1996) and linguistic elements, such as part of speech (Jones & Mewhort, 2007). The challenge of studying semantic memory lies in its complexity. Rather than possessing the rote taxonomy of a dictionary or encyclopedia, semantic memory is surprisingly nuanced and flexible. Besides being deeply integrated with episodic memory, semantic memory also informs most other cognitive processes, such as abstract reasoning, decision making, language processing and perception. And, in most cognitive theories of semantic memory, individual concepts are interactive, meaning that item-by-item memories are conceptually interdependent upon one another.

Connectionist Models of Semantic Memory

The connectedness of semantic memory gave rise to varied computational models of semantic memory which translate semantic memory networks into mathematical information. The first of these models, designed by Collins and Quillian (1969) and Collins and Loftus (1975), showed great success at understanding the hierarchical structure and spreading activation in memory using the conceptual parts of memory as nodes, and the propositional parts of memory as connections between nodes. Propositions possess Boolean logic; that is, they can be either true or false. Mathematically, this property is utilized in several semantic memory models, such as the connectionist networks of Rumelhart and Todd (1993). Within Rumelhart networks, concepts and propositions are both presented as nodes in an interconnected model of semantic memory (McClelland & Rumelhart, 1989; McClelland, Rumelhart, & Hinton, 1986; Rogers & McClelland, 2006). These nodes are then joined by weights that determine relatedness to one another, which creates a connected architecture and hence, the connectionist name associated with these models.

These weighted connections have been crucial to understanding neural connections in memory (Moss & Tyler, 2000; O'Reilly, Munakata, Frank, & Hazy, 2012; Rogers et al., 2004). Models are built with input nodes that lead through a hidden, unseen layer of nodes to an output nodes. The input nodes are fed information by activating sets of nodes, which is processed through the hidden layer, and the output layer produces an answer. For example, a *dog* output might be found with an input of *tail*, *fur*, and *ears*. These models are meant to mimic learning, as different forms of adjustment on the weights are implemented to study changes as the model is trained (McClelland et al., 1986; Regier, 2005). The implementation of learning has distinct advantages over previous models. However, connectionist models are built around an expected environment rather than an extant, specified corpus of linguistic data (Jones et al., 2015).

Distributional Models: Language and Semantic Memory

As a medium for large-scale semantic memory observation and modeling, written language has shown to be quite usable. Written language, as a complex, emergent process of human cognition patterns well onto existing mathematical models. Lexical analysis, which focuses on count aspects of vocabulary, comparing observed frequency in a body of text (i.e. corpus) with theoretical distributions based on research assumptions, is an example of early statistics-based textual analysis (Kucera & Francis, 1967). Discourse analysis, conversely, uses grammatical algorithms to examine syntactic structure within a group of documents, or corpus (Guerin-Pace, 1998). In terms of psychological data, written language possesses extremely low-volatility and is often available in physical or electronic archives (Brysbaert & New, 2009). This ease of access and reliability encouraged the development of a computational linguistic approach to semantic memory modeling which focused on extant linguistic structures within each text as opposed to conceptual network constraints. In this study, we focused on one such area, distributional models.

Distributional models of semantic memory are based on a simple linguistic theory: words with similar meanings often co-occur in each section of text (Harris, 1981). This phenomenon is statistically advantageous because it allows the semantic architecture of an individual text to be determined from the co-occurrence frequency of its individual words. By nature, all distributional models of semantic memory are centered on word frequency distribution, which is itself a component of lexical analysis. However, a wealth of mathematical techniques have been developed to convert these small-world lexical effects into large, complex models of discourse analysis. These larger, mathematically complex models have proven to be quite accurate in modeling many cognitive phenomena, including semantic memory architecture (Cree & Armstrong, 2012; Rogers & McClelland, 2006).

Latent Semantic Analysis

For this study, we used a distributional modeling technique called Latent Semantic Analysis (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). Latent Semantic Analysis (LSA) is a method of analyzing multi-document collections or archives (text corpora) using matrix algebra, singular values decomposition, and trigonometric eigenvector calculations. Before discussing the mechanics of LSA, it is important to note that it is not the only distributional model capable of semantically modeling large text corpora; however, LSA excels at creating comprehensible models of static semantic memory within a given set of documents and has been applied to many areas of psychological study, from problem solving (Quesada, 2007) to knowledge acquisition (Landauer & Dumais, 1997). Pioneered by Landauer et al. (1998), LSA computes the contextual semantics of a given word based on the terms which co-occur with it, and by nature, those which do not. In computational linguistics, context refers to the general linguistic environment which surrounds a given unit of language (i.e. word, paragraph, document). Thus, contextual semantics is the referential meaning of a given word or phrase based on nearby co-occurring terms. As an example, take the party game charades. If someone said, *it runs and barks and likes to play*, you would immediately shout, *dog*. Contextual semantics functions in a similar manner: derive a word's meaning simply by examining the words surrounding it. This concept is central to all distributional models.

By understanding how concepts are related, we can also use LSA's quantification of multiple terms to build models of large semantic and thematic information in documents. This interpretation is accomplished using a document-by-term matrix, with rows corresponding to term frequency and columns representing user defined bodies of text within a corpus. Thus, the initial text matrix utilized by LSA is nothing more than a record of word frequency within the corpora. This matrix contains the raw linguistic data from which the contextual semantics of individual words will be derived. However, computationally, the early text matrices of LSA resemble nothing more than a simple frequentist table of word occurrence in a corpus (Landauer et al., 1998). Next, the original text matrix is manipulated to create a high-dimensional space. In this semantic space model, words are represented as non-linear points. Contextual semantic meaning, based on frequency, is modeled as intersecting vector values between these points. Following complex dimension reduction, the angles produced by the meeting of these vectors are then calculated with a simple cosine function, with larger cosines corresponding to greater semantic similarity and vice-versa (Günther, Dudschig, & Kaup, 2016). Overall, much of this process is similar or even identical to common statistical research methods in behavioral science (i.e. correlation), although cosine functions have a distinct advantage of representing multi-dimensional space, rather than linear relationships.

What characterizes LSA is its use of singular-value decomposition, an algebraic technique which reduces the size of a matrix while maintaining row-to-column congruence (Berry, Dumais, & O'Brien, 1995). Using eigendecomposition (a generalized means of matrix factorization), singular-value decomposition factors the original m x n text matrix M into three separate matrices. These are: U, a unitary, m x m matrix which models an orthonormal space of the semantic model; V, a unitary, n x n matrix which models a document space analogous to U; and Σ, a rectangular, diagonal matrix of singular values which intersects U and V (Jones et al., 2015). Thus, the original text matrix M can be represented as a product of its factorized matrices U, V and Σ. After singular value decomposition, the resulting factorized matrices are used to create a Euclidean, three-dimensional semantic space. Individual words are then represented as points in this lower dimensional space, which utilizes the diagonal singular values matrix S to relate the orthonormal word occurrence matrix U to the term-to-document frequency matrix $V*$. Thus, a word's orientation in this resulting semantic space is a geometric expression of its expected meaning versus its contextual semantic meaning. Moreover, semantic similarity can easily be computed based on the cosine of the vectors between word points which are expressions of the singular values contained in the Σ matrix.

Textual Analysis Using LSA

LSA's ability to transform high dimensional, complex text matrices into three-dimensional semantic spaces is the core of its usability. As a means of large data set manipulation, LSA is multifunctional, with applications from testing reading skill with greater precision in traditional read-aloud experiments (Magliano & Millis, 2003) to quantifying context-asymmetry and item comparison, as used by Foltz, Kintsch, and Landauer (1998) to measure document coherence. Foltz et al. (1998) compared the vectors created by LSA to predict participants' perception of document coherence, which is the overall similarity between separate bodies of text. While the specific purpose of their study was to predict and measure document-by-document coherence, they also demonstrated that participants' perception of overall document coherence is formed by interrelating their cognitive understanding of contextual semantics. The more semantic overlap two documents share, the greater likelihood that participants would perceive document-to-document coherence. Thus, by comparing the corresponding documents' eigenvectors produced by singular-value decomposition, LSA was demonstrated to be a reliable predictor for human judgments of document coherence. In theory, this finding occurred because the vectors produced by singular-value decomposition in LSA seem to approximate human understanding of contextual semantic meaning.

Practical Applications of LSA: The Isaiah Scrolls and Deutero-Isaiah

The application of LSA to textual analysis is tantalizing, especially with the extreme processing power and modeling flexibility which singular-value decomposition affords the computational linguistic researcher. The Foltz et al. (1998) research demonstrates LSA performs well in predicting human perceptions of textual coherence, as well as more recent studies into the application of LSA (Hofmann, 2001; Kulkarni, Apte, & Evangelopoulos, 2014; Landauer, 2002; Wang, Peng, & Liu, 2015). And, statistically, there is no difference in methodology between proactively applying LSA as a predictive measure and retroactively measuring the contextual semantic similarity of bodies of text in a corpus. The question then becomes: which already-established corpora would benefit from such a technique? Using vector comparison similar to Foltz et al. (1998), we retroactively measured the document-by-document contextual semantics of a pre-existing corpora: the transliteral English Translation of the Book of Isaiah. Surprisingly, many sections of the Hebrew Bible have proven especially difficult to date, organize, and even translate. The Book of Isaiah is one such challenge for Biblical Scholars, mainly because there are no surviving original scrolls which contain the Isaiah text in its entirety. The fragmentary history of the Isaiah scrolls raises serious questions of document coherence and authorship. These doubts are especially troubling since the text is traditionally presented as a unified, single-author work (Brettler, 2005).

Within Biblical Studies, these doubts coalesced around a theory of multiple authorship for the Isaiah scrolls known as the Deutero-Isaiah hypothesis. This theory posits that the Isaiah scrolls were the product of three separate authors, each of whom existed in a distinct time-period and geographic location. The Deutero-Isaiah hypothesis is quite popular among Biblical Scholars (Sharp & Baltzer, 2003). Disagreement exists, especially among traditional scholars (Coggins, 1998), as well as questions of term significance (Sargent, 2014), and the precise location of authorship (Goulder, 2004). However, earlier statistical analysis of the Isaiah scroll fragments has supported this theory of multiple-authorship (Pollatschek & Radday, 1981). Therefore, this study sought to explore the Deutero-Isaiah hypothesis using LSA as an objective measure of semantic and thematic (Maki & Buchanan, 2008) relations between chapters of proposed authors. Specific hypotheses are described below.

Method

Data Analysis

Each chapter of Isaiah was converted to plain text files and imported in to *R* using *textmatrix()* command found in the *tm* package (Feinerer & Hornik, 2017), excluding English stopwords, such as *the*, *an*, *but*, etc. The term by document matrices were then log transformed and weighted to control for text-size differences. These files were then processed into latent semantic spaces where then created using *lsa()* in the *lsa* package (Wild, 2015), which provided corpora specific eigenvector values corresponding to the contextual semantics of each chapter of Isaiah. Following LSA, these 66 latent semantic spaces were logged as transformed text matrices which were used to calculate chapter-to-chapter cosine values as the variable of interest. For each of these cosines, we also calculated chapter distance, defined as the subtraction of the chapter number (i.e. 1-2 is a distance of 1, while 1-50 is a distance of 49). Using the divisions advocated by the Deutero-Isaiah hypothesis (chapters 1-39, 40-55, 56-66), we also coded each chapter combination as within author or across authors. The following hypotheses were tested using this coding system:

Hypothesis 1. This hypothesis was used to show the applicability of LSA to understanding semantic spaces of literature. Cosine values between chapters within suggested author should be greater than zero, thus, indicating semantic space similarity across one author's writing. Cosine values across authors may be greater than zero, due to common thematic material across authors. This hypothesis will be tested with a single sample *t*-test.

Hypothesis 2. Given support for hypothesis one, we sought to use the cosine values to examine the Deutero-Isaiah hypothesis by examining how cosine values for within author chapter combinations should be different from across chapter combinations. This hypothesis will be tested with specific *a priori* pairwise independent *t*-tests rather than all possible combinations (i.e. Author 1 will be compared to Author 1-2 and Author 1-3, but not Author 2-3).

Hypothesis 3. Hypotheses one and two focused on differences between average semantic space relatedness for proposed author systems in Isaiah. Hypothesis 3, instead, examined how the semantic space for each chapter changed with chapter distance by correlating cosine values with chapter distance. We expected to find negative correlations between chapter distance and cosine as further chapters would be less semantically related to each other.

Each chapter-to-chapter combination was considered an independent value; however, because chapters do repeat across these pairs, we also examined using a multilevel model controlling for chapter number as a random factor, with no discernible differences. Therefore, the simpler t-test analyses are presented below.

Results

Hypothesis 1

For Hypothesis one, each cosine combination of within author and across author was compared against zero using a single sample t-test (two-tailed), and the results are presented in Table 1. We hypothesized that within author cosines would be greater than zero, as this result would imply a related set of chapters creating a semantic space. Across author cosines were hypothesized to be potentially greater than zero, which suggests common thematic material and possibly one authorship. This hypothesis was supported, as all average cosines were significantly greater than zero, as shown in Table 1. These values are significant even after controlling for Type I error using a Bonferroni correction (i.e. 05 / 6 = .008). Precise p values can be found by viewing and running the *R* markdown file at http://osf.io/jywa6. Cohen's d values and their non-central confidence intervals (Cumming, 2014; Kelley & Preacher, 2012; Smithson, 2001) were calculated for each t-test as additional evidence for each test. Together, these values indicate large effect sizes to support our hypothesis (Cohen, 1992).

Hypothesis 2

For Hypothesis two, we compared matching within author cosines to across author cosines to determine if there is support for different semantic spaces in the Deutero-Isaiah Hypothesis. We expected internal within author cosine values to be larger than across author cosine values, as this result would indicate more cohesive semantic spaces within each proposed author over separate author spaces. Table 2 includes the independent t-test and Cohen's d values for these comparisons. Author 1's internal cosine values were significantly larger than the across Author 1 comparisons (see Table 1 for means and standard deviations); however, the effect sizes and their confidence intervals indicate that this difference was likely significant due to sample size, as effects are small with ranges close to zero. In contrast, Authors 2 and 3 showed significantly larger internal cosine averages than across author cosine averages with large effect sizes and corresponding confidence intervals.

Hypothesis 3

Last, we examined how semantic space relatedness changed across chapters, herein called semantic drift. The correlation between chapter distance and cosine was calculated for each chapter pairing, and a negative correlation was expected. Table 3 indicates the t-values, correlations, and their 95% confidence intervals. Because of the differences in sample size, we examined the strength of the correlation as an indicator of interest. This hypothesis was partially supported, as the overall correlation of chapter distance and cosine was significant and negative, with a small to medium effect size. Within Author 2 showed the most semantic drift across the semantic space, followed by within Author 3, and then within Author 1. While the average cosines were significantly greater than zero from Hypothesis one, the across author correlations for Author 1 to 2 and 1 to 3 were found to be approximately zero. Interestingly, across Author 2 and 3, a small negative correlation appeared.

Discussion

Deutero-Isaiah

This study systematically examined the semantic architecture of Isaiah using Latent Semantic Analysis (Landauer & Dumais, 1997; Landauer et al., 1998). First, chapter-to-chapter cosines were calculated for relatedness, and we examined if they were statistically different from zero using single sample t-tests. This analysis provided a standardized measure for the semantic structures within Isaiah and a basis for further statistical modeling of the text, as these average cosine values were different from zero. Hypothesis two was a natural extension of this concept, comparing within-section cosines to cross-section cosines to determine group similarities within Isaiah. This result led to hypothesis three and the introduction of *semantic drift* across the entirety of Isaiah. Combined with the effect size measurements from previous experiments, quantifying semantic drift gives an incremental measurement of the semantic differences across Isaiah.

Based on the t-test results of hypothesis one, it can be concluded that within-author cosines are significantly interrelated to each other. This finding implies that each sub-section of Isaiah forms a thematic group. Also, examining effect size measurements shows that the first authored section of Isaiah demonstrates the least cohesive thematic cosine effect when compared to sections two and three, which demonstrated the largest effect sizes. As a base measure of relatedness, hypothesis one supported strong cohesion within sections two and three with comparatively weaker thematic cosine similarities in section one. The results of hypothesis two portrayed group relatedness among Isaiah's sub-sections. Significant differences were found across all between-group average cosine values. However, in examining effect sizes, we see that within-group cosines of section one yield smaller effects than within-group cosines of sections two and three. Moreover, in examining within-groups cosines of sections two and three against between-groups cosines with section one, we find the largest effects. Effect size presents strong evidence for thematic asymmetry between section one to sections two and three. This result is consistent with scholarly opinion regarding Isaiah, especially regarding the Deutero-Isaiah hypothesis.

While hypothesis two observed larger group effects in Isaiah, hypothesis three quantified incremental semantic changes across the entirety of Isaiah, which we termed *semantic drift*. Overall, there was a significant, small-moderate negative correlation between chapter location and cosine similarity. Moreover, sections two and three demonstrated the largest negative correlations, with section two being statistically significant. While significant, section one's smaller correlation coefficient coupled with the smaller thematic cohesion demonstrated in hypothesis one presents difficulty in interpreting the semantic drift across section one. This result might imply different authorships within section one or a single author with different thematic focuses mashed together. When compared to the more cohesive and significant author two (or even the non-significant but similarly sized author three), the difference in effect sizes is apparent. Practically, these results suggest a clear, directed narrative in author two and, to a lesser extent, author three's writing which is either less prominent in author one, or entirely non-existent.

Conclusion

In this manuscript, we have demonstrated the usefulness of LSA to hypotheses that are normally subject to only qualitative analyses, which contributes to the scientific literature by performing necessary replication and extension studies (Schmidt, 2009), while not exclusively relying on traditional null hypothesis testing criterion (Cumming, 2008, 2014). One obvious limitation to this study is the use of the English translation of Isaiah; however, the convergent results of our study along with Biblical scholars (Sharp & Baltzer, 2003) provides promising support for the use of LSA in understanding many types of text. Recent studies have shown that both business (Kulkarni et al., 2014) and internet applications (Wang et al., 2015) have benefited from using LSA as an analysis tool. This manuscript extends that literature, thus, allowing researchers multiple avenues to explore their hypotheses in both the qualitative and quantitative realm.

Moreover, this study demonstrates that statistical modeling of a complex text can fortify scholarly opinion within the humanities and other fields. This work suggests new avenues for replication, especially in fields where statistical modeling is less frequently utilized. In this study, we referenced the work of Ioannidis (2005) in order to interpret our inferential statistical findings, relying not only on traditional p-value results, but also on effect sizes between groups and a clear delineation of hypotheses. Our work reinforces the reliability of findings in traditionally less statistically driven research areas, such as religious studies and linguistics and fits in line with recent movements in statistical thinking (Wasserstein & Lazar, 2016). In summation, statistical methodology is widely applicable, both on the edge of scientific development, but also in new and exciting areas of study as a tool for replicating previous findings to reaffirm the theories and work of our fellow researchers.

References

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, *37*(4), 573–595. doi:10.1137/1037127

Brettler, M. Z. (2005). *How to read the bible*. Philidelphia: The Jewish Publication Society.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. doi:10.3758/BRM.41.4.977

Coggins, R. J. (1998). Do we still need Deutero-Isaiah? *Journal for the Study of the Old Testament*, *23*(81), 77–92. doi:10.1177/030908929802308106

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:10.1037//0033-2909.112.1.155

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. doi:10.1037/0033-295X.82.6.407

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247. doi:10.1016/S0022-5371(69)80069-1

Cree, G. S., & Armstrong, B. C. (2012). Computational models of semantic memory. In M. Spivey, K. McRae, & M. Joanisse (Eds.), *The cambridge handbook of psycholinguistics* (Vol. 4, pp. 259–282). Cambridge, MA: Cambridge University Press. doi:10.1017/CBO9781139029377.018

Cumming, G. (2008). Replication and p intervals. *Perspectives on Psychological Science*, *3*(4), 286–300. doi:10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. doi:10.1177/0956797613504966

Feinerer, I., & Hornik, K. (2017). Text mining package. Retrieved from http://tm.r-forge.r-project.org/

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, *25*(2-3), 285–307. doi:10.1080/01638539809545029

Goulder, M. (2004). Deutero-Isaiah of Jerusalem. *Journal for the Study of the Old Testament*, *28*(3), 350–362. doi:10.1177/030908920402800306

Guerin-Pace, F. (1998). Textual statistics. *Social Sciences*, *10*(1), 73–95.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, *69*(4), 626–653. doi:10.1080/17470218.2015.1038280

Harris, Z. S. (1981). Distributional structure. In *Papers on syntax* (Vol. 10, pp. 3–22). Dordrecht: Springer Netherlands. doi:10.1007/978-94-009-8467-7_1

Hofmann, T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, *42*, 177–196. doi:10.1023/A:1007617005950

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37. doi:10.1037/0033-295X.114.1.1

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J. T. Townsend & J. R. Busemeyer (Eds.), *The oxford handbook of computational and mathematical psychology* (pp. 232–254). Oxford University Press. doi:10.1093/oxfordhb/9780199957996.013.11

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–52. doi:10.1037/a0028086

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day English.*

Providence, RI: Brown University Press.

Kulkarni, S. S., Apte, U. M., & Evangelopoulos, N. E. (2014). The use of Latent Semantic Analysis in operations management research. *Decision Sciences*, *45*(5), 971–994. doi:10.1111/deci.12095

Landauer, T. K. (2002). Applications of Latent Semantic Analysis. In *24th annual meeting of the cognitive science society* (Vol. 24).

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. doi:10.1037//0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, *25*(2-3), 259–284. doi:10.1080/01638539809545028

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. doi:10.3758/BF03204766

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction*, *21*(3), 251–283. doi:10.1207/S1532690XCI2103_02

Maki, W. S., & Buchanan, E. M. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review*, *15*(3), 598–603. doi:10.3758/PBR.15.3.598

McClelland, J. L., & Rumelhart, D. E. (1989). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. Cambridge, MA: MIT Press.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. (1986). The appeal of parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations* (pp. 3–44). Cambridge, MA: MIT Press.

doi:10.1016/B978-1-4832-1446-7.50010-8

Moss, H., & Tyler, L. (2000). A progressive category-specific semantic deficit for non-living things. *Neuropsychologia*, *38*(1), 60–82. doi:10.1016/S0028-3932(99)00044-5

O'Reilly, R., Munakata, Y., Frank, M., & Hazy, T. (2012). *Computational cognitive neuroscience* (1st ed.). Wikibooks.

Pollatschek, M., & Radday, Y. (1981). Vocabulary richness and concentration in Hebrew biblical literature. *Bulletin of the Association for Literary and Linguistic Computing*, *8*(3), 217–231.

Quesada, J. (2007). Spaces for Problem Solving. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 117–131). Routledge. doi:10.4324/9780203936399.ch10

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*(6), 819–865. doi:10.1207/s15516709cog0000_31

Rogers, T. T., & McClelland, J. L. (2006). *Semantic cognition*. Cambridge, MA: MIT Press.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205–235. doi:10.1037/0033-295X.111.1.205

Rumelhart, D. E., & Todd, P. (1993). Learning and connectionist representations. In D. Meyer & S. Kornblum (Eds.), *Attention and performance xiv: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 3–30). Cambridge, MA: MIT Press.

Sargent, B. (2014). 'The coastlands wait for me, and for my arm they hope': The sea and eschatology in Deutero-Isaiah. *The Expository Times*, *126*(3), 122–130. doi:10.1177/0014524613499485

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100.

doi:10.1037/a0015108

Sharp, C. J., & Baltzer, K. (2003). Deutero-Isaiah: A Commentary on Isaiah 40-55. *Scottish Journal of Theology*, *56*(1), 101–130. doi:10.1017/S0336930603220182

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*(4), 605–632. doi:10.1177/00131640121971392

Tulving, E. (1972). Organization of memory. In *Episodic and semantic memory* (pp. 381–402). New York, NY: Academic Press.

Wang, J., Peng, J., & Liu, O. (2015). A classification approach for less popular webpages based on latent semantic analysis and rough set model. *Expert Systems with Applications*, *42*(1), 642–648. doi:10.1016/j.eswa.2014.08.013

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi:10.1080/00031305.2016.1154108

Wild, F. (2015). Latent semantic analysis package. Retrieved from https://cran.r-project.org/package=lsa

Table 1

Summary and t Statistics for Hypothesis 1

Comparison	M	SD	t	df	p	d	95%CI
Author 1	.30	.20	40.73	740	$< .001$	1.50	1.39 - 1.60
Author 2	.52	.23	25.17	119	$< .001$	2.30	1.95 - 2.64
Author 3	.57	.22	19.01	54	$< .001$	2.56	2.01 - 3.11
Author 1 - 2	.26	.15	43.50	623	$< .001$	1.74	1.62 - 1.87
Author 1 - 3	.25	.15	34.21	428	$< .001$	1.65	1.51 - 1.80
Author 2 - 3	.35	.16	28.43	175	$< .001$	2.14	1.87 - 2.41

Note. Average scores are the mean cosine value for each pair of chapters by author.

Table 2

t Statistics for Hypothesis 2

Comparison	t	df	p	d	$95\% CI$
Author 1 v Author 1 - 2	3.74	1363	$< .001$	0.20	0.10 - 0.31
Author 1 v Author 1 - 3	4.18	1168	$< .001$	0.25	0.13 - 0.37
Author 2 v Author 1 - 2	15.63	742	$< .001$	1.56	1.35 - 1.77
Author 2 v Author 2 - 3	7.76	294	$< .001$	0.92	0.67 - 1.16
Author 3 v Author 1 - 3	13.74	482	$< .001$	1.97	1.66 - 2.27
Author 3 v Author 2 - 3	8.25	229	$< .001$	1.27	0.95 - 1.60

Note. Average scores and standard deviations are presented in Table 1.

Table 3

Correlation and t Statistics for Hypothesis 3

Correlation	t	df	p	r	$95\% CI$
Overall	-10.51	2143	$<$ .001	-.22	-.26 - -.18
Author 1	-2.50	739	.013	-.09	-.16 - -.02
Author 2	-2.88	118	.005	-.26	-.42 - -.08
Author 3	-1.33	53	.191	-.18	-.42 - .09
Author 1 - 2	0.93	622	.354	.04	-.04 - .12
Author 1 - 3	0.23	427	.820	.01	-.08 - .11
Author 2 - 3	-2.15	174	.033	-.16	-.30 - -.01