Count or Context: Investigating Methods of Text Analysis

By
Anissa Neal

Honors Thesis
Linguistics
University of North Carolina at Chapel Hill

2016

Approved:

Dr. Katya Pertsova, Advisor
Dr. Peter C. Gordon, Reader
Dr. Michael J. Terry, Reader

ACKNOWLEDGEMENTS

Many, many thanks to everyone who helped me complete this project. Thank you to Dr. Peter C. Gordon, Dr. Katya Pertsova, and Dr. Michel J. Terry for being a wonderful thesis committee. I could not have written this without all your help, advice and constant support. This project has given me a deeper love for linguistics and research, and for that, I cannot thank my committee enough. I would also like to thank my sister, Leah, for reading my drafts, giving me helpful advice, and putting up with my late night panics and linguistic ramblings. She was a constant source of strength when I needed it. This project has been one of the most difficult and rewarding journeys I have ever undertaken, and thanks to the wonderful and tremendous support from all these people, I would do it again. Thank you, every single one of you, for your patience and encouragement. I cannot express how grateful I am that I got to share this experience with you all.

ABSTRACT

Using text as a source of psychological and cognitive information has become a popular subject (Robinson, Navea & Ickes, 2013; Donahue, Liang & Druckman, 2014; Wolfe & Goldman, 2003). To do this, researchers use a variety of methods to analyze text, but Linguistic Inquiry Word Count (LIWC) has become one the more common techniques. LIWC is a token-based method that contains multiple dictionaries representing various psychological states (positive affect, leisure, religion, social words) and keeps a running tabulation of how many words in a given text occur in each category. Latent Semantic Analysis (LSA) is a context-based method that uses statistics to calculate similarity between different texts based off the surrounding words. As a common strategy of analyzing text for psychological states, it is important LIWC be truly representative of the aspects it explores. The dictionaries must accurately represent the categories they measure to be an authentic assessment of the analyzed psychological and cognitive states.

This current study seeks to use LSA to improve LIWC. The hypothesis is that a combination method of the two will perform better than the application of a single token-based method. LIWC and two other token-based methods were compared to a combination LSA-token method. The two techniques were applied to a set of headlines that had been previously judged by humans in terms of emotion and positive/negative valence. The first part of the experiment compared the token-based methods to confirm that they were different from each other but still successful measures of the stimuli. The second part of the experiment compared the correlation between the token-based method and the correct response of the pre-tagged data against the correlation of the combination method and the pre-tagged stimuli. The findings did not support the hypothesis, as the combination method performed worse than the token-based methods. These results, however, suggest further investigation into the power of LSA and its reliance on context. Specifically, LSA may be suited for analysis of longer, more semantically complex texts, not short, basic samples, like the headlines used in this study.

1. INTRODUCTION

Semantic analysis of text has become a popular mode of investigation for cognitive and psychological states, such as depression and suicidal tendencies (Bucci & Freedman, 1981; Stirman & Pennebaker, 2001). Linguistic Inquiry Word Count (LIWC) and Latent Semantic Analysis (LSA) have both been used as methods to extract pertinent cognitive data. LIWC focuses on categorizing word use into a variety of psychological categories, such as social processes, cognitive processes, perceptual processes, while LSA calculates the similarity between words and texts. The types of data they return are different, as are the two methods themselves, but both have been shown to be accurate, powerful ways to analyze large amounts of text. Furthermore, there is evidence to suggest that text, from forum posts to written essays, can contain important signifiers of psychological and cognitive states, like happiness and performance (Alpers et al., 2005; Sexton & Helmreich, 2000; Semin & Fiedler, 1998). This study assesses a combination method of both LIWC and LSA as a way to improve the analytical power of LIWC.

LIWC and LSA take vastly different approaches to semantic analysis. LIWC is a token-based method that is built mainly on its comprehensive dictionaries, composed of words associated with various cognitive and psychological states, while LSA is a form of vector space semantics that uses a function similar to factor analysis to compute similarity between and within texts. As text analysis tools, both are advantageous because they work easily on large amounts of text. Further discussion of the development and applications of each program will be explored in later sections. But despite their different approaches to text analysis, LIWC and LSA have been used to provide insightful information about the relationship between word use and cognitive functions (Landauer, Foltz, & Laham, 1998; Pennebaker, Mayne, & Francis, 1997).

1.1 Purpose of Study

This study seeks to investigate the performance difference between token-based methods and a context-token combination method. Token-based methods, such as LIWC, are a popular choice for text analysis, and it is important to investigate the ways the analysis it does could be further improved. The two methods will be compared against each other based on their performance on analyzing a set of headlines tagged for positive and negative emotion. I hypothesize that the combination method will perform better than the singular token-based method. The combined efforts of both a token-based method and a context-based method will produce results that are closest to the correct values determined by the pre-tagged stimuli. Performance will be calculated through correlation. The correlation between the token-based methods and the actual pre-tagged results will be compared to the correlation between the combination method and the actual pre-tagged results.

The token-based method of focus for this study is LIWC, though two affective databases will be used as a means of comparison. LIWC is built on multiple dictionaries that contain words matched to a variety of psychological and cognitive states (such as emotional affect, personal relationships, etc). I chose to investigate emotion because it can be conceptualized in binary terms, positive (happy, joy, thrill) or negative (depression, sadness, hate) affect. The affect will also be referred to as valence throughout this paper. Furthermore, while LIWC is a popular method, few studies have been done focusing solely on how well it handles emotional stimuli. A study by Bantum & Owen (2009) investigates LIWC's ability to categorize emotion, but not in conjunction with any other method, as this study does. LSA has been proven to be an accurate method of establishing textual coherence, or how similar text is to one another (Foltz, Kintsch, & Landauer, 1998). Given emotional stimuli, then, LSA should be able to classify similar emotions together. With positive and negative affect being such divided categories, it should be easier for LSA to separate stimuli into their respective categories. There is little to no literature on LSA in reference to emotion as well. The purpose of the combination of both token-based methods and context-based methods is to demonstrate how context-based methods can be used in conjunction with token-based methods to produce clearer and more accurate data. Since LIWC is such a popular method of investigation, finding ways in which it can be improved is an important endeavor. By combining it with LSA, this study hopes to enhance the analytical power of LIWC.

The experiment entails two parts. The first part centers on the token-based methods. LIWC will be the token-based method of focus, as it is widely used throughout the psychological community. This first part of the experiment determines that LIWC can be used for analysis of emotion. To do this, I use two affective databases: Affective Norms for English Words (ANEW) by Bradley and Lang (1999) and an updated version of ANEW by Warriner, Kuperman, and Brysbaert (2013), referred to as Updated ANEW. The purpose of the ANEWs was to create standardized materials for researchers of emotion to use so that cross-study comparisons could be explored. LIWC, on the other hand, was built as a program of analysis. Both focus on emotion, but LIWC's goal is to count word occurrences in text, and the ANEWs' goal is to provide a source of material for researchers of emotion. The purpose of comparing these affective databases to LIWC in this study is to determine how successful the affective dictionaries of LIWC are at classifying emotional stimuli but also allow for comparison across multiple token-based methods. The affective databases describe the valence, as well as arousal and dominance, for words averaged from on a scale from 1 to 9 from a multitude of speakers.

LIWC places the words into either the positive or negative affect category. In Part I of the experiment, I apply LIWC and the affective databases to the set of pre-tagged emotional stimuli, headlines from the SemEval Affective Task (Strapparava & Mihalcea, 2007). The hypothesis is that LIWC and the affective databases will have a not have a strong correlation, but the correlation will be positive. If there were a perfect positive correlation, the methods would essentially be the same and not diverse enough to compare as separate methods. The methods need to be similar enough to each other that they can be considered different methods of analyzing affect, but they also need to be different enough that they all do not produce the same results when applied to the SemEval headlines.

Part II of the experiment evaluates the combination method. I apply LSA and the token-based methods to the same set of pre-tagged headlines. Human raters judged these headlines on both emotion (anger, disgust, fear, joy, sadness, surprise) and valence, though only valence is used in this study. The valence results of the combination method will be correlated with human judged valence results of the pre-tagged stimuli, referred to as key valence. For Part II of the experiment, I hypothesize that (a) the combination methods will perform better than the token-based methods and (b) LIWC will perform better than the affective databases. Hypothesis (a) is the overall hypothesis of this entire experiment. With both a token-based method and a combination method, the results should be closer to the actual results of the pre-tagged stimuli. Combining context to an already somewhat successful token-based method should improve the performance of LIWC. I propose (b) because of the different uses for LIWC and the affective databases. LIWC is built to already analyze psychological and cognitive states, of which emotion is only one. Furthermore, LIWC forces affect to fall into one category, either positive or negative. The affective databases, on the other hand, were made to help in the production of experimental material. They also calculate valence as an averaged number, instead of an either/or dichotomy, that will have to be translated into a binary to compare to LIWC. Therefore, LIWC should perform better when faced with the task of separating stimuli in to positive or negative valence. Below is a summary of the hypotheses that form the core of this study.

Hypotheses:

Part I

(i) The correlation between LIWC and the affective databases, ANEW and Updated ANEW will not be strong, but it will be positive.

Part II

(iia) The combination method will have a higher correlation to the actual results of the pre-tagged stimuli than the token-based methods.

(iib) LIWC will perform better than the token-based methods built from the ANEW databases.

The following section, Section 2, will discuss LIWC and LSA in greater detail and contain discussions of past studies that are of particular relevance to this current experiment. ANEW and Updated ANEW will be discussed in this section as well. Section 3 will contain the methodology for Part I and Part II of the experiment and describe the SemEval headlines used for stimuli. Section 4 will present the results of the experiment, and Section 5 will discuss the findings reported and how they can be analyzed in semantic terms.

2. LITERATURE REVIEW

2.1 Linguistic Inquiry Word Count

Pennebaker, Booth, and Francis developed the Linguistic Inquiry Word Count (LIWC) in the mid-1990s as a text analysis tool of different psychological categories, such as social words, positive/negative emotions, insight words etc. LIWC reads in a text and calculates the frequency of words found in each category. It is composed of two key features: the processing component and the dictionaries (Tausczik & Pennebaker, 2010). The processing component is the function that allows LIWC to work its way through large amounts of textual data, moving through each file of text word by word. As it does this, each word is compared to the words contained in dictionaries. LIWC keeps a running tabulation of the words found in each dictionary category, and, once it finishes a file, returns the percentage of words in each category. LIWC dictionaries are built and edited through a complex process involving word generation and human judgment.

2.1.1 LIWC Dictionaries

Initially, the creators of LIWC began by looking for basic groups of words to represent the psychological and cognitive dimensions often studied in "social, health, and personality psychology" (Pennebaker, Booth, and Francis, 2001). They pulled words from several sources, such as the PANAS emotion scale (Watson, Clark, & Tellegen, 1988), Roget's Thesaurus, and English dictionaries. After the creation of preliminary word-lists, judges determined the most relevant words and added them to the initial lists. Once they had broad word lists, the words in the categories of Psychological Processes, Personal Concern, and most in Relativity (excluding verb tense) were judged by three independent judges (Pennebaker, Booth, and Francis, 2001).

The judges were instructed to focus on adding and excluding words. They indicated whether an individual word should be included on that specific list, and suggested other words to be added to the list. For example, the study done in this paper focuses on two categories: positive emotion and negative emotion, both contained under the affective category. Other categories include insight, causation, and achievement. Following this process, all the word lists were updated under the following conditions:

1) a word remained if two of the three judges agreed
2) a word was deleted if two of the three judges determined it should be deleted
3) a word was added if two of the three judges agreed it should added

(Pennebaker, Booth, and Francis, 2001)

Judge ratings for the Standard Language Dimensions category, which includes articles, pronouns, and prepositions, were not collected because of their objective nature.

In the second phase, the judges were given alphabetized word lists of entire categories. They were then asked to determine if each word should or should not be included in the high-level category, and then if any of the words should be included in the mid-level category. For example, the judge would first determine if a word belonged in the high-level category, such as cognitive processes, and then determine if any of the words could be placed in the mid-level categories, such as insight or causation. The dictionaries were then updated following these conditions:

1) a word remained in the category if two of the three judges agreed
2) a word was deleted from the category if two of the three judges agreed

(Pennebaker, Booth, and Francis, 2001)

The original judging was done from 1992-1994, but a major revision was done in 1997. Categories that were used with very low rates were removed, and new categories were added by following the same criteria detailed above.

An external validity experiment was then performed using the updated LIWC. A group of 72 college students was instructed to write either about their thoughts and feelings on coming to college (the experimental group) or to describe a particular object or event in an unemotional way (the control group) (Pennebaker, Booth, and Francis, 2001). Judges then scored the essays on dimensions of emotion and cognition following criteria that were selected to correspond with LIWC categories. A Pearson correlation was done to compare LIWC results to the judges, and a high correlation was found. The authors interpreted this result as support for LIWC validity, as it was able to capture psychological phenomena to the same level as human judges.

2.1.2 Past Studies

LIWC is an oft-utilized tool of text analysis. This section presents studies that specifically address the ways in which LIWC is successful as a text analysis program and the ways in which it could be further strengthened.

Robinson, Navea, & Ickes' (2013) study focuses on function words and how their usage can predict final course performance. Undergraduate students were asked to write self-introductory essays. LIWC analyzed the essays after their conversion into computer readable text. Predictors for course performance were divided into eight overall factors, some employing of LIWC categories. Factor 1, Acting in the Present, contained LIWC categories of auxiliary verbs, common verbs, and use of present tense. Factor 3, Negative/Critical, had the categories of word count and negative emotion words. Factor 5, Gender and Sexuality, contained the sexual words category. Factor 7, Literary Punctiliousness, had the quote and comma categories. Factor 8, Eating and Drinking, had ingestion words. Using these factors, the authors found that end-of semester performance could be predicted from self-introductions written in the beginning of the semester with low-performing students showing stronger predictors of (a) egocentricity (based on use of first person pronouns), (b) linguistic simplicity (fewer words per sentence), (c) narrow focus on their day-to-day lives (less focus on friends and family), and (d) preoccupation with hedonistic creature comforts (use of ingestion and pleasure words). While the researchers used LIWC in addition to factor analysis in this study, the study nonetheless helps impart valuable information about the variety ways in which researchers can implement LIWC in conjunction with other methods of analysis. In fact, the central conclusion from their study is that using LIWC as a multi-tiered approach to analysis can generate comprehensive and insightful results.

Bantum & Owen (2009) discuss the validation of computerized text-analysis programs, and how such programs' validation can be improved for identification of emotional expression. The authors looked at two forms of computational analysis: LIWC and Psychiatric Content Analysis and Diagnosis (PCAD). They note that one major difference between PCAD and LIWC is that PCAD evaluates the word in the context of the surrounding clause. In their study, the two primary aims were to determine if LIWC and PCAD were accurate in detecting emotional expression and exploring the relationship between coding methods and self-report measures. This study used all the LIWC emotion categories, and overall found that the LIWC was comparable to PCAD or significantly better in all emotion categories, thus determining that LIWC was superior to PCAD in detection of emotion. However, the authors also found that LIWC is more likely to over-identify than under-identify. This means that LIWC is more likely than PCAD to identify a word as being part of a category it is not, in fact, a member of. While the authors note “PCAD’s attempts to use context to disambiguate word meanings did not appear to be particularly effective,” they state that LIWC’s accuracy could be improved through adding more sophisticated computational strategies, such as word-disambiguation (86). The study I undertake in this paper uses LSA to address this caveat. Neither LIWC nor LSA look at surrounding clause structure, but LSA does look at surrounding words, which allows for some disambiguation.

Researchers have also investigated the dictionary categories themselves. The integral study by Donahue, Liang, & Druckman (2014) describes a validation process for the addition of new dictionaries. LIWC allows researchers to edit the default dictionaries, but the Donahue et al. study focused on the manner in which these new dictionaries are validated. Their study used rhetoric preceding the 1993 Oslo I Accords, such as speeches, as stimuli. They present three methodological questions that need to be addressed when validating the dictionaries. First, is the creation of new dictionaries necessary or if the default dictionaries will suffice? The authors found that the dictionaries did not fully capture their stimuli and updated them. Second, is using word count an accurate representation of what the speakers are actually expressing? The authors cite Tausczik and Pennebaker (2010) and their study determining that context-free counting can accurately represent psychological and cognitive constructs, as long as the dictionaries these counts have been based off of have been validated. Third, how does one determine the most effective method of validation, since it is the validation of the dictionaries that the results so heavily rely on? According to the authors, the abstract or concrete nature of the validation process will vary depending on how abstract or concrete the investigative constructs are. An abstract validation process would have judges make decisions on the words context-free, while a concrete validation process would be contextually bound.

Donahue, Liang & Druckman suggest three methods as a means of validation. Method 1 is an abstract approach that uses five statistical methods, such as t-tests and factor analysis, to determine whether or not words should be removed, and it produces an updated list from these varying statistical criteria. For example, in regards to the mean analysis, any mean lower than 3.5 was eliminated. For the t-tests, any word significantly different at the 0.05 level was removed. Factor analysis removed any word with a factor loading less than 0.6. In Method 2, the process was similar to the validation strategy used for the LIWC dictionaries. Students were asked to produce essays on one of the six constructs of interest. The lists from Method 1 were then applied to each of the essays. This second approach was more effective than Method 1. Lastly, Method 3 uses coders trained on each of the six constructs to analyze the stimuli, speeches of Palestinian and Israeli leaders. The overall results found that the three methods together failed to successfully validate the dictionaries. That is, the three methods were unsuccessful in determining how the dictionaries of LIWC should be edited and updated when creating novel dictionaries. A mixed approach of both abstract and concrete methods may be best, although the authors warn that mixing methods can lead to ambiguous data.

2.2 Affective Norms for English Words (ANEW)

The first affective database used in this study is the one developed by Bradley and Lang (1999). It is a comprehensive database of English words with affective coding. The database acts as a set of verbal material rated in pleasure (valence), arousal, and dominance, following the dimensional views of emotion that have been widely used throughout the past years. Bradley ad Lange presented an Introductory Psychology class with the words and asked to rate them for all of the above three qualities. The Self-Assessment Manikin (SAM) scale, developed by Lang (1999), was used to rate the affect. The scale ranged from 1 to 9, with 1 being the lowest and 9 being the highest. SAM uses pictorial ranges (excited/bored faces, sleepy/awake faces, happy/sad faces), and allows for the participant to darken bubbles on Scantrons corresponding to the various dimensions. The resulting database contains 1211 words and ratings. Bradley and Lang's full work contains the words further broken down into gender. The variable of focus for this study is pleasure/valence.

2.3 Updated ANEW

The second affective database share similarities to the first. Warriner, Kuperman, and Brysbaert (2013) further developed the original ANEW database. Their version was created from Bradley and Lang's (1999) ANEW database, Van Overschelde, Rawson, and Dunlosky's (2004) category norms, and the SUBTLEX-US corpus (Brysbaert & New, 2009). The remaining words came from a list of lemmas composed by Kuperman, Stadthagen- Gonzalez, and Brysbaert (2012). The participants completed the experiment through Amazon Turk, a total of 1,827 respondents, and rated the word on only one dimension, following the same 1 to 9 scale as in the original ANEW. Overall 1,085,998 ratings were collected, but after trimming and reductions, the resulting database contained 13,915 rated words. The authors compared their results to that of other studies, particularly Bradley and Lang (1999) from which this updated database was based, and they found that valence generalized well across the studies. The variable of focus in Updated ANEW for this study is also valence.

2.4 Latent Semantic Analysis

Landuaer, Foltz, & Laham (1998) developed Latent Semantic Analysis (LSA), a form of text analysis that uses Vector Space Semantics (VSS). It takes large text corpora and returns similarities between words, passages, or full texts. The latent aspect of LSA is that it does not simply compute co-occurrence of the surrounding words, but rather uses a statistical technique called Singular Value Decomposition (SVD), which is similar to factor analysis, to represent human knowledge and judgment of words. Unlike LIWC, LSA considers context in that it takes the surrounding words into account when computing the similarity. As a form of VSS, LSA calculates the similarity in a dimensional "semantic space" built from a training corpus.

LSA can be used in two ways: first, as a means of comparing similarities in and between texts and passages, and second, as a model of acquisition of knowledge. Foltz, Kintsch, & Landauer (1998) use LSA in this first way when they investigate textual coherence, that is how similar texts are to one another, and find that LSA can, in fact, provide an accurate model of coherence. Under the second use, Landauer & Dumais (1997) use LSA to propose a solution to the "poverty of the stimulus" in language acquisition. This study will be using LSA in the first manner, textual coherence. LSA is built on a variety of complicated statistical processes that will not be fully discussed in this paper, but a more detailed account can be found in the Landauer, Foltz, & Laham (1998) study. Below is a brief description of the process:

1. Text is represented as a single cell in a matrix. Each cell contains the frequency with which the word of its row appears in the passage denoted by its column.
2. Cell frequency is weighted by a function that expresses both the word's importance in the particular passage and the degree to which the word type carries information in the domain of discourse in general

3. Singular Value Decomposition (SVD) applied to matrix.
4. Linear decomposition applied next.
5. Reconstruction on two dimensions that approximates the original matrix.

(Landauer, Folts, and Laham 1998)

The following section will discuss some past studies that have used LSA to explore a variety of topics.

2.4.1 Past Studies

LSA considers surrounding words at a level more complex than co-occurrence. Unlike LIWC, which looks only at specific words, LSA considers the environment around the word being analyzed. However, neither method takes word order on a syntactic level into account. Landauer, Laham, Rehder, & Schreiner (1997) investigate how much meaning can be derived without word order. Although the authors recognize that syntax and word order play significant roles in human comprehension, they state that it is difficult to determine just *how much* of a person's information extraction is reliant on word order. To test this, the authors compare the performance of a computational model of understanding that does not use word order, LSA, to judgments of human readers. The comparison is done in two experiments. In the first one, undergraduate students were asked to write essays on the anatomy and function of the human heart. Professional readers at the Educational Testing Service, Inc made the human judgments. The LSA was trained on a corpus created from relevant articles taken from *Grolier's Academic American Encyclopedia*. In the second experiment, psychology students in an introductory course had ten minutes to write essays on one of three psychological topics. A professor and two teaching assistants made the human judgments. The LSA was trained on a corpus made from the course textbook. Human raters read and judged the essays for how much the student knew and how much information was correctly conveyed about the subject using a 1 to 5 quality range. The authors found that LSA measures of the essays were closely related to the human judgments. They conclude that while syntactic knowledge is essential to human comprehension, it is unclear to what degree this syntactic knowledge is needed to gain meaning, and, furthermore, that LSA operates similarly to human judgments of wider meaning.

The Wolfe & Goldman (2003) study focuses on how LSA can be implemented to predict psychological phenomena and the factors a researcher must consider before using LSA for psychological analysis. The authors present two major issues that researchers must consider when using LSA for psychological matters. First, one must determine if semantic relatedness is an accurate measure of the psychological phenomena. Although LSA can be used for a variety of tasks, at its core it is a powerful tool for exploring semantic relatedness. Therefore, when using LSA in a study it is integral that semantic relatedness can reflect some sort of information on the psychological items. The relatedness found by LSA between semantic items must have some interpretable relation to psychological states. For example, determining the relatedness between *cat* and *mouse* in a text does not impart any deep psychological information about the writer. However, semantic relatedness between words such as *joy* and *laugh* in a text suggests something about the affective state of the writer of the analyzed text. In both examples, there is a semantic relation between the words, but only in the second example can that relatedness be extrapolated to cognitive and psychological states.

Second, one must ensure that the “knowledge” of the LSA is representative of the variable being investigated. LSA is trained on a corpus, which builds the semantic space from which the relatedness measures are calculated. As a result, it is important to choose a corpus that is not too specific or too general. Aside from presenting these two theoretical and methodological issues, Wolfe & Goldman also undertake a study involving LSA. They explore how well the complexity of a historical event can be captured by student essays on contradicting causes of the event. The essays were both hand-coded and applied to the LSA, and they found that the LSA predicted the student's reason slightly better than the hand coding. Of central importance in this study are the issues that researchers must consider. Emotion words, the stimuli of this study, are of psychological relevance in that emotions can impart important information about mental states. In regards to the second concern, the corpus, which will be discussed later, is built from movie subtitles, which allows it to reflect human language use in a somewhat naturalistic way.

3. METHODOLOGY

The methodology is broken into two different sections. The first explains the process and results of the first attempt of applying LSA at a word level to determine the cohesion and distinction within and between words in the LIWC dictionaries. Since this method was unsuccessful, I use LSA at a phrase level and apply it to the SemEval headlines. The second attempt and its results form the bulk of this study.

3.1 LSA at a Word Level

Initially I planned to use LSA as a method to determine how consistent the LIWC affective dictionaries were, and, therefore, how well it classified emotional stimuli. If the affective dictionaries were judged adequately coherent within and distinct from each other by LSA, then the affective dictionaries would be sufficient for usage in the combination method. The affective categories of positive and negative emotion were pulled from LIWC and run through LSA. The resulting matrix would then, hopefully, distinguish the consistency of the internal categories by grouping them together in the semantic space and separating from the other category. The binary aspect of positive and negative affect would help exacerbate this pattern of internal consistency and external distinctness.

3.1.1 Word Level Methodology

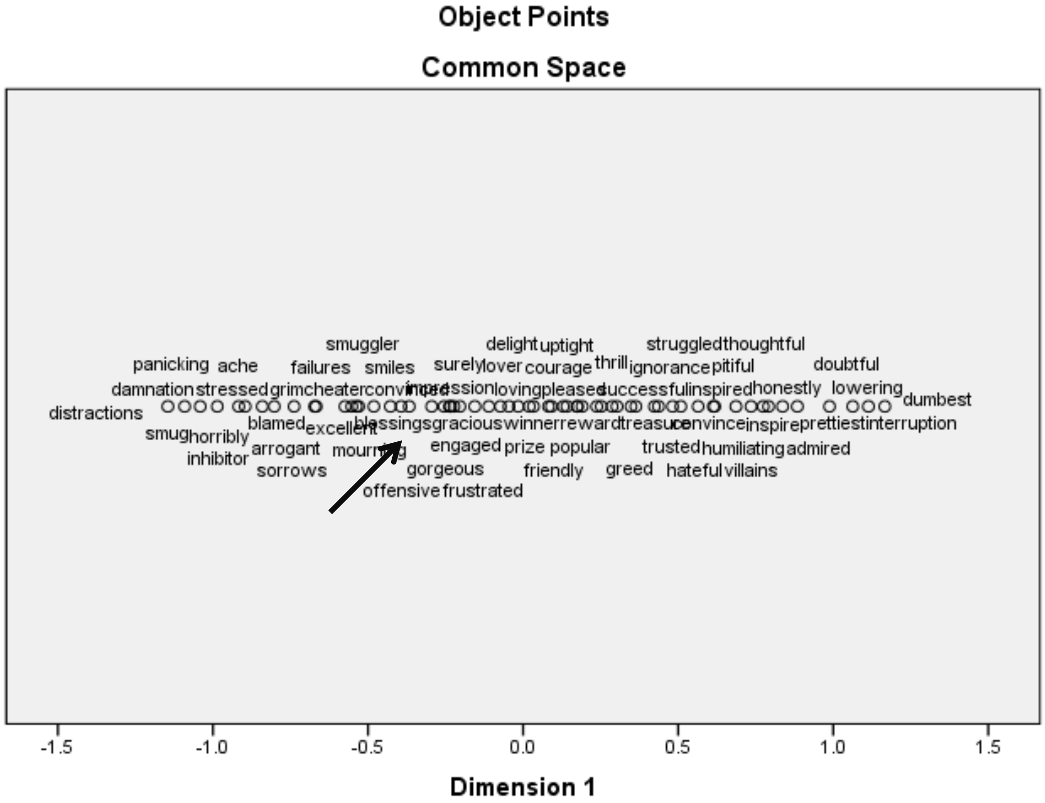
I expanded the wildcards (ex. *happ**, where *happy, happiness, happily* are all counted in the category *happ** is in) of the LIWC dictionaries as a means to encapsulate the full power of

LIWC and handle the incomplete dictionary forms. Furthermore, the wildcard forms could not run through LSA unless they were translated into their lemma forms or expanded into other similar words. To expand the wildcards, I matched them with a word frequency database. Focusing on high frequency words, I determined 20 central words for each category that I deemed representative of the valence. For example, *blessings* was a central word for the positive category. I would use these 20 central words to ascertain the internal consistency of the two categories. I added the 20 central words to a list of 40 other words randomly pulled from the same valence dictionary. With these 120 words, 60 for each valence category, I applied LSA. The goal was to determine which of the 20 central words remained fairly consistent in the semantic space over four different trials. If a word remained in a fairly constant semantic space over the trials, it could be argued that such a word was central to the category, for it must be fairly representative of the category to remain in the same space despite the change in surrounding words. For each trial, the surrounding 40 words were changed and the LSA run again. The words that remained consistent would then compose the positive and negative categories that would be run with LSA on the SemEval headlines in the combination method.

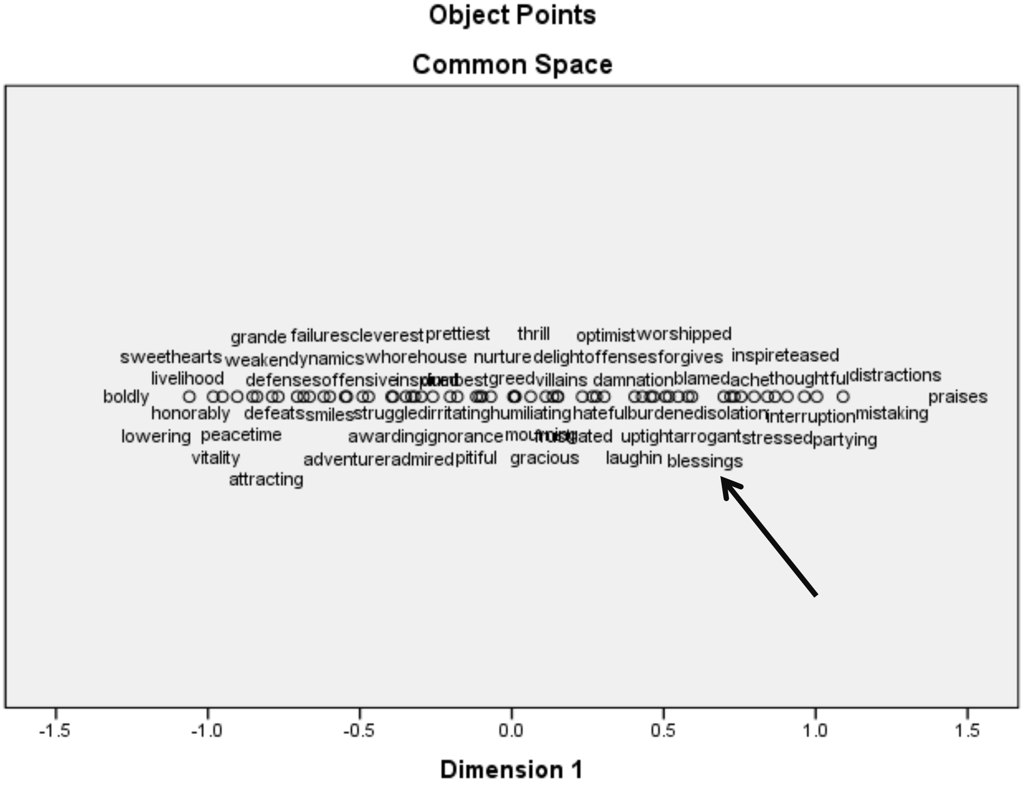
3.1.2 Word Level Results

The resulting LSA matrix for each trial was interpreted using a multidimensional analysis to determine which words remain consistent throughout the trials. Graphs 3.1 and 3.2 show the results from two trials.

Graph 3.1 *Trial 1*



Graph 3.2 *Trial 2*



Across the multiple trials, the results continued to be inconclusive. No words emerged from the original set of 20 as consistent across multiple trial runs. One of the 20 central words, *blessings*, is marked on the graphs, and, as shown, it does not remain constant. The other 20 central words presented similar results. Given more time, it would be interesting to further pursue this line of thought. Due to time constraints, four trial runs were the most doable. However, with a greater number of trials, a clearer picture may emerge.

3.1.3 Word Level Discussion

The length of the string LSA is analyzing in this instance is an important consideration. LSA can read a large file as a single text and compute similarities to other texts. Despite the fact that LSA is trained on a large corpus, the items being analyzed in this study are still single strings of text. There might not be enough semantic information contained in a single word for the LSA to work successfully at this degree. For example, take the word *greed*, one of the chosen 20 central words of the negative valence category. Then take the SemEval headline *Ancient coin shows Cleopatra was no beauty*. The headline example is not particularly rife with affective information, and the single word, *greed*, has the advantage of being, most likely, used in a negative context, but the headline is a seven-word sentence. When LSA reads both examples in as items to be analyzed, the sentence has the advantage of added semantic weight due to its sentence structure and other surrounding words. This idea will be further developed in the discussion of the LSA results in §5.3.

3.2 LSA at a Phrase Level

3.2.1 Stimuli: SemEval 2007 Affective Task

The stimuli of this study come from the 2007 Semantic Evaluation (SemEval) Affective Task done at Swarthmore College. It uses headlines to rate and classify emotions to investigate the relationship between emotion and lexical semantics. Headlines were chosen for this task due to their nature to provoke a reaction, usually an emotional one. Researchers pulled headlines from Google News, CNN, and other news organizations. Participants had two affective tasks relating to the headlines and could choose to do either or both. The emotion annotation task ranked the emotion (anger, disgust, fear, joy, sadness, surprise) of a headline on a scale from 0-100, 0 indicating the emotion is not present and 100 representing the maximum amount of emotion contained in the headline. The valence annotation task ranked headlines on their overall valence, -100 being the most negative and 100 being the most positive. There were a total of 1,000 headlines. The valence variable was the one used in this current study.

3.2.2 Part I

ANEW and Updated ANEW were edited to contain only the word and its respective valence. Using a Python function, I separated the SemEval headlines into single text files containing only the headline. I ran the headlines through LIWC using only the positive and negative LIWC dictionaries. The LIWC program also allows for the usage of novel dictionaries, so I formatted ANEW and Updated ANEW in the appropriate manner and read them into LIWC as dictionaries. The separation of words in the affective databases into two dictionaries, positive and negative, was decided on the basis of their valence. If a word's valence was greater than or equal to five, I placed the word in the positive dictionary. If less than 5, I placed the word in the negative dictionary. The following formula calculates the valence for LIWC, ANEW, and Updated ANEW.

Valence = Positive_Valence – Negative_Valence

Therefore: Valence > 0 Positive (ex. 15-3=12)

Valence = 0 Neutral (ex. 4-4=0)

Valence < 0 Negative (ex. 2-18= -16)

where Positive/Negative_Valence is the frequency of words calculated to be in the Positive/Negative dictionary.

Using SPSS, all the files were merged to create a single file containing the headline, the key valence, and the valence as determined by LIWC, ANEW, and Updated ANEW. I then ran an analysis to determine the correlation amongst the three token-based methods (results in Table 4.1). I also ran a correlation between the valence of each method and the key valence, for comparison to the combination method (results in Table 4.2 and 4.3). I first computed the correlation with all the words in the SemEval headlines (Table 4.2). Then, I filtered the words so that only words contained within LIWC, ANEW, or UpdatedANEW were present (Table 4.3).

3.2.3 Part II

The process for running the LSA involves more steps. The Java program that runs LSA requires three arguments, one of which is a training corpus. This study uses the SubTLex corpus. SubTLex is a corpus built from multiple movie subtitles. I chose this corpus because while movies are not natural speech, they are a close approximation to how humans use language in an everyday sense. The rest of the input consisted of the individual headline files and the dictionary files. LSA reads each file as a single item to be analyzed for comparison. In the case of the individual headline files, each headline was entered as a single item. In the case of the dictionary files, the entire dictionary, containing the entire list of words in said dictionary, was entered as a single item. LSA evaluates the similarity between all the items and returns a matrix of their similarities.

To determine the valence value of the LSA combination method, I used the dictionary text files. LSA returns a computed similarity of each headline file to each dictionary file. To calculate the LSA valence, the similarity of the negative dictionary was subtracted from the similarity of the positive dictionary; it is similar to the formula used in Part I. For each headline, the computed similarity between that headline and the negative dictionary was subtracted from the computed similarity between each headline and the positive dictionary.

LSA_Valence = Headline_PositiveDictionary – Headline_NegativeDictionary

where Headline_Postive/NegativeDictionary refers to the computed similarity between a specific headline and the positive/negative dictionary.

Using SPSS, I created a file containing the headlines, the key valence, and their LSA valences and ran a correlational analysis on key valence and LSA valences (results in Table 4.4).

4. RESULTS AND DISCUSSION

4.1 Part I Results

Below are the results of Part I. Table 4.1 contains the correlation between the three token-based methods.

Table 4.1 *Token-based method correlations*

		LIWCValence	ANEWValence	UpdatedANEW Valence
LIWCValence	Pearson Correlation	1	.265	.290
	Sig. (2-tailed)		.000	.000
	N	1000	1000	1000
ANEWValence	Pearson Correlation	.265	1	.470
	Sig. (2-tailed)	.000		.000
	N	1000	1000	1000
UpdatedANEW Valence	Pearson Correlation	.290	.470	1
	Sig. (2-tailed)	.000	.000	
	N	1000	1000	1000

The results in Table 4.1 depict the correlation of LIWC, ANEW, and Updated ANEW after being applied to the SemEval headlines. The correlations between the ANEWs and LIWC are weak, with all values being less than 0.3, but positive. ANEW, it should be noted, has a stronger correlation with Updated ANEW because Updated ANEW is a further extension of ANEW. None of the values are close to a perfect positive correlation, satisfying hypothesis (i). The weak correlation suggests that there is a correlational relationship, though weak. Furthermore, the correlations are also positive. This is not too surprising as both methods contain similar words, though LIWC was built more with a focus on determining psychological states from a given text and the ANEWs for the creation of stimuli. Regardless, both were built with psychological and cognitive properties in mind. The weak correlation also implies that while the methods are similar in their focus on psychological and cognitive responses, they are not *too* similar. This is also supported by their shared variance. The shared variance between LIWC and ANEW is 0.0702, and the shared variance between LIWC and Updated ANEW is 0.0841. This means that LIWC can account for about 7% of ANEW and about 8% of Updated ANEW. For ANEW, there is about 93% that LIWC cannot predict, 92% for the Updated ANEW. In this unpredictable space, there is the possibility that the other methods, ANEW and Updated ANEW, might capture properties of the SemEval headlines not captured by LIWC. The variance further depicts that while the token-based methods are alike, they are still distinct enough to be comparable. There would be no point contrasting them, especially on the same set of headlines, if they are completely homogeneous in their predictions. The correlation is weak, but this weakness solidifies the assumption that the ANEWs and LIWC should be slightly similar, due to their affective natures, and confirms the hope that the ANEWs and LIWC are not too similar, remaining distinct affective methods.

Table 4.2 *Token Methods: All word correlation*

		KeyValence
LIWCValence	Pearson Correlation	.441
	Sig. (2-tailed)	.000
	N	1000
ANEWValence	Pearson Correlation	.197
	Sig. (2-tailed)	.000
	N	1000
UpdatedANEWValence	Pearson Correlation	.352
	Sig. (2-tailed)	.000
	N	1000

Table 4.3 *Token Methods: Select word correlation*

		KeyValence
LIWCValence	Pearson Correlation	.638
	Sig. (2-tailed)	.000
	N	401
ANEWValence	Pearson Correlation	.308
	Sig. (2-tailed)	.000
	N	450
UpdatedANEWValence	Pearson Correlation	.366
	Sig. (2-tailed)	.000
	N	960

The two tables above present the correlations of LIWC, ANEW, and Updated ANEW and the SemEval headlines. In Tables 4.2 and 4.3, LIWC had the stronger correlation to the correct items, with a correlation of 0.441 and 0.638, all-words and select-words, respectively. The value of note, however, is the 0.638 correlation of LIWC in a select-word context. Table 4.2 depicts the correlation of each database using all words, including words that may not be in that particular database. Table 4.3 shows correlation for only the words contained within LIWC and the affective databases. For example, *tiger* may have been in a SemEval headline and not in LIWC, ANEW, or Updated ANEW, but it was still considered when correlations were computed. The results show that the correlations, while positive, are low. These correlations consider all 1,000 headlines. Table 4.3 removes headlines with words such as *tiger* and looks only at headlines with words that are known to be in the token-based methods. This filtering process has quite the effect on the correlation. LIWC has the strongest correlation, particularly in the select-words case, across both tables, and shows the greatest increase between the two tables as well. These results, both select-words and all-words, support hypothesis (iib) in that LIWC would perform better than the other token-based methods. However, LIWC's high correlation comes from only 401 words, and ANEW and Updated ANEW considers 450 and 960 words, respectively. Furthermore, the increase in correlation from all-words to select words is expected because if a method is given words that it has valence for, then the calculated correlation should be stronger when it does not have to consider words it does not have valence for. While this information is helpful in expressing LIWC's high correlation with the key valence, thus supporting (iib), it does not help with generalization as this higher correlation is found only in the select-word scenario. The scenario of greater interest is how well the token-based methods do in a general context, without the benefits of filtering. Table 4.3 indicates that the token-based methods do not do as well. The purpose of the combination method with LSA, discussed in the next section, is to alleviate this issue of generalization by adding another layer of disambiguation.

4.2 Part II Results

Below are the results of Part II, the combination method.

Table 4.4 *Combination Method: LSA Correlations*

		KeyValence
LIWC07_LSA_Val	Pearson Correlation	.175
	Sig. (2-tailed)	.000
	N	1000
ANEW_LSA_Val	Pearson Correlation	.164
	Sig. (2-tailed)	.000
	N	1000
UpdatedANEW_LSA_Val	Pearson Correlation	.191
	Sig. (2-tailed)	.000
	N	1000

Table 4.4 shows the results of the combination method using LSA. The LSA method presents weaker results. The Updated ANEW-LSA pair showed the strongest correlation overall with 0.191, but it is still less than the weakest correlation for token-based methods applied to all words. Furthermore, this higher correlation could be attributed to the shared nature of the Updated ANEW and the training corpus of the LSA; both use SubTLex. The combination method with LIWC was the next highest correlation. The singular token-based method is more successful than the combination LSA method, disproving hypothesis (iia), in which I believed the context-based nature of LSA would act as an amplifier and produce stronger and more consistent results than the token-based methods. The combination method, it appears, was unsuccessful in improving the token-based methods' ability to generalize.

5. GENERAL DISCUSSION

LSA's classification power is another area of interest to this study, especially given the unpredicted results. The positive and negative affective categories were chosen for investigation because of their binary nature. In LIWC dictionaries, a word cannot be in both categories. I edited ANEW and Updated ANEW to act in a similar manner. If the valence of a word was five or higher, the word had positive affect; if lower, negative affect. This was done in order to fit the binary of the investigated categories of LIWC. LSA, however, requires no such approach since it determines the similarity of items in reference to each other, not based off a pre-determined binary assignation. The two methods, while perhaps equally powerful, are suited for measuring different types of classification. LIWC, as evidenced by the results, does exceptionally well when it can analyze every word of the stimuli. When there are words in the stimuli not present in LIWC, the correlation decreases. The same pattern can be seen for the ANEW and Updated ANEW. It is unsurprising that these methods work well on the SemEval headlines because, while representing a range of emotion and ratings, they are still only headlines, which are usually fairly descriptive but also very succinct. This length issue arose earlier with the word level analysis of LSA on LIWC (§3.1). Even in the LSA-combination method, the headlines are compared against a dictionary file containing lists of singular words that, while related to each other based off their positive or negative affect, are still not full sentences. The LSA valence is determined based of headline similarity to the affective dictionaries. While LSA does not work through the dictionary files one word at a time, but instead reads the whole file as a single source of text, it still runs into the problem of lacking context because the words are presented in isolation. Lists of words, while longer than single words, still do not contain much in the way of context. Even the word order study done by Landauer, Laham, Rehder, & Schreiner (1997) used essays, much more substantial than headlines. Length alone is not the issue, as there were a total of 1,000 headlines all of varying length, but the surrounding semantic context is.

Headlines are known for their ability to capture the reader's attention and can occasionally use metaphorical language to do so. The headline *Poison Pill to Swallow: Hawks Hurting After Loss to Vikes* uses *poison pill* metaphorically quite successfully. However, the surrounding words are still not enough for the metaphorical meaning to be fully captured. The metaphorical interpretation comes easily to me because I have enough contextual and experiential background to understand the meaning. Furthermore, I have access to other similar metaphors, such as *that's a hard pill to swallow*, and I can conceptualize the negative affect that would be associated with such a metaphor. LSA does not have access to all this nuanced meaning from the single headline. The training corpus does help augment this knowledge gap somewhat, but the stimuli itself needs to be of enough contextual significance that LSA can correctly calculate similarity. The token-based methods would be able to analyze words like *loss, hurting,* and *poison* as negative, finding them contained on their negative dictionary list. Even with a few examples of metaphorical language, the token-based methods have an advantage over LSA. The token-based methods look at each word comprising the headline and determine if a given word is, first, in the dictionary, and, second, is positive or negative valence. LSA, on the other hand, must consider the surrounding words comprising the headline. So even without instances of metaphorical language, headlines such as *UK announces immigration restrictions*, LSA must compare a single word to the surrounding words of the headline.

When one considers that human judges rated the SemEval headlines, it may seem counterintuitive that LIWC and other token-methods are better suited than LSA, which has shown to somewhat resemble human knowledge (Landauer, Laham, Rehder, & Schreiner 1997).

However, humans, while evidently capable of higher-level thinking and functioning, especially in regards to stimuli such as metaphors, are also perfectly capable of performing simpler tasks. I do not doubt that human raters could judge both metaphorical speech and non-metaphorical headlines just as well. The LSA, on the other hand, may not be suited for this more basic level of classifying analysis. The advantage of LSA is that it can consider surrounding context. But this is only an advantage when the surrounding context is rich enough. The context provided by the headlines alone is too limited. Further studies in this area are necessary as they would be highly illuminating. LSA needs a nuanced source of text to work to its full potential. The reason LSA was unsuccessful in validating the dictionaries of LWIC (§3.1) is the same reason it is shows weaker performance than the token-based methods on the headline stimuli; it unsuccessful at both a word level and a phrase level because there is not enough surrounding context for the context-based method to work to its full potential. LSA is trained on a corpus, but this is only half of the battle. The stimuli read into the program must also be sufficient enough if it is to produce constructive results.

6. CONCLUSION AND FURTHER CONSIDERATIONS

The token-based method performed better than the context-based combination method. These findings did not support the hypothesis that the combination method would perform better than the singular token-based method. They did, however, support the hypothesis that LIWC would perform better overall. The token-based and context-based combination methods are suited to measuring different aspects of meaning, making it difficult and ineffective to compare them. The token-based methods work well on emotionally dense, succinct stimuli. LSA appears better suited for more complicated semantic structures, but not more basic level tasks, such as the binary emotional classification done in this study. Additionally, LSA seems too powerful a tool to work well with the stimuli used in this experiment. Further investigation into this issue would use context-based methods, like LSA, with longer bodies of textual stimuli to see if performance improves as context becomes richer. This is to ensure that there is enough surrounding context to fully support the true power of LSA. Another avenue of research would be to investigate if LSA is truly unsuccessful with the types of stimuli used in this study and suggest reasons why or ways to improve upon it. Text analysis and semantic analysis continue to be important areas of linguistic research, and capturing the full extent of meaning, not just emotion, remains highly relevant to exploring the ways in which meaning interacts with mental states. The methods used for semantic analysis of text must continue to be evaluated to ensure the accurate assessment of the variables it is measuring.

7. REFERENCES

Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., et al. (2005). Evaluation of computerized text analysis in an Internet breast cancer support group. *Computers in Human Behavior, 21,* 361-376.

Bradley, M.M., & Lang, P.J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida

Bucci, W. & Freedman N. (1981). The language of depression. *Bulletin of the Menninger Clinic, 45,* 334–358.

Donahue, W.A., Liang, Y., Druckman, D. (2014). Validating LIWC dictionaries: The Oslo I Accords. *Journal of Language and Social Psychology, 33,* 282-301.

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.

Gottschalk, L.A., & Gleser, G.C. (1969). *The measurement of psychological states through the content analysis of verbal behavior*. Berkeley: University of California Press.

Landauer, T.K., Dumais, S.T. (1997). A solution to Plato's Problem: The latent semantic theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E., (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M. G. Shafto & P. Langley (Eds.), Proceedings of the 19th annual meeting of the Cognitive Science Society (412-417). Mawhwah, NJ: Erlbaum.

Pennebaker J.W., Chung, C.K., Ireland, M., Gonzales, A., Booth, R.J. (2007). *The Development and Psychometric Properties of LIWC2007*. Austin, TX: LIWC.net.

Pennebaker, J. W., Mayne, T. J., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology, 72,* 863-871.

Robinson, R.L., Navea, R., Ickes, W. (2013). Predicting Final Course Performance From Students' Written Self- Introductions: A LIWC Analysis. *Journal of Language and Social Psychology, 32,* 469-479.

Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology, 54,* 558-568.

Sexton, J. B., & Helmreich, R. L. (2000). Analyzing cockpit communications: The links between language, performance, and workload. *Human Performance in Extreme Environments, 5,* 63-68.

St-Hilaire, A., Cohen, A.S. & Docherty, N.M. (2008). Emotion word use in the conversational speech of schizophrenia patients. *Cognitive Neuropsychiatry, 13,* 343-356.

Stirman S.W. & Pennebaker J.W. (2001). Word use in the poetry of suicidal and non-suicidal poets. *Psychosomatic Medincine, 63,* 517–22.

Strapparava, C. & Mihalcea, R. (2007) SemEval Task 14: Affective Task. Swarthmore College.

Tausczik, Y.L., Pennebaker, J.W. (2010). The Psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29,* 24-54.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45,* 1191-1207.

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063-1070.

Wolfe, M.B.W., Goldman, S.R. (2003). Uses of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behavior Research Methods, Instruments, & Computers, 35,* 22-31.