Does God Speak in Threes? Applying Latent Semantic Analysis to Questions of Authorship in Isaiah

Caleb Z. Marshall

Missouri State University, Springfield, MO USA.

E-mail: Marshall628@live.missouristate.edu

Erin M. Buchanan

Missouri State University, Springfield, MO USA.

E-mail: erinbuchanan@missouristate.edu

Summary. We shall write an abstract that is about 120 words that covers the topic and results well.

Keywords: latent semantic analysis, religion

1. Introduction

The key question in computational linguistics is how do we theorize, test and manipulate language without relying on language to explain our findings? It is the classic circular argument of who came first: the chicken or the egg? The processes and rules of a given language or the understanding of these mechanics in a scientific environment? Of course, while research is principally an intellectual endeavor, it is impossible to deny that good research is both internally valid in its tests and measures, but also externally clever, capable of tackling abstruse problems in novel ways. It is this meeting of soundness and creativity that propels scientific discovery forward?but first comes the inevitable: the complex problem which evades preconceived measurements and techniques.

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 (henceforth: corpus) with theoretical distributions based on research assumptions (Guerin-Pace 1998), is an example of early statistics-based textual analysis. Discourse analysis, conversely, uses grammatical algorithms to examine syntactic structure within a single corpus (Guerin-Pace 1998).

For this study, we used a statistical modeling technique called latent semantic analysis (LSA for short). Pioneered by Landauer and Dumais, latent semantic analysis computes the contextual semantics of a given word based on the terms which co-occur with it, and by nature, those which do not (Landauer, Foltz and Laham 1998). LSA uses a document-by-term matrix to create a multi-dimensional semantic space which records term frequency per document. What separates LSA from other multi-dimensional models is its use of singular-values decomposition, an algebraic technique which reduces the size of a matrix while maintaining row-to-column congruence. After singular values decomposition, a new, three-dimensional Euclidean space results from the smaller, congruent matrix. Individual words are then represented as points in this lower dimensional space. Finally, semantic relationships are computed as vectors between word points.

Latent semantic analysis?s ability to transform high dimensional, complex spaces into three-dimensional models is the core of its usability. As a means of large dataset manipulation, latent semantic analysis is multifunctional, with applications from understanding human processing of conflicting discourses (Wolfe and Goldman 2003) to testing reading skill with greater precision in traditional read-aloud experiments (Magliano and Millis 2003). The application of context-asymmetry and item comparison, as used by Foltz, Kintsch and Landauer (1998) to measure document coherence, was of significant importance to this study.

Using similar methods as Foltz et al, we retroactively applied item comparison and context matching to a pre-existing corpora: the transliteral English Translation of the Book of Isaiah. However, rather than measuring emergent linguistic phenomena, we examined contextual trends and semantic patterns within the Isaiah corpora. This was to statistically test the Deutero-Isaiah hypothesis, which cites significant lexical and thematic differences within Isaiah as evidence for tripartite authorship. The Deutero-Isaiah hypothesis quite popular among Biblical Scholars (Goulder 2004; Kohl 2002). 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). While an unconventional application of latent semantic analysis, this study attempts to apply a demonstrated statistical tool to a decidedly non-statistical issue to foster inter-disciplinary collaboration and a continued excitement for applied mathematics.

2. Method

This quantitative exploration used latent semantic analysis to examine thematic change throughout the Book of Isaiah. As described previously, Latent semantic analysis represents words as high-dimensional vectors travelling across contrived lower-dimensional semantic space. Because of the Euclidean nature of the resulting latent, three-dimensional space, allows semantic similarity can be measured across corpora as the cosine of the angle between associate terms (Gunther, Dudschig and Kaup 2015).

Cohesive authorship of Isaiah was tested using the tertiary split advocated by the Deutero-Isaiah Hypothesis (Chapters 1-39; 40-55; 56-66 respectively). Following Latent Semantic Analysis, a simple correlation was calculated using chapter distance as the independent variable with larger cosine values corresponding to higher rates of relatedness. It is important here to make the distinction between overall cosine correlation, which refers to the relatedness of one entire corpora to another complete corpora, versus item-specific cosine correlation, which refers to the individual words? vectors and the resulting angles produced by their intersection. The former takes into account the matrices? whole cosine values, whereas the item-specific cosines refers to the comparison cosine values produced by term-to-term vector intersection. This experiment employed overall cosine correlation to test similarity between the three matrices, each of which corresponded to a hypothesized single author.

3. Results

Following LSA and cosine correlation, distance was negatively correlated with cosine values (r = -0.2213). Multivariate normality and linearity were tested for the chapter-to-cosine correlations and were normal. Heteroscedasticity is still being processed at this time. A three-way ANOVA was used to test distance-to-cosine correlation significance and determined that distance was an accurate predictor of cosine value (p ¡ .0001).

4. Discussion

This experiment demonstrated that chapter distance within Isaiah is an accurate predictor of thematic cosine strength. In conjunction with the general thematic asymmetry between each of the hypothesized sections of Isaiah observed by traditional Biblical scholarship, it is reasonable to conclude that a portion of Isaiah?s thematic change is the result of multiple authorship. This conclusion, while in keeping with modern scholarly opinion, also reflects the observed, homogenous linguistic change within the independent sections of Isaiah. Rather than possessing a cohesive thematic narrative, Isaiah expresses a richly variegated linguistic texture. This is more in keeping with a multiple authorship hypothesis as opposed to typical thematic cohesion across a single author?s work.

As a statistical foray into textual analysis, this experiment demonstrates that multi-dimensional modelling has diverse applications, one of which is analysis of authorship in ancient texts. Future studies utilizing Latent Semantic Analysis could perhaps focus on more small-world effects, such as item-by-item cosine change across smaller sections of a work such as the Book of Isaiah. What is more, since research of this nature integrates well with other disciplines, future researchers might welcome the opportunity to challenge themselves with the quandaries and quagmires of other areas of study, such as Textual Analysis, Historical Records or Archaeology.