A Semantic Unsupervised Learning Approach to Word Sense Disambiguation

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Dian I. Martin
May 2018

ACKNOWLEDGEMENTS

I kindly thank my research committee members, Dr. Michael Berry, Dr. Audris Mockus, Dr. Kevin Reilly, and Dr. Brad Vander Zanden for supporting my doctoral candidacy. I extend my deepest gratitude to Dr. Michael Berry who has supported and encouraged me in my academic research and career endeavors over many years. Without him, I would not be where I am today professionally nor would this dissertation have ever come to fruition.

While I can no longer express my thanks directly, I owe a debt of gratitude to the late Dr. Tom Landauer. I thank him for the opportunities he provided me to work with him in many pioneering efforts developing LSA as a theory of learning and applying it to many different areas of application. His insights and in-depth knowledge of LSA, as well as his personal encouragement and our conversations, have shaped the direction of this research. The last project in which I had the pleasure of working with him as a colleague involved word maturity and the development of an automatic reading tutor using LSA. Those projects inspired the idea of using LSA as a learning system for inducing word senses. One of my regrets is that I did not finish this work before his passing.

Special thanks go to my immediate family. I could not have accomplished this work without them. I am forever grateful for the love and patience of my three children, Julia, John Everett, and Wyatt. An extraordinary thanks to my dear husband John, for his enduring love, support, help, suggestions, and hours of proofreading. This research would have never been possible without him

Finally, I would like to acknowledge my Lord and Savior Jesus Christ. Without Him, I could have never made it through this journey.

ABSTRACT

Word Sense Disambiguation (WSD) is the identification of the particular meaning for a word based on the context of its usage. WSD is a complex task that is an important component of language processing and information analysis systems in several fields. The best current methods for WSD rely on human input and are limited to a finite set of words. Complicating matters further, language is dynamic and over time usage changes and new words are introduced. Static definitions created by previously defined analyses become outdated or are inadequate to deal with current usage. Fully automated methods are needed both for sense discovery and for distinguishing the sense being used for a word in context to efficiently realize the benefits of WSD across a broader spectrum of language. Latent Semantic Analysis (LSA) is a powerful automated unsupervised learning system that has not been widely applied in this area. The research described in this proposal will apply advanced LSA techniques in a novel way to the WSD tasks of sense discovery and distinguishing senses in use.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Lexical ambiguity is natural in text and is a fundamental characteristic of language. Words have more than one meaning. Given the 121 most frequently used words in the English language, which occur approximately 20 percent of the time in text, these words have an average of 7.8 different meanings, or senses, associated with each of them (Agirre and Edmonds 2007a). Yet to a human reader these words have very little ambiguity when actually read in context.

Word sense disambiguation is the process of determining which sense of a word is being used in a given text passage. Essentially, it is a classification problem (Agirre and Edmonds 2007a, Yarowsky 2000). The challenge in Natural Language Processing is the automation of this classification decision. What is essentially a simple task for a human reader is one that is very difficult to accomplish using algorithmic systems (Pal and Saha 2015, Bhala and Abirami 2014, Agirre and Edmonds 2007a).

1.1 Word Sense Disambiguation

Automated word sense disambiguation (WSD) is a challenging task because word meanings do not necessarily divide into a certain number of discrete senses. A word sense is the particular meaning associated with a word in a given context of usage while the definition of a word encompasses all of the possible meanings or senses. Word senses can be considered either coarse-grain or fine-grain depending on the level of detailed distinction that is required to divide them. Coarse-grain senses exhibit major, unrelated differences in meaning. Words that possess only coarse-grain senses are called homographs. Fine-grain senses have a more subtle or nuanced distinction between them, and they are often

interrelated. Words that have mainly fine-grained senses are considered to be polysemous (Yarowsky 2000). Frequently, words have both coarse and fine-grain senses associated with them. For example, the word *bank* may be viewed as having two clear distinct senses as a noun with one meaning being "a slope beside a body of water", while another sense would be "a financial institution". These are coarse-grained senses. However, the word *bank* when taken in the sense of "a financial institution" can be divided into more subtle, fine-grained, distinctions (i.e.: "the physical building where financial transactions are performed", "the financial institution that accepts deposits and participates in different lending activities", "a reserve of money", or "a container for keeping money") (Agirre and Edmonds 2007a). The text phrase "*I am going to put my money in the bank*." has several possible interpretations for the sense of the word *bank*. Is the money being put in a piggy bank? Is the money going to be deposited at a financial institution? Or, is *bank* referencing a fund that is being put aside for future use or emergencies. Determining the level of granularity depends on the application and context of the word being used.

Interestingly, human readers seem to be able to do this innately, differentiating between multiple senses of an ambiguous word given sufficient context. Of course, individual humans also have differences in their ability to perform this disambiguation task given their cognitive background, exposure to spoken and written language, and domain specific vocabulary. Even for human experts, lexicographers, determining the number of senses of a word and giving those senses a definition is a challenging and subjective task (Pederson 2007).

1.1.1 Historical Background

Formal research in the area of word sense disambiguation began as far back as the late 1940s and early 1950s both in computational linguistics and artificial

intelligence (AI). Alan Turing in his 1950 paper "Computing Machinery and Intelligence" expressed indirectly the importance of WSD. He described the primary indicator of the existence of intelligence as being the ability to understand language (Turing 1950). One of the tests of understanding language is the ability to disambiguate the meaning of words (Hirst 2007). WSD is the one of oldest problems in the field of Natural Language Processing (NLP) (Yarowsky 2000). It is also a facilitator in tasks and applications in computational linguistics (Jurafsky and Martin 2009). WSD was recognized as a crucial aspect of machine translation due to the fact that differentiating between senses for words being translated became a major hurdle. Often there is one word in a given language that has the potential to translate into multiple words in another language each with a different meaning. For example, the English word *sentence* in the legal context translates to one of two different words in Spanish, *sentencia* or *condena*, depending on subtle differences in the context in which the word *sentence* is used (Yarowsky 2000). A human speaker may not even consciously differentiate the senses involved unless faced with the task of translating to the other language.

Original work in WSD was performed manually by human lexicographers. Attempts to automate WSD first began to appear in the 1970s. These systems involved artificial intelligence (AI) approaches to address the problem of sense discovery. The first AI work in the field was the development of a "preference semantics" system, and the research in this area led to "word experts". In the AI paradigm, proper knowledge representation as a ruled based system was needed (Yarowsky 2000). Since this information could not be automatically gathered, this meant that knowledge sources had to be hand-coded. Since then, four major categories of WSD approaches have developed: dictionary- or knowledge-based methods, supervised methods, minimally supervised methods, unsupervised methods (generally named word sense discrimination (Schütze 1998, Pedersen 2007))

(Agirre and Edmonds 2007b, Jurafsky and Martin 2009, Pal and Saha 2015, Bhala and Abirami 2014, Navigili 2009).

1.1.2 Solution Approaches

There are two major aspects of the WSD problem: Sense definition where the possible senses for a word are described and then sense identification where the particular sense in use for a given context is determined. Dictionary- or knowledge-based methods rely on existing dictionaries or some a priori lexical knowledge base, such as WordNet (Felbaum 2012, WordNet), that has been built up manually to inform their processing decisions for the identification task. Supervised methods for WSD use sense-annotated corpora produced by human lexicographers for training an automated system which is then used for the identification task. Semi- or minimally-supervised methods involve building a disambiguation model based on small amount of human annotated text or word-aligned bilingual corpora and then bootstrapping from this seed data to build additional sense indicators which are then used to identify a sense used in a given context. Unsupervised methods for WSD use no external information and work directly with raw non-annotated corpora to induce the senses for words.

Currently, the best results for automated WSD are achieved using supervised methods (Pal and Saha 2015, Navigli 2009, Zhou and Han 2005). These methods, however, require extensive sense-tagging training data to be available beforehand. This requirement poses a problem both in time and expense, as well as the lack of flexibility in dealing with ever changing language. Systems based on knowledge-based, supervised, and semi-supervised methods all require significant a priori knowledge and must be custom built by humans for a specific target language. These systems are also capacity limited in the number of terms for which they can distinguish senses. This leads to problems when using larger

amounts of text, new domains, or new languages. A recent survey of WSD algorithms in 2014 (Bhala and Abirami 2014) suggests that WSD algorithms work best on large volumes of data. This finding further restricts the use of supervised methods. Unsupervised systems for WSD exist, but they have not been well developed or their accuracy lags that of the other types of systems (DiMarco and Navigili 2012, Boyd-Graber et al. 2007, Gliozzo et al. 2004, Navigli and Lapata 2010, Pilehvar and Navigli 2014, Tomar et al. 2013).

1.1.3 Application Areas for WSD

Ambiguity in word meaning presents many challenges in text analysis, or any area involving natural language processing. As indicated by the ongoing workshops, SemEval (SemEval-2018), and work devoted to the evaluations of computational semantic analysis systems (Resnik and Lin 2010, Resnik and Yarowsky 1999), there is still a challenge in enabling a computer to derive meaning from natural language input. Senseval and SemEval are a series of workshops created specifically for the evaluation of WSD systems developed by different groups in the computational linguistics community (Edmonds and Cotton 2001, Edmonds and Kilgariff 2002, Kilgariff 1998, Kilgariff and Palmer 2000). The specific area of interest that prompted this research is text analysis in educational applications, specifically modeling vocabulary acquisition and reading comprehension to facilitate automatic tutors to maximize learning (Biemiller et al. 2014, Landauer and Way 2012, Kireyev and Landauer 2011, Landauer et al. 2011, Foltz et al. 1998). In the knowledge acquisition or education domains where automated systems, such as auto-tutors, need to be able to understand the abilities of the user or student by analyzing their responses and providing feedback to assist in the process of learning, comprehending a target text, acquiring certain vocabulary, etc., WSD is a critical capability.

Other application areas where WSD is a key factor include information retrieval, text mining, message understanding, information extraction, and lexicography (Navigli 2009, Agirre and Edmonds 2007a, Agirre and Edmonds 2007b, Yarowsky 2000). In information retrieval, WSD contributes to distinguishing between relevant and non-relevant documents for a given query. Text mining benefits from WSD in building systems that analyze text because being able to understand word meanings is essential to text understanding. The capability of extracting certain target information from a body of texts, such as news wire reports, patents, email databases, etc., requires systems that are capable of understanding the language contained in the texts, and therefore the ability to identify the correct sense of a word in a given usage. WSD is vital to lexicography, building intelligent dictionaries, thesauri, and grammar checkers. An automated WSD system would have a potentially broad impact for different communities in computational linguistics, sentiment analysis, and machine learning as well. Any automated system that needs the ability to process text and distinguish the meaning of certain words or contexts would benefit from robust WSD facilities.

Another possible application of the technology used in automated WSD is the analysis of the learning system that automated WSD process is built upon. The ability of unsupervised systems to induce word senses from a corpus provides an opportunity to view and analyze the meaning or knowledge contained in that corpus. The particular approach to WSD described in this dissertation is based on a learning system that utilizes Latent Semantic Analysis (LSA) to construct this underlying learning system. While the intention here is to employ this learning system for sense induction and sense disambiguation, it is also providing insight into the learning system itself. To the extent that the learning system represents an AI, the induced word senses serve to describe the knowledge that the AI has obtained. This same learning system may be used for other purposes such as text

analysis or other applications where it is used to represent a body of knowledge for different processing objectives. The ability to describe or analyze the learning system and the way it understands language can be leveraged to better tune or fit the learning system for those other uses.

1.2 Overview

This research describes the application of LSA as an unsupervised learning system for the tasks of word sense induction and disambiguation. A description of the background technology is presented in Chapter 2 along with the innovative developments to be tested and refined for the purposes of discovering word senses and distinguishing them in context. The document sources used for these experiments are also described. Three main aspects of this work are then discussed in subsequent chapters. Part of the development of this approach focused on the concept of word importance and the impact of individual words on the collective meaning of a sentence and is presented as the focus of Chapter 3. The task of word sense induction and the observations from the experiments using the learning system to automatically induce word senses is the subject of Chapter 4. In Chapter 5, the final task of word sense disambiguation, identifying the sense of a word as it is used in a particular context, is covered. Finally, Chapter 6 presents general conclusions arrived at from the experimentation and recommendations for additional follow on research in this area.

CHAPTER 2
ALGORITHMIC INNOVATION

In order to facilitate automated WSD there is a need to develop unsupervised algorithms that rely on no resources beyond original non-annotated monolingual corpora and yield comparable results to existing supervised systems or human understanding. The main focus of this research is to develop an unsupervised WSD algorithm that is based on the following assumptions that have been observed in linguistic research:

- Words with similar meanings occur in similar context across a large body of text (Landauer 2007, Landauer 2002, Foltz 1996)

- Words exhibit the same sense within a document (one sense per discourse) (Stevenson and Wilks 2005, Yarowsky 1995)

- Words exhibit only one sense within the context of the few words immediately around them (one sense per collocation) (Stevenson and Wilks 2005, Yarowsky 1995)

Given these characteristics, a system capable of clustering words based on contexts given in the corpora, rather than on pre-existing sense inventories, would provide an automated way to both discover and distinguish word senses in a given text corpus. One such system is LSA. LSA is used to form a cognitive model that is able to exploit patterns of word usage across a corpus to learn the meaning of words relative to each other within that body of text (Deerwester et al. 1990, Landauer and Dumais 1997, Landauer et al. 1998a, Martin et al. 2016). It has been noted that, given a context in which an ambiguous word is used, LSA can determine its sense for information comparison purposes (Landauer 2007, Kintsch 2007). Since LSA is a fully automated process taking non-annotated text as input

to produce its cognitive model (or LSA semantic space), it is a viable, promising approach for the construction of an unsupervised WSD algorithm. This research will utilize LSA as a foundation for an automated WSD system, leveraging the word meaning that is captured in the cognitive model to investigate the idea of word senses and develop a process for the identification of the specific word sense within the context of a particular usage. This analysis process will also provide insight about the knowledge base represented in the underlying semantic space.

2.1 Latent Semantic Analysis

Latent Semantic Analysis is an unsupervised learning algorithm that has been shown to mimic human learning and thought. There are many applications in which the performance of a LSA-based learning system on certain cognitive tasks has simulated human knowledge and the understanding of words and meanings of text (Martin et al. 2016, Landauer 2007, Landauer 2002, Landauer 1998, Landauer et al. 1998a, Landauer et al. 1998b, Landauer and Dumais 1997, Foltz 1996). One exemplary application of using the LSA-based learning system is the Intelligent Essay Assessor which has been used in many areas to score essays as well as provide analysis on essays (Hearst 2000, Landauer et al. 2003a, Landauer et al. 2003b, Foltz et al. 2013). The LSA-based cognitive model has also been used in auto-tutors which interact with users to help them gain more knowledge or understanding about a certain subject (D'mello and Graessar 2012, Graessar et al. 2007, Landauer et al. 2009, Kintsch et al. 2007). For more extensive lists of example applications which have leveraged the LSA cognitive model to represent humanlike understanding see (Martin et al. 2016) and (Landauer et al. 2007).

LSA is based on the compositionality constraint: "The meaning of a document is the sum of the meaning of the terms that it contains." The corollary to that constraint also applies: "The meaning of a term is defined by all the contexts in

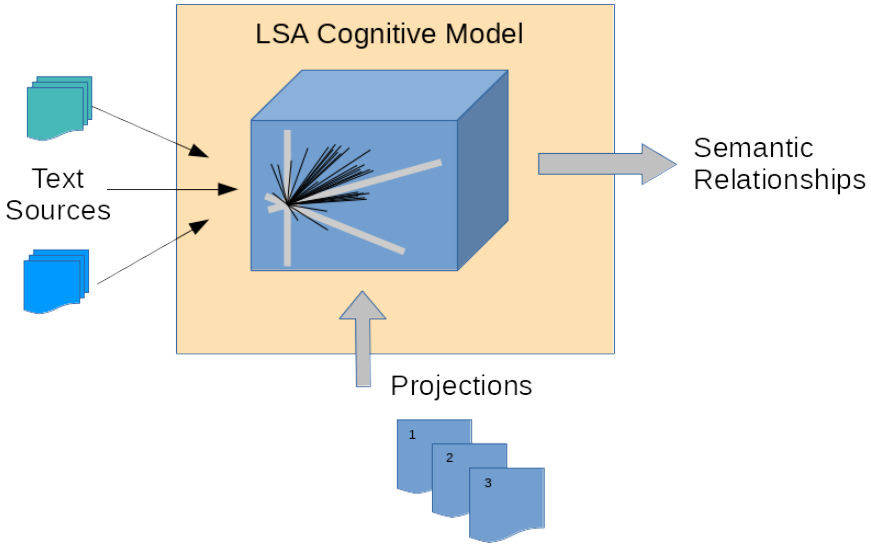

Figure 1: Visual representation of the LSA-based learning system

which it does and does not appear" (Landauer 2007, Landauer 2002). The LSA-based learning system, as illustrated in Figure 1, includes the input text from which to learn, the cognitive model represented by the LSA semantic space, and the functions necessary to interrogate or probe the cognitive model.

2.1.1 Mathematical Foundations

LSA uses a mathematical technique to represent the underlying semantic relationships between words and to model human understanding of language (Martin and Berry 2010, Martin and Berry 2007, Martin et al. 2016). The semantic relationships are derived by representing the input text in a high-dimensional vector space which is produced by the singular value decomposition (SVD). A LSA learning system is built by processing a text corpus that provides some representation of a body of knowledge, as well as background information for learning language.

To begin processing, the input text for the LSA-based learning system is transformed into a matrix, essentially a table, where the rows represent the unique terms and the columns correspond to the actual contexts in which those terms appear. The terms are generally single words, but they could also be different constructs derived from the contexts such as multi-word phrases known as *n*-grams. Contexts are generally paragraph-sized documents, but could be passages, sentences, phrases, or chapters. To capture the language and linguistic framework representative of a literate adult requires a text corpus consisting of a minimum of 100,00 paragraph-sized pieces of text, though better linguistic representation can be achieved with larger corpora (Landauer 2007).

The constructed term-by-document matrix is a sparse matrix where each cell contains a term frequency value which is the count of the number of times a term appears in a context. Generally, the cells in the matrix are then weighted by local and global weighting functions. A description of the different types of functions are given in (Martin et al. 2016), (Martin and Berry 2007), and (Dumais 1991). The local weighting function is applied to each cell to normalize term frequency value within each context. The global weighting function is applied to normalize the global term frequency (*gf*) across all contexts. In practice, the local-global weighting function of *log-entropy* is most often applied within a LSA-based learning system because it replicates the ideas of learning (Landauer and Dumais 1997). The local portion of the function, log of the term frequency ($\log(tf+1)$), is intended to approximate the simple growth of standard leaning. This reflects the effect that semantic association between two terms is greater if they appear together once in two different contexts than if they appear repetitively in the same. The global function, entropy

$$(1+\sum \frac{p_{ij}\ log_2 p_{ij}}{log_2(n)}\ , \text{where } p_{ij}=\frac{tf_{ij}}{gf_i}\ ,\ \text{for term } i \text{ and document } j)$$

yields an estimate of the degree to which observing a term indicates the context in which it appears. Stated differently, when a term has a smaller entropy value, suggesting that the word appears frequently in the corpus, that word carries less information about the contexts in which it has occurred; therefore, it has less usage-defined meaning (Landauer and Dumais 1997).

Once the matrix has been weighted, the SVD is applied to decompose the matrix and create a vector space in which all terms and documents are represented as hyper-dimensional vectors. The SVD takes the input matrix, A, and produces three matrices: $A = U\Sigma V^T$. The U matrix and the V matrix are orthogonal matrices containing the r orthonormal eigenvectors of AA^T and A^TA respectively, where r is the rank of matrix A. The Σ matrix is a diagonal matrix containing the nonzero square roots of the eigenvalues, the singular values, associated with AA^T and A^TA. Each of these three matrices can be truncated at any rank, $k < r$, which will yield the best k-rank approximation of A. The truncated SVD is defined as $A_k = U_k\Sigma_k V_k^T$. A pictorial representation of this can be seen in Figure 2 (Martin

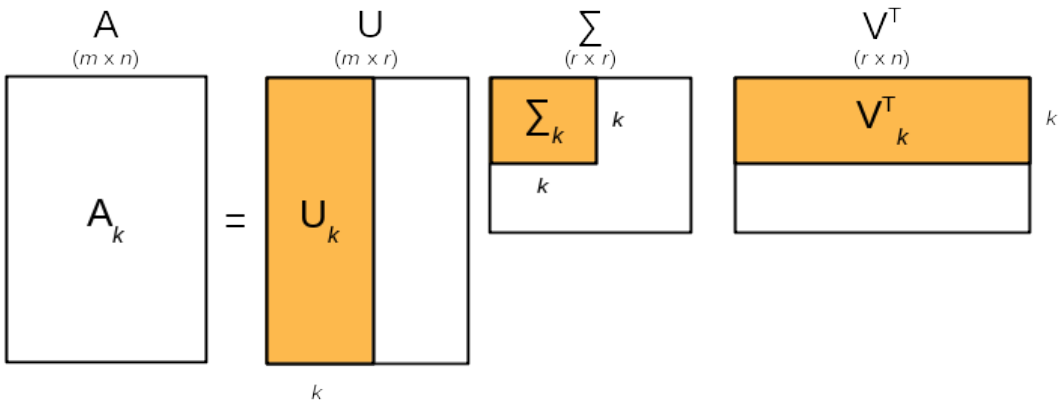

Figure 2: Depiction of the truncated SVD at rank-*k* (Martin and Berry 2007, Martin and Berry 2010).

and Berry 2007, Martin and Berry 2010). The rows of U_k represent the term vectors in the semantic space, and the rows of V_k represent the document vectors.

The truncated rank-k vector space captures the semantic meaning of both the terms and documents from the input corpus. The number of dimensions, k, is usually selected to be around 300 dimensions in typical use (Landauer et al. 2007, Martin and Berry 2015, Martin et al. 2016). The conversion of the weighted matrix by the SVD into a truncated vector space takes the initial information of terms and contexts and transforms it into a unified representation of knowledge, a semantic mapping. The truncation, or dimension reduction, of the initial matrix A is necessary to capture the underlying semantic meaning of terms and documents. The relationships between the resulting term and document vectors within the semantic mapping depend not only on direct associations of terms within a document but also with all the other contexts in which a term appears and does not appear. Consequently, within the mapping system, terms that are associated in meaning within a single document might not be similar in meaning in the context of other documents. Likewise, documents can be identified as similar in meaning even though they do not have any overlapping terms (Landauer and Dumais 1997). This semantic mapping system represents the cognitive model of the learning system as shown in Figure 1.

Once the cognitive model is produced, the LSA-based learning system can probe, interrogate, and investigate the meaning relationships between the term and document vectors in the semantic space. Generally, the cosine similarity measurement is used to quantify meaning relationships between the items within the vector space. Given two term vectors, say t_1 and t_2, within in the semantic space, the cosine similarity is defined as

$$cosine = \frac{t_1 \cdot t_2}{\|t_1\| \, \|t_2\|},$$

where $\|t_1\|$ is the vector length of t_1 and $\|t_2\|$ is the vector length of t_2. The Euclidean vector 2-norm suffices for these vector length calculations. It is worth noting that the cosine measures between two vectors in the semantic space are values that range from -1.0 to 1.0. A cosine value of 1.0 indicates that the vectors are identical in their mapping which indicates that they express the same meaning, while a cosine value of -1.0 indicates the vectors are opposite in their mapping of meaning. A cosine value of 0.0 indicates that the vectors are unrelated in meaning.

The LSA-based learning system provides a basis of knowledge from which to derive the meaning of other contexts or information not present in the original input corpus. This is done by projecting new contexts within the vector space leveraging the term meanings identified the cognitive model. Mathematically, a vector projection for a new context is the weighted sum of the component term vectors divided by the singular values. The formula for computing such a projection is:

$$projection = \frac{p^T U_k}{\Sigma_k},$$

where p is a term-frequency vector representation of the context to be projected weighted by *log-entropy* functions.

For more detailed explanation of the mathematical derivation of the LSA cognitive model see (Martin et al. 2016) and (Martin and Berry 2007).

2.1.2 Previous Attempts to use LSA for WSD

There have been previous attempts to use LSA in the construction of a WSD system. None of these have fully exploited the capability of LSA or even employed

it correctly. It is important to note that most sources claiming to use LSA for WSD are not using a LSA-based learning system but instead are generally referring to the application of the SVD to achieve dimensional reduction on a multi-dimensional dataset in an attempt to reveal latent relationships between the data items. Several researchers acknowledge that LSA has plausibility in discovering senses, but they either use another method (Van de Cruys and Apidianaki 2011), use LSA with other algorithms (Gaur and Jain 2013), use LSA with annotated or a priori data (Schumacher 2007, Katz and Goldsmith-Pinkham 2006, Gaur and Jain 2013), or use LSA on corpora that is too small and/or use too few dimensions to define the semantic space and construct a suitable learning system (Levin et al. 2006, Pino and Eskenazi 2009, Katz and Goldsmith-Pinkham 2006). In the papers by (Levin et al. 2006) and (Pino and Eskenazi 2009), even with the use of a small set of documents from which to derive different word sense meanings from contexts, the use of LSA in word sense disambiguation showed promise. Other published research has indicated that LSA would be good at finding the dominant sense for a target word but did not explore the use of LSA further (Pedersen 2007).

Other attempts to use LSA in this field have shown promise but were limited by the way in which LSA was applied. In one of the more significant works in this area (Schütze 1998), the singular value decomposition (SVD), which is a component of the LSA algorithmic approach, was used to discover some limited latent structures or meanings in a co-occurrence matrix in an attempt to detect underlying features that would better measure contextual similarity. Schütze's research in automatic word sense discrimination, the dividing of occurrences of words into groups based on whether or not they belong to the same sense, introduced the idea of context-group discrimination. This approach groups similar occurrences, or contexts, of ambiguous words into clusters. Words and context vectors are represented in a vector space model that is constructed by decomposing a word-to-word co-

occurrence matrix, made up of a selection of neighboring words to the target word being disambiguated, into reduced dimensional vector space using the same mathematical technique (SVD) employed by LSA. (This differs from LSA which decomposes the entire term by document matrix for a corpus to form the learning space as described in Section 2.1.1). The resulting word vectors obtained from Schütze's process are then used to form context vectors by computing the average of the word vectors corresponding to the words occurring in the context of the ambiguous word. Schütze clustered these context vectors to a predetermined number of context groups using the Buckshot algorithm such that each group would correspond to a sense of the word. This work is significant because it is the initial attempt to use a reduced dimensional vector space to address the problem of WSD with a clustering-based approach. Other follow-on attempts have been built upon this work with some success in limited applications (Pedersen and Kulkarni 2007, Pedersen and Kulkarni 2005, Pedersen et al. 2006, Pedersen et al. 2005, Kulkarni and Pedersen 2006, Purandare and Pedersen 2004a, 2004b, 2004c). Schütze recognized that contextual similarity plays a role in human semantic categorization where words are considered semantically similar because they are used in similar contexts. This resulted in the observation that "occurrences of an ambiguous word belong to the same sense to the extent that their contextual representations are similar" (Schütze 1998).

2.2 LSA-WSD Approach

This research hypothesizes that utilizing a LSA-based learning system for WSD (LSA-WSD) would yield improvements over previous unsupervised WSD algorithms and produce results comparable to human rater agreements for both discovering word senses and disambiguating target words. Some of the potential advantages of using a LSA-based learning system as the supporting technology for WSD is that it is fully automated, can be used for any word, could easily be

applied to another language, and is updateable as language usage changes. Additionally, the information obtained in the LSA-WSD process will also serve to characterize the knowledge represented in the LSA cognitive model being used.

The primary tasks for WSD are sense discovery and sense identification which are discussed in the context of LSA-WSD in Sections 2.2.1 and 2.2.2. This research will also explore the use of a modified clustering technique that takes advantage of the properties of the LSA cognitive model for relating semantically similar items (Section 2.2.3). All of this is combined into a process for inducing word senses and disambiguating individual words in context, forming the LSA-WSD system.

2.2.1 Sense Discovery

The first task in automated WSD utilizing the LSA-WSD approach is to determine the senses associated with a given word within an input corpus. The ideal input corpus would provide a general linguistic framework of meaning and would not be domain specific, although for different applications it could be augmented or targeted for a particular domain or to investigate a specific cognitive model.

Once the general LSA semantic space has been produced, individual input sentences would be mapped into this cognitive model for processing a target word. Based on the mapping within the LSA semantic space, sentences can be filtered by the impact of the target word in the sentence. This employs the notion of word importance within a sentence context as described in Section 3.2. Noted sentences where the target word has been identified as important to the meaning of the sentence can then be grouped as determined by the similarity of their contexts into different clusters. The hypothesis being that sentences where the target word makes a large difference in the meaning of the sentence will provide stronger indication of the sense of the target word for those contexts. The clustering approach used in this research is described in Section 2.2.3. Following

Schütze's hypothesis, these sentence clusters (sentclusters) should represent possible senses for the target word.

A second approach toward sense discovery within the LSA-based learning system is the analysis of synonyms to a target word using the same clustering technique as applied to sentences. This exploits the notion that each term vector in a LSA sematic space carries all the possible meanings or senses for that term (Landauer 2007). The challenge is then to separate these senses into individual identified senses. Examining the top *s* synonym words, the synonym clusters (synclusters) should represent possible senses for the target word.

The advantage of the unsupervised LSA-WSD approach for sense discovery is that it yields a true representation of the definition of a word with respect to the specific senses found in the learning system. This LSA-WSD method could be applied to any word, and the method could also be extended to different domains and languages while dynamically adapting to the content being analyzed. In this approach, there could be a different number of senses identified for a given word than in an annotated source such as WordNet (Felbaum 2012, WordNet).

2.2.2 Sense Identification

Once senses have been induced, the second task of the LSA-WSD work would be to distinguish the sense in which a word is being used within a particular context. One way this can be accomplished is by taking the sentence, or context, containing the target word in question, mapping it into the LSA semantic space, and then finding the closest cluster representing one of the previously discovered senses for the word. The closest cluster would serve to identify the sense of the target word, disambiguating the sense of the word in question. This is one of the approaches that is tested in this research, see Section 5.1. Other methods for

disambiguating the sense of a word in context using the induced sentclusters or synclusters are possible and are left as a subject for future research.

2.2.3 Semantic Mean Clustering Approach

For this LSA-WSD research a clustering model that is an integration of connectivity-based clustering and centroid-based clustering was developed for use in sense induction. There are many different clustering models and techniques available for use in categorizing data into groups around on some characteristic it exhibits. The most popular models are the connectivity models and the centroid models (other less common models include distribution, density, grid- based, etc.). (Aggarwal and Reddy 2014, Everitt et al. 2011).

Connectivity models connect data objects based on some distance measure, and the most prominent algorithm for connectivity-based models is hierarchical clustering. Hierarchical clustering can be divided into agglomerative methods and divisive methods and are represented by a dendrogram, which is essentially a tree structure. Agglomerative hierarchical clustering starts with all the n data objects each in their own cluster. Clusters are merged based on a distance measure in successive levels of groupings until all data objects are grouped into one cluster. Divisive hierarchical clustering is performed similarly in reverse. All objects start in a single common group which is successively split into smaller groups or clusters until all the data objects are contained each in their own separate cluster. For both methods the resulting number of clusters is determined by examining the dendrogram and picking a level of the hierarchy to use for cluster identification based on perhaps the similarity between the clusters at that level or some other criterion. The major drawbacks to these clustering techniques are complexity, with the best case being $O(n^2)$ and the worst case being $O(n^3)$. Also, once each

hierarchical grouping is made, the merging of data objects or splitting of data objects between clusters at that level cannot be modified (Everitt et al. 2011).

Centroid-based models for clustering involve iterative algorithms where n data objects are combined into a predetermined number of groups, k, based on the proximity of a data object to the cluster center, usually the centroid, of each group. The most well-known technique for this model is k-means clustering. The complexity of a k-means clustering algorithm is $O(n^2)$, and the selection of k is arbitrary. The resulting partitioning can be sensitive to processing order, the order in which each data object is compared to the centroids to determine to which group it belongs (Aggarwal and Reddy 2014). Also, because of the random selection of the initial groupings, it is difficult to reproduce the same clusters containing the same objects if the data is reprocessed.

To facilitate word sense induction using the LSA-based learning system in this research, a different approach to clustering vector mappings of sentences or words was explored, semantic mean clustering (SMC). This was driven by the desire to leverage the fact that cosine measurements in the LSA cognitive model represent a similarity of meaning and have certain thresholds and bounds that can be interpreted with some useful implications. A centroid model for representing the clusters was needed to generalize the meaning representation of the member items, whether they be documents, sentences, or individual terms. It is not necessarily important that the individual items have a close relationship with each other that is of interest for this application, but that they have a relationship to certain meanings as manifested in the centroid. Additionally, the meanings that are found should drive the number of clusters identified, not an arbitrary predetermined k number of groups. This also requires that cluster membership

for specific items may change as the cluster definition encompassed in the centroid evolves as each item is processed.

SMC is similar to k-means clustering in that it is based on centroids, however, it is not initialized with a random grouping of item clusters so the clustering outcome is reproducible. Also, there is no pre-defined value for k and the number of clusters is determined by the associations captured in the learning system. Like hierarchical clustering, SMC produces reproducible results and allows the data to decide how many clusters there are, but it differs in that items are assigned to a cluster by their similarity to cluster centroids not their measurement to other items. The SMC algorithm takes one input parameter, the cluster inclusion threshold, which defines the minimum similarity measurement required when comparing an item to the cluster centroid. Because of the constant recalculation of the cluster centroid, the cluster assignment for an item is not permanent until the entire process has completed. During processing an item can fall outside of the similarity threshold for its cluster as new items get added and the centroid for the cluster is recomputed. Multiple passes are required to reprocess any items that fall out of their initial cluster assignments. These are performed until all items have been assigned membership in some cluster. The complexity of the clustering used by the SMC algorithm is at worst case no greater than $O(kn)$, where $k < n$, but in practice only one fallout processing pass, $k = 1$, has been observed to be occasionally necessary.

The SMC algorithm using the LSA cognitive model is defined as follows:

1. For each item in the set of items (documents, sentences, or terms) being processed:
    a. Project, or identify, the item in the LSA semantic space.

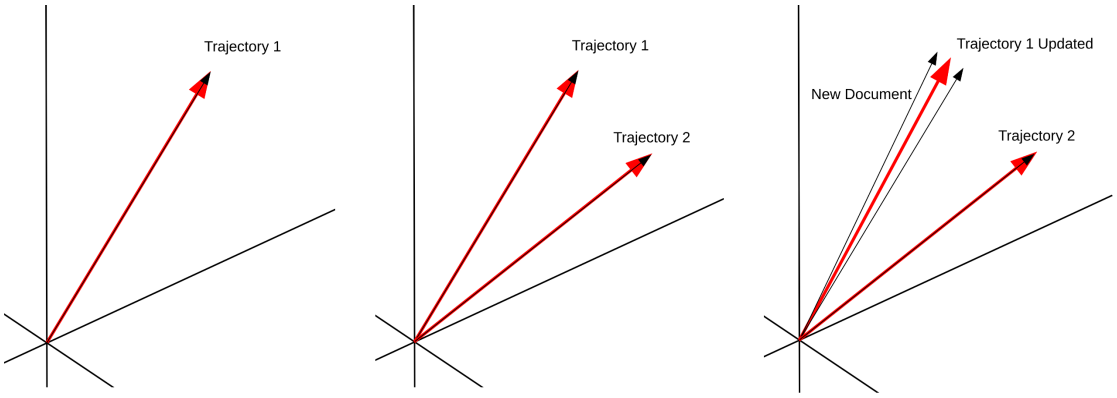

Figure 3: Steps in SMC cluster construction

b. Compare the item mapping to all identified cluster trajectories.
    i. If this is the first item, it will be recorded as the first cluster trajectory (see Figure 3 left).
    ii. If the item mapping is not close to, outside of the similarity threshold of, any previously identified cluster trajectories, then the item mapping is recorded as a new cluster trajectory (see Figure 3 middle).
    iii. If the item mapping is close to an existing cluster trajectory, within a certain similarity threshold, then it is grouped with items associated with the nearest cluster trajectory and the cluster trajectory (centroid) is updated by averaging all the member item mappings (see Figure 3 right).

2. After the full set of items is processed the items are examined to determine if any of them fall outside of the similarity threshold.
    a. If an item falls outside the similarity threshold for the cluster trajectory of which it is a member:
        i. The item is removed from the cluster group membership and the cluster trajectory is recalculated.

ii. The item is cataloged for reprocessing.

b. If an item is within the similarity threshold for the cluster trajectory of which it is a member, no action is necessary.

3. Repeat step 1 with the items cataloged for reprocessing and step 2 until no fallout items are identified in step 2.

The SMC process automatically determines the number and formation of the clusters where each cluster is a group of items falling within a set similarity threshold of the centroid and representing a consolidated semantic trajectory within the LSA cognitive model as depicted in Figure 4. These properties are used in the LSA-WSD system for word sense induction as described in Chapter 4.

2.3 Document Collection Sources

For this research two document sources were used. The first, a 3.5 million document corpus, where each document is a single paragraph sized text, was used to provide material for building broad based LSA spaces. The corpus

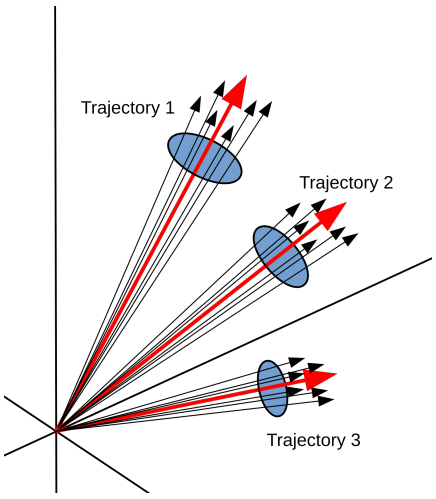

Figure 4: Clustered trajectories identified by SMC.

contains educational materials (books, periodicals, etc.) each noted with an associated Lexile level, which is an estimate of the reading difficulty level for the text (Landauer et al. 2011, Landauer and Way 2012). The texts in the corpus are representative of texts encountered by a typical learner of language over time (Biemiller et al. 2014, Landauer et al. 2011). They range from elementary to adult reading difficulty level and are made up of reading material often used in American schools. A LSA semantic space, or cognitive model, was constructed from a subset of documents from this corpus selected to include items ranging from elementary to adult difficulty level to serve as a basis for linguistic meaning. The goal is to implement the LSA-WSD system using a semantic space representative of the linguistic knowledge an average adult.

The second collection used was the Reuters Text Retrieval Corpus (RTRC) RCV1 collection (Lewis et al., 2004). This collection consists of over 800,000 English language news articles from the Reuters newswire which have been made publicly available for research use. The RTRC corpus consists of various articles published during the 1996-1997 timeframe spanning many different categories. Each news article is tagged with several category indices, including a general content category identifier and specific content sub-category indicators. A selection of articles was randomly picked from a narrowed group of articles from the Government/Social content category. This restriction was used to eliminate articles containing very little linguistic content such as financial market reports and other numerical financial data. This restricted group consists of 234,873 news articles.

These two collections were chosen for various reasons. As explained in Section 2.1, a LSA semantic space must be trained on sufficiently large and adequately representative text to learn language, understand the meaning of words, and represent a certain cognitive model. In this research, a general linguistic

knowledge of an average adult human was desired. The grade level collection allowed for the construction of a cognitive model representative of an adult human, containing texts with elementary or previous knowledge upon which an adult learns. Documents were selected from reading complexity levels from grade 1 – adult, with each level being equally represented or chosen. From this group there was enough content to select multiple, possibly nonoverlapping document sets to use as training text. The news collection also contains varied content deemed to be typical for adult reading level. These two document collections allowed for an adequate base LSA semantic space, a representation of a typical adult cognitive model (Martin 2016, Martin et al. 2016).

2.4 Document Sets

From the grade level and news document collections, a total of ten different document sub-sets were extracted. Six different document sets were randomly chosen from the grade leveled collection as two sets each of 150,000, 200,000, and 250,000 documents to create six different LSA-based learning systems. Additionally, since some duplication of content was allowed in the selection of these six sets, two more documents sets of 200,000 items each were randomly chosen from the grade level content with the restriction that they contain no duplicate documents between them. Two more document sets were selected from the news document collection each containing 200,000 documents.

Table 1 shows the size of each document set used as input texts. The objective in choosing these different training sets was to investigate whether size of the training set or variation in content would influence the overall learning and its effects on determination of word importance, sense discovery, and sense identification.

Table 1: Document sets used in finding sentences for a target word and creating LSA semantic space. Two different document sets were chosen for each corpus size. Two of the grade level document sets were specifically chosen to contain unique (non-shared) documents.

LSA Semantic Space	# Documents	# Unique Words
Grade Level A 150K	162777	141252
Grade Level B 150K	162845	141774
Grade Level A 200K	209365	162295
Grade Level B 200K	209423	162308
Grade Level Unique A 200K	196261	164940
Grade Level Unique B 200K	196262	164975
Grade Level A 250K	259847	182492
Grade Level B 250K	260059	182311
News A 200K	200000	254236
News B 200K	200000	255640

The document overlap ratio was calculated to verify the uniqueness of the documents chosen for the document sets used to create the LSA semantic spaces. The document overlap ratio reflects the degree to which two document sets include the same documents between them. It is defined as the number of common content documents between the two document sets to the total number of unique documents contained in both sets (Martin 2016). A low document overlap ratio indicates a lower amount of content is shared between the two sets. The document overlap ratio between the document sets for this research is shown in Figure 5. The two news document sets have a high document overlap ratio so they are expected to behave very similarly since the LSA semantic spaces for them are built using almost the same training text as input. The question to be examined is how do document sets, and their corresponding LSA spaces differ with respect to their

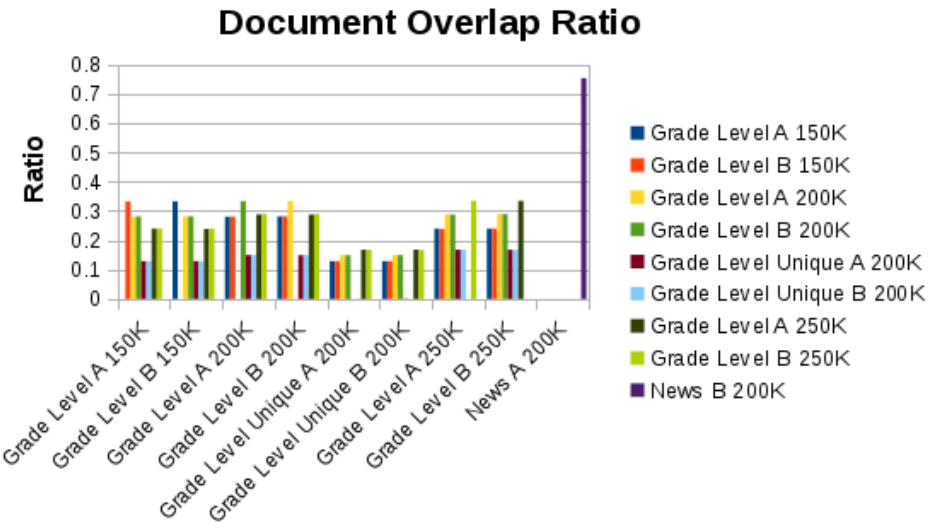

Figure 5 : This graph depicts the document overlap ratio between all the document sets used in experimentation. There was no document overlap between the grade level and newswire collections nor was there any overlap between grade level document sets that were selected specifically to have no shared documents between them.

sentences, the meanings of the words the sentences contain, and the word senses that are exhibited.

CHAPTER 3
WORD IMPORTANCE IN A SENTENCE

Discovering senses for a target word requires finding adequate contexts for the word. There are several units of context to consider, such as a phrase, sentence, paragraph, chapter, document, or a longer unit of text. A sentence is a syntactic unit, or a group of words, that expresses a complete thought. For a target word, a sentence should give good context for determining its meaning, unlike a whole paragraph that describes an idea or theme. A paragraph is another context unit, but it might not be specific enough or might contain too much varying information to distinguish the sense of a particular word. Of course, all sentences are not created equal in giving adequate context for a word. For example, even these short sentences “Oh shoot!” and “Shoot the basketball!” give different context and meaning to the word “shoot”. In the first sentence “Oh shoot!” could mean several things: disappointment, an exclamation over a mistake, or even a person telling another person to shoot something. The second sentence gives better context. One person or persons is telling another to shoot the basketball allegedly at the basket hoop. Determining which sentences to use to induce the senses for a word is a nontrivial decision and worth some investigation.

3.1 Identifying Sentences

The first task in determining which sentences to use for word sense induction is to identify the individual sentences contained in the documents that were used to build the LSA cognitive model. Breaking a document into sentences is a complex task. The obvious sentence breaks, the end of sentence punctuation markers “.”, “?”, “!”, do not always indicate an end of a sentence. This is particularly true when it comes to the use of a period. Periods can be found at the end of many different abbreviations such as titles (Dr., Mrs., Gov., Jr.), months, days, streets (St., Ave.,

Rd.), etc. Measurements often have periods after them: oz., kg., lbs. as well as corporate abbreviations: Inc., Co. Then, there are single letter abbreviations to consider as well such as Ph.D. Quotation marks also complicate the process of automatically finding a sentence break. A portion of this research for identifying word importance was the development of string parsing functions to automatically identify sentence breaks based on the various punctuation conditions that may be encountered in the text.

After documents were broken down into sentences, each sentence was then stripped of punctuation and parsed again to identify the actual words comprising the sentence. After processing all of the sentences for a document set, some statistics were gathered. The minimum, maximum, median, and average sentence lengths for each corpus were calculated along with the standard deviation of the average sentence length. Those results are shown in Table 2. For these statistics, sentence length specifies the number of unique words appearing in the sentence. Repeated words are only counted once. For example, the sentence “The girl ran down the street.” has sentence length of 5.

It is interesting to note that the sentences for the grade level document sets all have a minimum sentence length of one and about the same median sentence length, average sentence length, and standard deviation. The maximum sentence length for these document sets varies slightly between these sets. Given how close the results are for sentences in these sets for the minimum, median, and average sentence lengths, the variation in the maximum sentence lengths suggests the presence of atypical outliers for each document set. The same pattern is true for the sets of news articles where the maximum sentence lengths can be attributed to an atypical sentence occurring in the collection. The sentence in this case was taken from an article that had very little sentence punctuation.

Table 2: Sentence statistics for each of the document sets.

Document Set	Number of Sentences	Maximum Sentence Length	Minimum Sentence Length	Median Sentence Length	Average Sentence Length	Standard Deviation
Grade Level A 150K	1955690	152	1	8	10.4	8.1
Grade Level B 150K	1958077	146	1	8	10.4	8.1
Grade Level A 200K	2503308	152	1	8	10.5	8.1
Grade Level B 200K	2503697	142	1	8	10.5	8.1
Grade Level Unique A 200K	2309345	152	1	9	10.7	8.0
Grade Level Unique B 200K	2306918	129	1	9	10.7	8.0
Grade Level A 250K	3099118	146	1	8	10.5	8.1
Grade Level B 250K	3097901	152	1	8	10.5	8.1
News A 200K	2782399	2781	1	21	21.1	10.6
News B 200K	2781141	2781	1	21	21.1	10.7

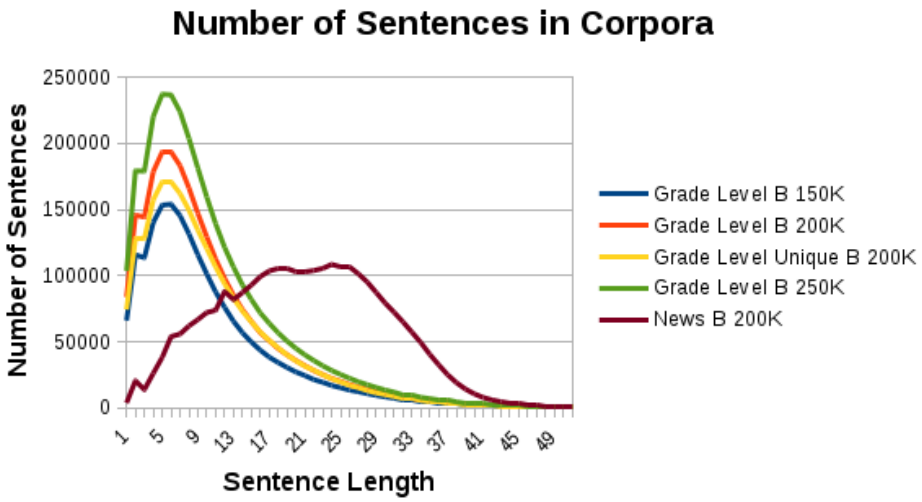

Figure 6: The number of sentences for sentence lengths 1-50 found the various document sets. Only the B sets are shown because the corresponding A sets are nearly identical.

Articles like this were generally a sports update with a list of scores where the content was all run together after processing for sentence identification.

The grade level document sets appear to have highly similar sentence lengths (See Table 2). The document overlap ratio between the sets was not above .35 for any of the sets which indicates largely different document content in each of the sets. The sentence lengths of the news articles differ quite a bit from those of the graded content sets. Figure 6 illustrates the number of sentences out of the total number sentences at various sentence lengths for each document set. A further look at the percentage of sentences in each of the different sentence lengths shown in Figure 7 reveals this same conclusion. This figure shows that the average or the median sentence length account for less than 8% of the total sentences for the grade level sets and 4% for the news sets. Over 90% of all the sentences in each of the grade level document sets and about 60% of all the

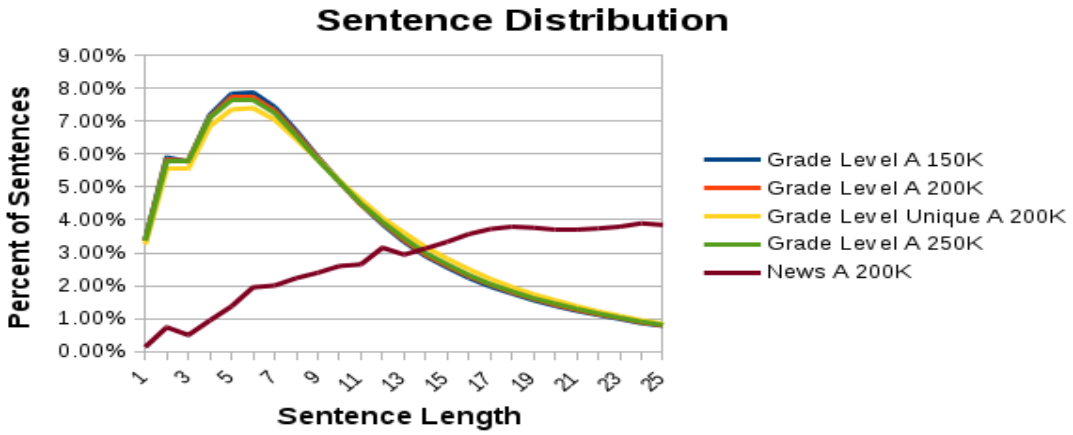

Figure 7: The percent of sentences at each sentence length for the various document sets. Only the A sets are shown because the corresponding B sets are nearly identical.

sentences in the news set have sentence lengths of 25 or less. For the grade level document sets, the observed sentence lengths and their distribution are similar regardless of the number documents present in the set. The number of sentences at any one particular sentence length is a small percentage out the total number of sentences for the set. The same is true for the news document sets. The difference between the document sources is the average and median sentence lengths. The news document sets have more words per sentence on average than the grade level sets. Varying the size of the document sets in the grade level collections had little to no effect on the overall distribution of the sentence lengths.

3.2 Determining Importance of a Word

The sentence clustering (sentclusters) approach to word sense discovery, introduced in Section 2.2.1 and the subject of Chapter 4, includes the notion of word importance within a sentence. The hypothesis for sense discovery is that

sentences where the target word makes a large difference in the meaning of the sentence will provide stronger indication of the sense of the target word for those contexts. Reasoning about word importance requires a means of quantifying the impact of each word in a sentence on the overall meaning of the sentence. There has been no notable research discussing the impact of individual words on sentence meaning to date. The LSA-based learning system provides a workable framework for making this measurement by using the additive analysis method mentioned in (Martin et al. 2016).

Using the additive analysis method to calculate the impact of the meaning of a word in a sentence containing that word is done by examining the underlying meaning captured in the LSA cognitive model. This determination involves projecting two versions of the sentence, one with the target word in place and one with the target word removed, into the LSA semantic space as two separate projection vectors. The cosine similarity is computed between the two projections, and this cosine is referred to as the cosine impact value (CIV). The CIV has an inverse relationship with the impact of a word on the meaning of a sentence. If the CIV is high, which is the case when the two projections are very similar, then the target word has minimal impact on the mapping of the sentence. This suggests that the word is less important to the meaning of the sentence. If instead the CIV is low, as when the two projections are dissimilar, this indicates the mapping of the sentence changes significantly when the word is removed; and therefore, the word impacts the meaning of the sentence to a larger degree. This in turn suggests that the word is more important to the overall meaning of the sentence. This can be visualized as in Figure 8.

To proceed it was necessary to establish an interpretation of the CIV that can be used as a general indicator of word importance. Ideally a specific CIV or range of CIVs could be identified as a threshold indicating at which point a word was truly

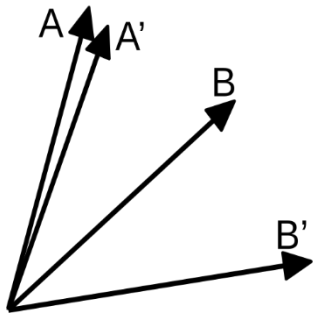

Figure 8: Depiction of the cosine impact on two sentences with a target word and without. Sentences A and A’ show a minimal impact from the removal of the word while B and B’ show a larger impact.

significant to the meaning of a sentence. For this step of the research the documents in each corpus were broken down into individual sentences. Sentences of length one were disregarded for further processing. In such cases the single word makes up the whole meaning of the sentence and does not give any context as to what the word means or in what sense it is being used.

The bulk of the sentences for the grade level document sets range in length from 2 to 19 (±1 standard deviations around the mean length of 10.5), while for the news document sets sentence lengths range from 10 to 32. The CIV was calculated for each word in each of the sentences with lengths in these ranges. The resulting 234,568,429 CIVs were then examined.

3.2.1 Effect of Sentence Length on Importance

Considering words and the corresponding CIV on their containing sentences, the question of whether sentence length has an impact on the importance of a word in a sentence arises. If so, then it is also necessary to determine if the effect of

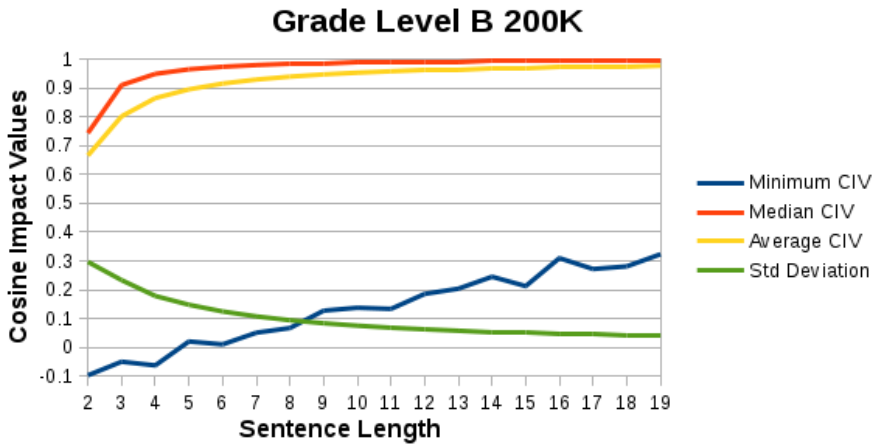

Figure 9: The minimum, median, and average CIVs for all words in sentences of lengths 2-19 for document set grade level B containing ~200,000 documents.

sentence length differs among the various cognitive models that were constructed. To address this question, an experiment was performed to evaluate the minimum, median, and average CIV for all words at each sentence length increment. In Figure 9 and Figure 10, examples of these values are shown along with the standard deviation for the mean cosine for two different document sets with approximately the same number of documents. The minimum CIV for the news source document set is a little bit higher, around 0.4, compared to the grade level set in which the minimum CIV is around 0.3. It can be noted that the minimum CIV does increase as sentences get longer, though this increase slows and levels off at the higher sentence lengths for all collections. Overall this implies that as sentence length increases the impact of any one word on sentence meaning decreases, as might be intuitively expected. The median and average CIVs are about the same for all collections at each sentence length. Sentences of shorter length, from two to four words, have lower average and median CIVs since each word carries more meaning in shorter sentences. As sentence length increases,

the average effect of any single word decreases. Additionally, the standard deviation decreases rapidly as sentence length increases, further suggesting that more of the words are impacting sentence meaning to a lesser degree as sentence length increases. The generalization of these results would indicate that while most words have a minimal impact on the meaning of the sentences containing them, some words do individually influence the meaning to a notable degree. These characteristics were observed to be similar across all the LSA semantic spaces (see results in Appendix).

To further analyze the distribution of the CIVs, the values at each sentence length were examined. The CIVs were grouped in 50 bins at a step interval of 0.02, counting the number of word appearances in each group. CIV values of less than 0.0 were included in the group for the first interval. The news document sets have fewer word appearances for the shorter sentences in this analysis because they contain fewer sentences at this sentence length. Most of the words at the sentence

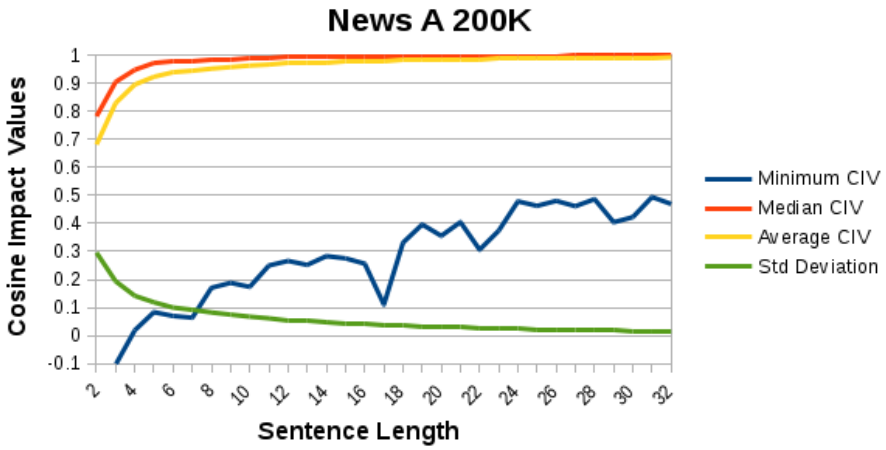

Figure 10: The minimum, median, and average CIVs for all words in sentences of lengths 2 to 32 for document set news articles A containing 200,0000 documents.

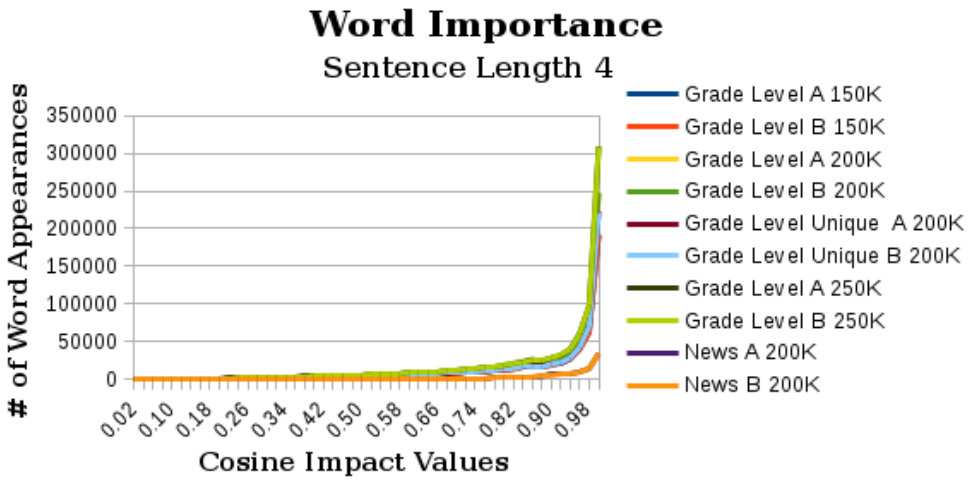

Figure 11: For each document set, the CIV for all the words in sentences of length four in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

length of two have CIVs of less than 0.92, suggesting that majority of these words have a large impact on the meaning of the sentence, as would be expected. Given a sentence of length two, when one word is removed the meaning of the sentence typically becomes quite different. For example, the sentence “Go Home!” has a different collective meaning than the individual words taken by themselves. “Go Home!” gives a command to walk or travel to one’s house. If the sentence is just “Go!”, it could mean that a person should go somewhere or imply a cheer for a team. The word “Home” by itself is not even a complete sentence and does not have much meaning other than to indicate a place by itself. The graph in Figure 11, showing the distribution of the CIVs for the different corpora at sentence length four, exhibits an inverted L shape except it is less pronounced in the news document sets which have longer sentences overall. The CIV distribution was examined for all sentence lengths. This same inverted L distribution pattern is exhibited for CIVs at all the other sentence lengths. It becomes more pronounced

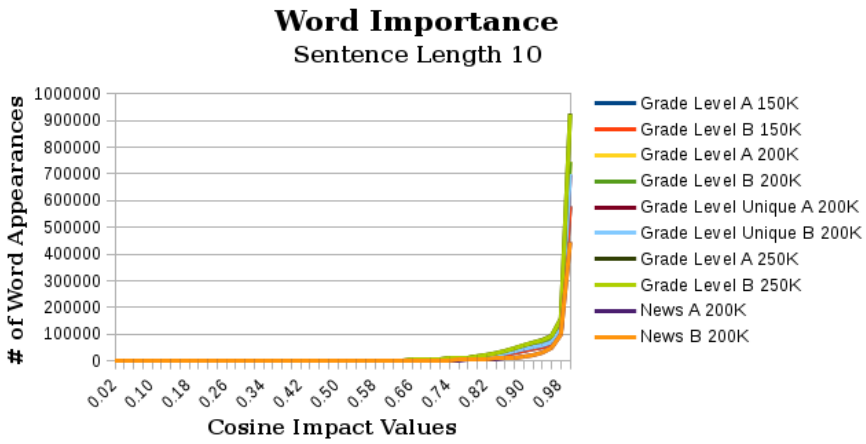

Figure 12: For each document set, the CIV for all the words in sentences of length ten in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

at increased sentence lengths as can be seen in the example distribution graph of CIVs for sentence length ten in Figure 12.

As noted previously with the average and median CIVs (Figure 9 and Figure 10) at the various sentence lengths, most words have a high CIV, indicating a low impact on overall sentence meaning across all sentence lengths. CIVs below 0.9 are rarely observed, accounting for fewer than 12.5% of all the CIVs computed. 75.2% of the CIVs were 0.96 or higher. On the line graphs of the CIV distributions this trend is observable as a common inflection point for all sentence lengths across all collections becoming more obvious at longer sentence lengths. Interestingly, at sentence length ten which is well below the mean sentence length of the news set, the CIVs are similarly distributed for all document sets (see Figure 12). Also, comparing the CIVs at different sentence lengths within a single

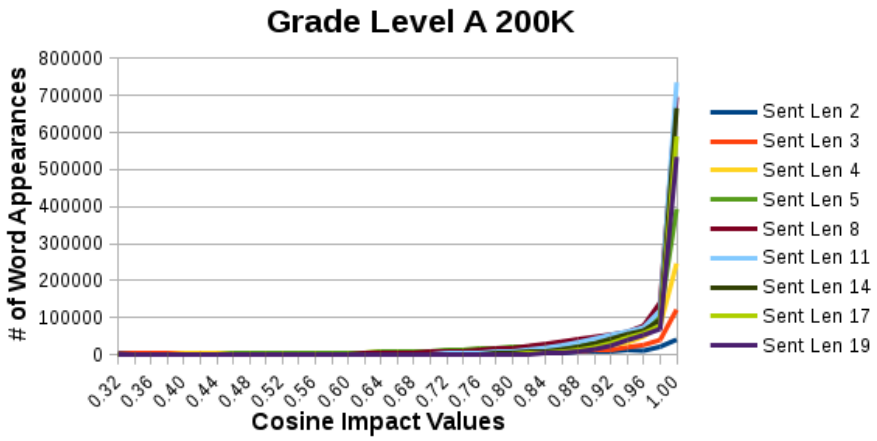

Figure 13: Distribution of CIVs for words at different sentence lengths for the grade level A document set.

collection again shows a similar distribution pattern. This can be seen for an example of the grade level document sets in Figure 13, and for one of the news collections in Figure 14. Graphs of the CIVs for all the sentence lengths across all collections and for individual collections can be found in the Appendix. For this research any word appearance with a CIV below 0.9 is considered to be an instance of an important word within the context of that sentence.

3.2.2 Word Characteristics Effecting Importance

A second question to consider when evaluating word importance is determining if there are specific words that never contribute notably to the meaning of a sentence. Likewise, it is desirable to explore the possibility that there are specific words that always have a large impact on the meaning of any sentences in which they occur. Either condition will help to refine the evaluation of word importance for use in identifying sentences that should be considered for sentclusters as part of the sense induction process. If a word contributes to the meaning of all the

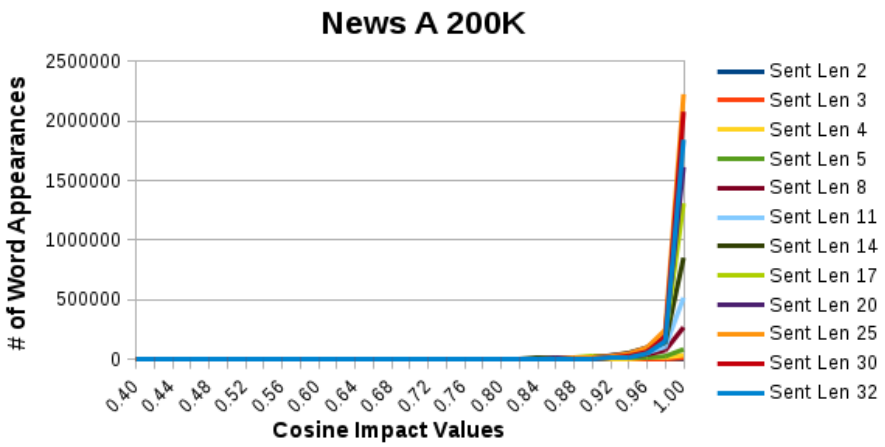

Figure 14: Distribution of the CIVs for words at different sentence lengths for the news articles A document set.

sentences including the word, then that word should have a group of sentences with sufficient context to distinguish word senses for the given LSA cognitive model.

To identify words that impact sentence meaning in a document set, the first step is a review of all the sentences that are associated with a target word. In any given document set there are generally some words that only appear in one sentence in the corpus. These words may impact sentence meaning but were disregarded for the purpose of this research as word senses cannot be found for them using sentclusters. For the document sets considered in this research, such words were rare, representing less than 1% of all unique words.

Examining the remaining words appearing in sentences with length greater than one, a small percentage, less than 7%, of the all the unique words in a corpus have a CIV below 0.9 for one or more sentences indicating that the word was important to the sentence meaning (see Figure 15). All these words, which can have an

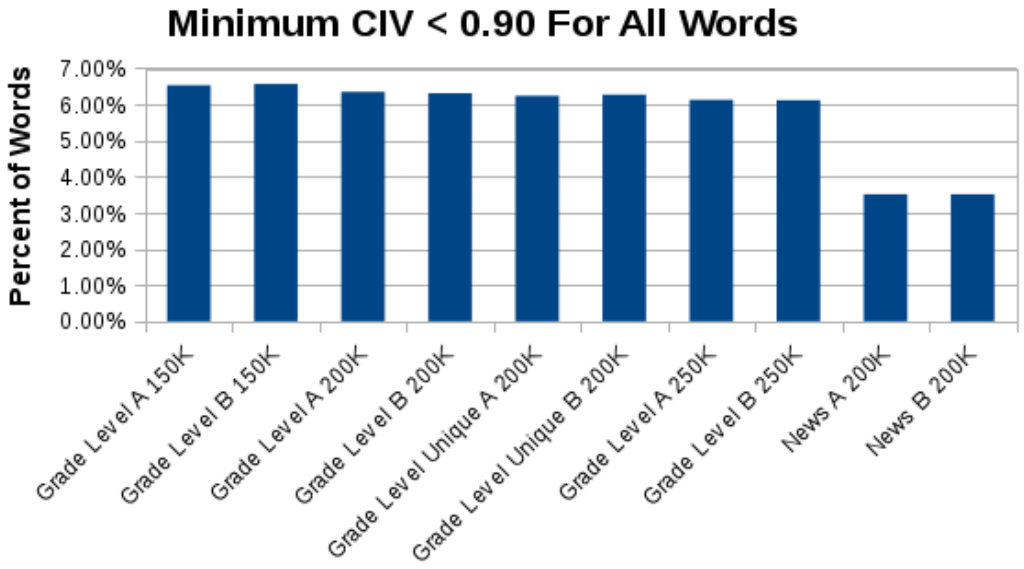

Figure 15: Examining all the unique words in each document set, only a small percentage of the words have significant CIVs on one or more of its sentences.

important effect on the meaning of a sentence, on average only show importance in less than 1% of the sentences in which they appear. Even fewer words, less than 1%, always have a CIV below 0.9 indicating the word is important in whatever sentence it occurs. Further investigation reveals that most of those words only appear in 1-3 sentences. These observations support the idea that there are a large number of words that never individually have more than a small impact on the meaning of a sentence, while there are a very small number of words that always have an important effect on the meaning of any sentence in which they appear.

The list of the 25 words with the lowest average CIVs, indicating that on average they have the most impact on the meaning of the sentences in which they are involved, consists primarily of proper names and a few verbs. Table 3 shows a

Table 3: A selection of words from the top 25 words that have the lowest average CIV, implying they have the best overall impact on sentences in which they appear.

Document Set	Words in Top 25 Set					
Grade Level A 150K	Annie	Abby	Arthur	Emily	cried	asked
Grade Level B 150K	Annie	Jack	Ben	Arthur	Emily	cried
Grade Level A 200K	earmuff	same	Annie	Jack	cried	asked
Grade Level B 200K	Sam	Annie	Abby	George	cried	asked
Grade Level Unique A 200K	Amy	Billy	Sam	supermen	Kate	asked
Grade Level Unique B 200K	Amy	Billy	Sam	Ben	Jack	asked
Grade Level A 250K	Sam	Amy	Ben	asked	cried	yes
Grade Level B 250K	Sam	Amy	Ben	Emily	asked	cried
News A 200K	page	Dole	nuclear	she	Vietnam	council
News B 200K	page	Dole	Taiwan	she	council	women

few of these words for each document collection analyzed in this research. This does not imply that these words always have an impact on the meaning of any sentence in which they are included.

The analysis in Section 3.2.1 shows that sentences of longer lengths do not have many words that contribute to the meaning of the sentence, therefore, just words that appear in sentences of length 4 to 10 were selected for further restricted analysis. Only approximately 45% of the overall words for the grade level sets and 27% of the overall words for the news document sets have sentences in this length range. Using just this subset of words and sentences, the percentage of unique words having a CIV less than 0.9 in one or more sentences is greater than was observed when considering all sentence lengths greater than one (shown in Figure 15), but is still small (see Figure 16). Even fewer of these words, on average, have

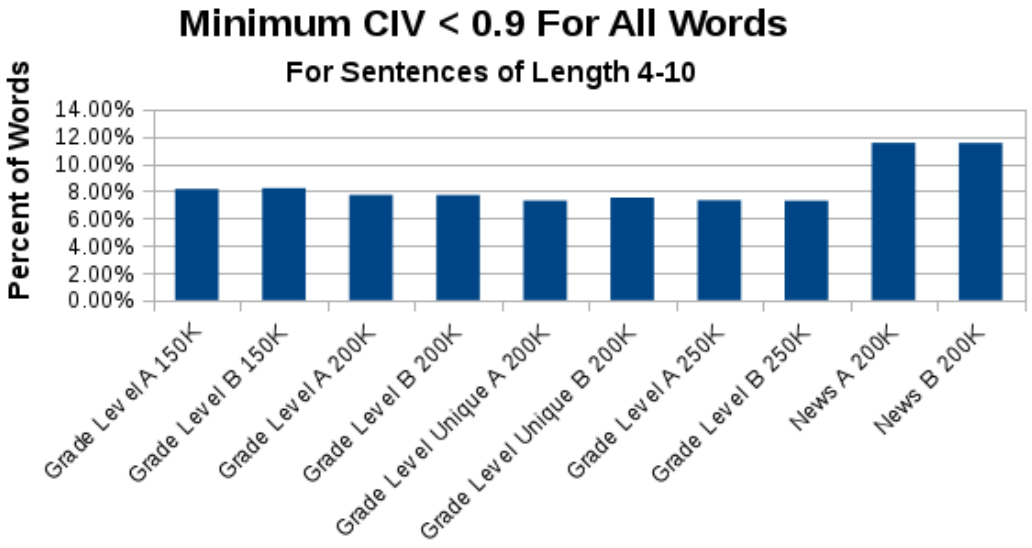

Figure 16: Percent of words that have sentences of lengths 4 to 10 that have an impact on the meaning on one or more of its sentences.

an important effect on meaning for all the sentences in which they occur, between 1% to 5% of the time, across all the examined corpora.

There are a very few words that always have an impact on the meaning of the sentences of length four to ten. Table 4 shows a few of them for each document set. It is notable to observe that these words are consistently nouns that are appearing in a considerable number of sentences.

All of these observations indicate that there is a relatively small portion of words that can have an important effect on the meaning of a sentence. Restricting the sentence lengths being considered marginally increases the proportion of important words, but overall less than 12% of the unique words in a corpus have individual importance in meaning. This limitation reduces the number of words that can be evaluated for sense induction using the sentcluster approach.

Table 4: Examples of words that are included in sentences of lengths 4 to 10 that always have an impact on the meaning of the sentences in which they appear.

Document Set	Word	# Sentences
Grade Level A 150K	Annie cat bear	2270 2029 1490
Grade Level B 150K	bear Jack Mike	1413 2744 1221
Grade Level A 200K	bear cat Sarah	1634 2358 1256
Grade Level B 200K	Billy bear cat	1349 1612 2370
Grade Level Uniq A 200K	Billy fish cat	1290 1437 1753
Grade Level Uniq B 200K	bear fish John	1338 1595 1356
Grade Level A 250K	Ben Billy cat	1972 1623 2724
Grade Level B 250K	bear cat John	1943 2747 1868
News A 200K	page UPS FBI	89 99 86
News B 200K	nuclear FBI Pope	152 76 122

3.2.3 Effect of Word Importance on Sentence Meaning

Given individual word importance measurements, the next consideration of this research was to characterize the appearance of important words in sentences. Each of the sentences was reviewed to determine the degree to which important words played a part in the overall meaning of the sentences, if important words appeared frequently in sentences and if there were sentences which contained no important words. In light of the compositionality constraint (see Section 2.1.1), it is expected that combinations of multiple words would have greater impact on sentence meaning, but it was unknown if individually important words would appear commonly in sentences.

Using sentences of length two and greater, all sentences were initially reviewed for the presence of important words, words with a CIV below 0.9. Sentences were counted where no words were individually important, all words were considered individually important, and where important and individually non-important words were mixed. The observed percentages of these groups out of the total number of sentences for each document corpus is illustrated in Figure 17.

In all of the grade level document sets most of the sentences, 85%, fall into the category of having both important and non-important words. Approximately 89% of these sentences have a length of four words or greater. In general, less than 4% of the sentences consist solely of individually important words. Of those sentences, approximately 98% are sentences of length two or three. Finally, 11% of the sentences in the grade level document sets contain no individually important words. These sentences were observed to contain four or more words 98% of the time.

The news document sets exhibit somewhat different characteristics. There are more sentences that contain no individually important words than was seen in the

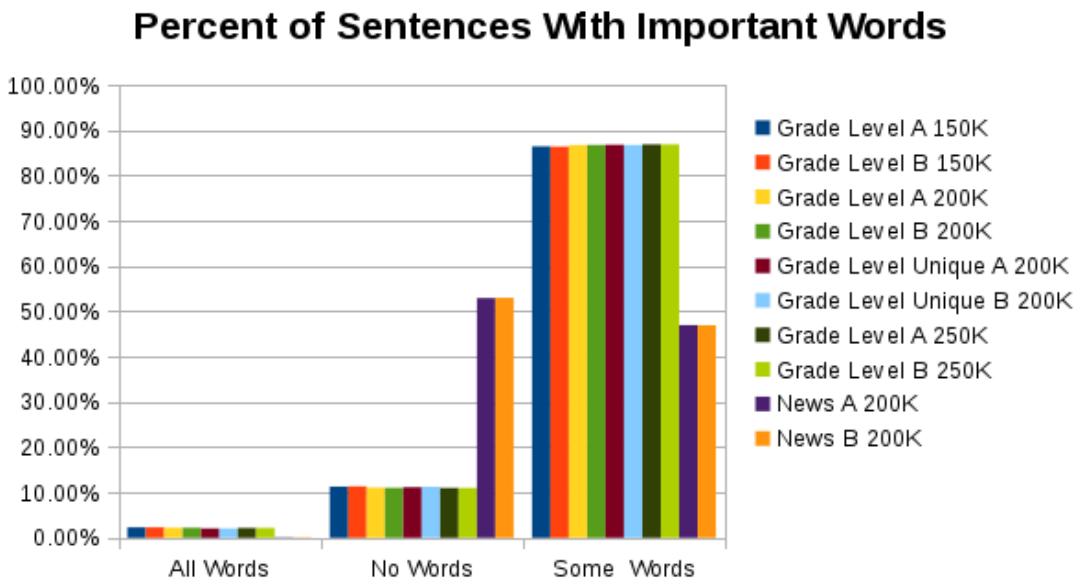

Figure 17: Percentage of sentences in which all the words, none of the words, and some of the words have an impact on the meaning of the sentence across document sets.

grade level spaces with 53% of the sentences being in this category. This is notably larger than the group of sentences containing both important and non-important words which only represent approximately 47% of the sentences. Less than 1% of the sentences in the news document sets consisted of words that were each individually important. These differences can be attributed to the observation that longer sentences have fewer individually important words since the combined meaning of other words in the sentence carries more weight than in shorter sentences. The news sets consist of much longer sentences than the grade level corpora as was shown previously in Figure 6. In the longer sentences of the news document sets, each of the words are contributing an individually smaller amount to the meaning of the whole sentence.

As Figure 17 illustrates, a large portion of the sentences across all the document sets have sentences containing a combination of individually important and individually non-important words. This confirms the notion that all words contribute a certain amount to the meaning of the sentence, albeit some more than others. For example, the sentence

I huff and puff and sit down with a thump.

found in the Grade Level A 150K document set has two words deemed important to the meaning of the sentence: “down” has CIV of 0.84 and “sit” has a CIV of 0.85. The rest of the words have CIVs above 0.90. Another example from the Grade Level A 200K document set

Beth laughs and claps her hands.

again has two words that measure as individually important to the meaning of the sentence: “Beth” with a CIV of 0.81, and “hands” which has a CIV of 0.6. As was previously noted in Section 3.2.2, words with high degrees of importance tend to be nouns. While parts of speech analysis is not considered as a component of this research it is an interesting observation.

3.3 Word Importance Observations

There were several notable observations from this portion of the research on word importance. The primary result was the identification of a CIV of less than 0.9 as the criteria for determining individual importance for a word based on its effect on the meaning of a sentence. Overall, relatively few words are individually important. Less than 7% of the unique words in the document collections were identified as important. There is a small set of words that are always important whenever they appear, and these words tend to be nouns. Even though the percentage of the vocabulary that classifies as important words is low, a general majority of

sentences do contain at least one important word. Sentence length varies widely across document collections, and there is no sentence length or group of sentence lengths that represent an overwhelming number of sentences. Sentences of less than four words tend to contain all important words, but this is expected and attributable to the limited meaning that is expressed in such short sentences. As sentence length increases, individual word importance decreases but there remain some instances of important words even in long sentences. It was also noted that corpus size and content did not have an observable effect on the word importance measurements in these experiments. The experiments and analyses presented in this chapter helped to guide parameters to use in word sense induction using sentclusters described in Chapter 4.

CHAPTER 4
WORD SENSE INDUCTION

Word sense induction (WSI), the automatic discovery of possible word senses, begins with a learning system, a system that will derive the senses for a target word. Therefore, WSI is only as good as the learning system upon which the induction is obtained. This research on WSI is based on the LSA derived learning system (see Section 2.1). The LSA-based learning system includes the input text from which to learn, the cognitive model represented by the LSA space, and the functions necessary to interrogate or probe the cognitive model.

4.1 Creating the Learning System

The two document collections used as input into the learning system for WSI are discussed in Sections 2.3 and 2.4. The grade level data source containing paragraph sized texts from educational books with noted reading difficulty levels was used to create a LSA cognitive model representative of a typical American learner of language. The other document source that was used as a basis for a comparable learning system was the Reuters newswire collection, news. Based on the results from the word importance experiments described in Chapter 3 where it was observed that the characteristics of the individual semantic spaces were very similar to the others selected from the same source, only two semantic spaces, one from the grade level source and one from the news source, were used for the WSI experiments described in this chapter. Each of the selected semantic spaces were built using approximately 200,000 input documents.

The ability to distinguish senses depends on the knowledge contained in the learning system. The cognitive model represented by the LSA semantic space is only as complete as the language usage reflected in the input text. Ideally a

learning system would discover all the possible senses for a word, but in practice that is difficult to achieve. Just as not every human has the same identification of senses for a word due to language development, content exposure, or other domain knowledge, the model can only represent what it has learned from the text to which it has been exposed. To this extent, automatic WSI based on a LSA learning system serves not only to derive possible word senses to be used in disambiguation or other applications, but also as an indicator of the knowledge contained within the learning system. Inducing senses from a learning system such as this can be useful in establishing senses related to a specific domain or a representation of a particular learner. In other words, the identified senses for a given learner can be used as an indicator for how well the person represented by the cognitive model knows the different meanings of a word. This research demonstrates the ability to automatically determine word senses using the LSA-based learning system and to what degree those sense determinations are reliable.

To facilitate this research, the WSI software component of the overall LSA-WSD system, was written to use a LSA cognitive model, a target word, and a clustering threshold, provided as input, to apply the sentcluster (described in Section 4.3) or syncluster (described in Section 4.4) approach and identify candidate senses for the target word. Several additional outputs are provided for analysis and are used to evaluate the performance of the sense discovery output. The system employs the SMC algorithm to derive the word sense clusters (WSC).

4.2 Clustering Hypothesis

The clustering results are highly dependent on the sentences being processed and the underlying LSA-based learning system being used. Ideally for WSI, the sentences would yield a reasonable number of clusters for a target word, and the

sentences would be evenly distributed across the individual clusters regardless of the number of sentences. Outliers would be an anomaly, but worth further examination because they could indicate obscure senses for the target word. In most cases, it is undesirable to have each sentence be its own WSC because this would result in the splitting of senses too finely to be useful, or many senses being repetitively identified in different clusters. It is not necessarily a major detriment to have multiple clusters classifying the same sense, but it is not desirable. Additional review would be required to determine that the multiple clusters represented the same distinct sense. It would also require more comparisons to find the closest WSC when trying to determine a sense in the subsequent disambiguation process.

In the same respect, having all the sentences in one cluster would indicate only one sense had been found for the target word. This is certainly plausible, but expected to be uncommon. If sentences are clustered into one or two clusters containing the majority of the sentences and many outlying small or singleton clusters for the remainder, then the WSCs are suspect. For these cases, the outliers might be noise and the big clusters will quite likely be multi-sense clusters. A low value for the average and median cosine of each sentence in the cluster to the centroid for these larger clusters would suggest just that. Multi-sense clusters are not useful because they do not provide a distinct identification of word sense.

4.3 Sense Discovery with Sentclusters

The sentcluster WSI experimentation began by using sentences where the target word was deemed important to the meaning of the sentence as described in Section 3.2. The premise being that these sentences would best identify the senses associated with the given target word. This hypothesis was compared to an approach using all sentences with the target word to form sentclusters and induce senses.

The first step in forming sentclusters is the identification of sentences containing the target word. This step is requires locating all sentences within a corpus containing the target word. Finding the sentences containing the target word has a task complexity of $O(n)$, where n is the number of individual word appearances in the corpus. A word appearance is simply an instance of a word being included in a sentence. This step was performed for each of the individual target words being analyzed and limited to examining sentences of four words or greater. Sentences of three words or less were eliminated from consideration due to the findings described in Section 3.2.1. Next, the identified sentences are filtered based on importance of the target word, resulting in two sets: one of all sentences containing the target word and a second with only sentences where the target word is important to the meaning of the sentence. Finally, the sets of sentences are clustered using the SMC algorithm described in Section 2.2.3.

4.3.1 Target Words

Eighteen words were chosen to be target words for the sense discovery experiments. This set, shown in Table 5, consists of words with coarse and fine-grained senses and includes mostly nouns and verbs with a few adjectives and adverbs.

Table 5: Target words used in experiments for word sense discovery and identification

Words 1-6	Words 7-12	Words 13-18
bank	interest	pretty
batch	keep	raise
build	line	sentence
capital	masterpiece	serve
enjoy	monkey	turkey
hard	palm	work

4.3.2 Initial Experiment

Sentclusering was performed for each target word on the set of all sentences containing the word (referred to as the “all sentences set”) as well as the filtered set of sentences where the target word was deemed important (referred to as the “important word set”).

It was immediately evident that using the all sentences set for sentclustering is unworkable because the process generates an extremely large number of clusters. Across the set of target words, the number of generated clusters ranged from hundreds to thousands, with the exception of the words *monkey* and *masterpiece*, which still produced sets of 45 to 80 clusters each depending on the semantic model being used. A large percentage of the clusters generated contained only one sentence. This condition was observed in 90% to 97% of the clusters generated with grade level learning system and 57% to 88% of those generated with the news learning system.

Using the important word set resulted in a substantial reduction in the number of generated clusters, in some cases producing only 1% of the number of clusters generated using the all sentences set. Even with the reduction in the number of clusters using the important word set, there still existed a high percentage of clusters with just one sentence, an average of 75% for all the words examined. These clusters were also not evenly distributed, with generally a few very large dense clusters.

4.3.3 Determining Appropriate Clusters

Trying to obtain an appropriate number and adequate sense representations for a given target word required several experiments. The first was to toggle the cosine cluster inclusion threshold. Varying the threshold from cosine values of 0.1 to 0.9

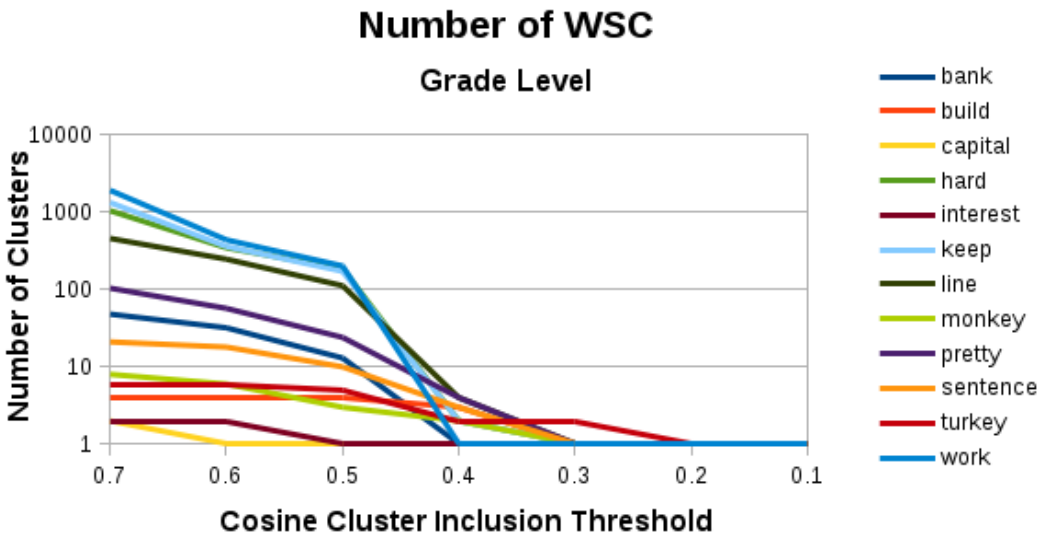

Figure 18: The number of WSCs induced from the important word set for the target word at different thresholds for the grade level learning system (log scaled).

in steps of 0.1, did yield a reduction in the number of WSC as the threshold was decreased as shown in Figure 18 and Figure 19.

Each learning system shows a significant drop in the number of WSCs formed around a threshold of 0.5 or 0.4. Specifically, there is a notable consolidation of the important word set into a few WSCs. This suggests that WSCs induced at those cluster thresholds should be candidates for identifying the different senses for the target word.

For all the different target words, the candidate WSCs induced using the sentcluster approach showed some promise, and some distinct senses were represented by the discovered WSCs. However, in each case there existed a multi-sense cluster which was typically rather large. Also, the number of WSCs

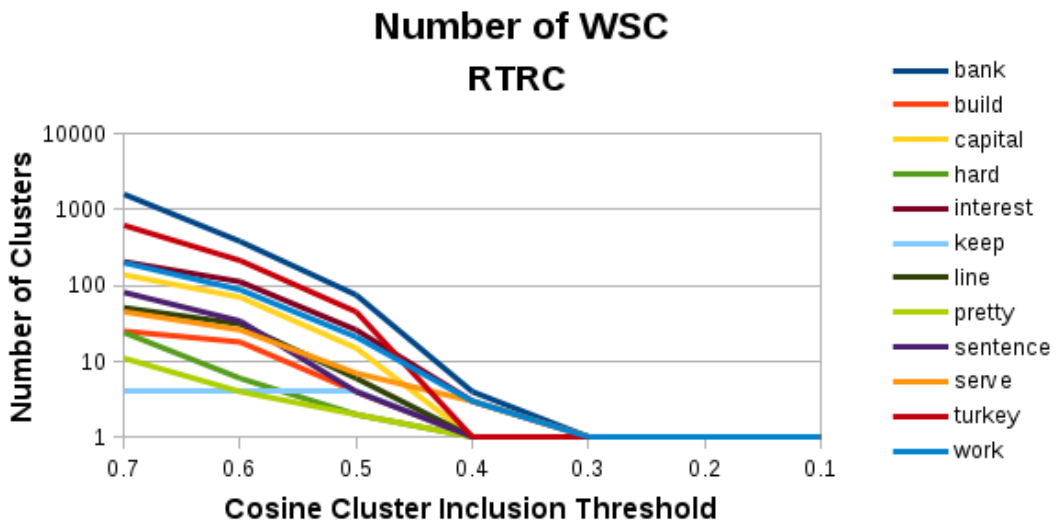

Figure 19:The number of WSCs induced from the important word set for the target word at different thresholds for the news learning system.

formed often seemed unreasonably high, with several clusters representing the same word sense when manually evaluated. For example, examining the target word *bank* and the corresponding sentences containing the word *bank* obtained from the grade level corpus, there were 88 important sentences found. Using the grade level learning system at thresholds of 0.9, 0.8, 0.7, 0.6, 0.5, and 0.4, groups of 75, 64, 48, 32, 13, and 1 WSCs were found, respectively. A reasonable number of WSCs was needed and setting the threshold to 0.5 accomplished that, but of the set of 13 clusters produced there were several that implied the same sense, and a multi-sense cluster was also produced. Table 6 shows the top sentences corresponding to a few of the 13 candidate WSCs for *bank*. The first three clusters are distinct senses for *bank*: 1) land alongside a river or lake, 2) financial institution, and 3) a container for empty bottles. The fourth cluster is a bit ambiguous but

seems most like the sense indicated in cluster 1 as it indicates a place. The last WSC in Table 6 clearly demonstrates a multi-sense cluster for *bank*. The most notable concern with this multi-sense cluster makes no differentiation between coarse-grain senses for *bank*. Bank as a financial institution should be distinguished from a piggy bank, river bank, and memory bank. However, this WSC is considered distinct by the learning system. This cluster does overlap slightly within the threshold of cluster 1 (centroid similarity of 0.63) and cluster 2 (centroid similarity of 0.53), but all other clusters are more distant. The senses represented by clusters 1 and 2 are also represented by the multiple senses observed in cluster 5.

Even when using a threshold of 0.90, which produced 75 candidate WSCs, a multi-sense cluster was still induced. This multi-sense cluster contained two sentences:

Table 6: Using the grade level learning system and a threshold of 0.5 for cluster inclusion, an example of five of the 13 candidate WSCs are listed for the target word ***bank***.

WSC #	# in Cluster	Example Sentences
1	1	Bits of broken shell lie on the sunny bank.
2	2	The bank was held up. The bank held Arncaster's mortgage.
3	1	She retrieved the shopping bags and hurried to the bottle bank.
4	1	They walked from bank to bank.
5	74	The Brickster was a bank robber. In the bank, Mark goes up to a teller.
		In my bank, one quarter goes CLANK. "My piggy bank," Slither said.
		There's one hiding in the bushes on the bank. She does a perfect cannonball from the mossy bank.
		Sunny squinted, searching her memory bank.

“Manaus is on the bank of the Rio Negro.” and “Is it a bank robbery.” Clearly these two sentences employ *bank* in a different sense.

Findings of the same type were observed with the news corpus. The important word set from that corpus was processed for the target word *interest* and used to induce word senses. The important word set consisted of 458 sentences using the word *interest*, and at thresholds of 0.7, 0.6, 0.5, and 0.4, groups of 205, 112, 26, and 3 WSCs were found respectively. Examining the three candidate WSCs induced using the 0.4 threshold, one of these was a multi-sense cluster of 456 items. Obviously, almost all the sentences were being associated with this cluster. Manual inspection of the candidate WSCs produced using the 0.5 threshold revealed the identification of some distinct senses, but again a multi-sense cluster was present containing 418 of the sentences. A few of the candidate WSCs for *interest* are shown in Table 7. The senses identified in clusters 1 through 4 exhibit both fine and course-grained senses for *interest*: 1) common concerns or views of an organization or group, 2) finance charge on a debt, 3) a self-benefit, and 4) a subject of fascination. Four example sentences from Cluster 5, the multi-sense cluster, are shown depicting different types of senses found in the cluster items.

While this approach was capturing word senses for the target words to a degree, the issue with these first experiments was that there existed one or more multi-sense clusters for all of the target words. These multi-sense clusters were usually rather large, on average 62% of the sentences for the grade level learning system and 78% for the news learning system. It is interesting to note that within the grouping of sentences within these large clusters was generally tight, all sentences had a cosine similarity within 0.65 to 0.88 of the centroid. In an attempt to separate the multi-sense clusters, as well as to reduce the number of candidate WSCs, an

Table 7: Using the news learning system and a threshold of 0.5 for cluster inclusion, an example of five of the 26 candidate WSCs are listed for the target word ***interest***.

WSC #	# in Cluster	Example Sentences
1	9	It will require the wisdom of Solomon to persuade both wings to unite in the party's interest. Neither the national interest nor the party interest require us to ride roughshod over views deeply and sincerely held.
2	2	Further, it is proposed to treat as rent 15 percent per annum as imputed interest on interest free or concessional interest-bearing deposits in excess of six months received from the tenant. They will also earn a 10 percent annual interest.
3	1	Chinese self-interest is also a powerful motivation, Murckart argues.
4	1	There's also more interest in red flowers other than roses.
5	418	Who has any interest in reigniting this quarrel? It also rejected a suggestion that one of its MPs, Tutekawa Wyllie, had a conflict of interest in the matter.
		The interest of the country should prevail over the interests of a government.
		SEC is asking that Colin disgorge the funds, including interest.

additional series of experiments were performed as follows using the important word set:

1. Perform clustering based on a larger grade level learning system with more documents (315,565) as input text for inducing WSCs.
2. Perform clustering using a different source of sentences with the target word than the input to the learning system, i.e.: using the important word set from the news corpus with the grade level learning system.

3. Perform clustering using an augmented sentence vector with more context by including the sentence before and after the sentence with the target word.
4. Perform clustering using the sentence with the target word removed.

The reason for creating a larger grade level learning system in experiment 1 was to give more overall context upon which to derive the WSCs. The SMC algorithm did produce better results in this case, but multi-sense clusters were still induced. In experiment 2, clustering the sentences from a different source did not yield any better results. Augmenting or adjusting the sentences as described in experiments 3 and 4 were an attempt to give more context for the target word (experiment 3) and to make the meanings of the sentences more distinct (experiment 4). This was done in hopes of making the projection of the important sentence better identify the sense of the target word. These experiments did not produce notably different results. There was more refinement of the individual WSCs, but multi-sense clusters were still being regularly induced. It was observed that sentences tended to be grouped upon the meaning of the target word and one other dominant word in the sentence or the sentences were generally ambiguous in their use of the target word.

Overall the sentclustering approach, while showing some promise, was unable to separate some items into clearly distinct sense clusters. Multi-sense clusters were present in almost every case.

4.4 Sense Discovery with Synclusters

After the sentcluster experimentation, the syncluster approach, described in Section 2.2.1, for sense discovery was tested. Synclustering examines the actual target word and its meaning within the learning system by examining words close to the target word within the LSA semantic space, its synonyms. This was done

to explore specifically how the meanings of the target word are represented in the cognitive model. Within the LSA vector space, every unique word corresponds to a term vector (see Section 2.1). Embedded in that vector are all the senses of the word learned from the input text (Landauer 2007, Landauer and Dumais 1997). Synclustering is predicated on the idea of attempting to separate the senses for a target word embedded within its term vector by clustering the synonyms of that word based upon their similarity measurement with each other. Within the LSA cognitive model, words close to each other are considered synonyms, but they are really just closely related and may not all be true synonyms in that they do not have the exact same meaning. These "synonyms", or closely related words, indicate what associations of meaning exist for the target word within the learning system.

To perform synclustering, all the words present in the LSA-based learning system are examined to find the words closest to the target word using the cosine measurement for similarity. These top k words are then clustered using the SMC algorithm to produce candidate WSCs for the target word. Once the candidate WSCs are induced, the closest word to the centroid of each of the clusters is taken to be the identifier, or descriptor, for the cluster. For the synclustering experiments, the WSI software takes an additional input parameter for the k number of terms to cluster in addition to the other input parameters. This parameter was varied to test the clustering with 100 and 200 words. Synclustering was performed for the same eighteen target words examined for sentclustering and listed in Section 4.3.1. Only two cluster inclusion thresholds were tested, 0.4 and 0.3, based on experiments with sentclustering thresholds. These thresholds yielded good results and testing of other threshold values was left for future work. Results for five of these words are discussed in Section 4.4.1, with the remaining results being included in the Appendix.

Table 8: The top 100 closest words to the word ***bank*** in the grade level learning system.

Terms 1-20	Terms 21-40	Terms 41-60	Terms 61-80	Terms 81-100
bank	levee	riverbed	riffles	waterfall
banks	gorge	barges	snags	waded
downstream	flatboat	paddled	money	overhanging
riverbank	bend	tributaries	shallows	crossing
upstream	boatmen	thames	creek	sandbars
river	canoe	midstream	conononka	portage
rapids	steamboat	canal	savings	bills
downriver	footbridge	countercurrents	flowing	swift
dam	flood	monongahela	bottomland	sawmills
upriver	ferrymen	paddle	creeks	paddling
bridge	dammed	reeds	watercourse	mississippi
flowed	bottomlands	cash	poled	damming
current	sandbar	ferryman	wading	meander
raft	flatboats	boatman	riverside	murky
tributary	robb	riverbanks	narmada	platte
barge	stream	dams	rhadamnanthus	riverboat
steamboats	deposit	rafts	cocytus	uminpeachable
muddy	loan	headwaters	radarscope	potomac
eddies	willows	silt	insecttortured	marshy
bluffs	nashua	poling	shallow	spanned

4.4.1 Finding Synonyms and Producing Sense Clusters

Using the grade level learning system, the top 100 closest words, or synonyms, to the target word *bank*, are shown in Table 8 prior to clustering in the order of their proximity. All the words are lowercased, including proper nouns, because casing is not considered in the text processing. Simply inspecting the words manually, it is interesting to note that the top 36 words are associated in some way with a riverbank or a body of water. The first term that is not clearly in that sense category is the word *deposit*. This word is ambiguous because it could be used as a money deposit or a deposit on the bank of a river. Some of the term associations are not immediately obvious, however, the words such as *Nashua*, *Thames*, *Monongahela*, *Conononka*, etc. are actually proper names of rivers and should be associated with the river sense of *bank*. The only other terms in the list that would

Table 9: The WSCs discovered using synclusters on the top 100 synonyms for the word ***bank*** in the grade level learning system.

Word Sense Cluster	Number in WSC	WSC Descriptor	Next closest words	Cosine Between bank and WSC Centroid
WSC 1	93	downstream	river rapids upstream riverbank	0.78
WSC 2	6	money	bills cash savings loan	0.51

not be associated in some way with a river or riverbanks are *loan* (ranked 38th), *cash* (ranked 52nd), *money* (ranked 63rd), *savings* (ranked 67th), and *bills* (ranked 87th). These terms suggest an association of money or money related items with the word *bank*. There are two definite senses for the word *bank* that emerge just from manual inspection of the list. This suggest that synclustering might be expected to induce two senses from this list.

Using the grade level learning system, the synclustering of these 100 words yielded the results that are shown in Table 9. Two distinct clusters were induced for *bank*, one for the "riverbank" sense and one for the "money" sense. Both clusters have centroids that are relatively close to the target word *bank*, with cosines of 0.78 and 0.51. A cosine of 1.0 would indicate that the mapping of the associated cluster is identical to the mapping for the target word. The values for these candidate WSCs indicate that the cluster for the river sense is more closely associated with *bank* than the cluster for the money sense by this learning system.

Table 10: The WSCs using synclusters when clustering the top 100 synonyms for the word ***bank*** in the news learning system.

Word Sense Cluster	Number in WSC	WSC Descriptor	Next closest words	Cosine Between ***bank*** and WSC Centroid
WSC 1	88	banks	banking deposits bankers lending	0.78
WSC 2	9	rates	interest reserve mortgage discount	0.36
WSC 3	1	finance		0.21
WSC 4	1	manages		0.21

Synclustering with the news learning system exhibited somewhat different results for the word *bank*. The synonyms for *bank* in the news learning system were different than the ones identified in the grade level system. With the same parameters, using the top 100 synonyms to *bank* and a cluster inclusion threshold of 0.3, there were four candidate WSCs induced as described in Table 10. All four of them relate to the financial institution sense for the word *bank*, but only one, cluster 1, shows a strong association with the target word. The other three, with cosine similarities of 0.36 and 0.21 are more distant, suggesting that they may not represent senses for *bank*. Examining the descriptors suggests that there is really the only one primary sense for the *bank* that is understood by this learning system, though the second cluster could be taken to indicate a sense related to mortgages which would be a fine-grained sense of the word.

The next target word to be studied was the word *palm*. Using the grade level learning system, the top 100 synonyms, and a cluster inclusion threshold of 0.3, there were twelve candidate WSCs induced for *palm*. Centroids for three of the

Table 11: The WSC results for using synclusters on the word ***palm***.

Grade Level Learning System			News Learning System		
WSC Label	# in Cluster	Cosine to Centroid	WSC Label	# in Cluster	Cosine to Centroid
hand	29	0.65	beach	3	0.55
trees	40	0.50	cigarette	96	0.47
gripping	5	0.43			

candidate WSCs had a cosine similarity with the target word *palm* that was greater than 0.4, while the remaining ones had cosine similarities of less than 0.31. This suggested that there were three senses associated with *palm* by the synclustering process, as shown in Table 11. Upon inspection, it was observed that they correspond to the three coarse-grain senses of the dictionary definition of *palm*: 1. An unbranched, evergreen tree, 2. Inner surface of a hand between the wrist and the fingers, and 3. To hide or hold something in one's hand (English Oxford Living Dictionaries n.d.). It is worth noting that the first two senses correspond to the noun part of speech for *palm*, while the third corresponds to the verb part of speech for *palm*.

The candidate WSCs induced for *palm* using the news learning system were different. Only two clusters emerged using the same parameters. One cluster appears to correspond to *Palm Beach* in Florida. This is not surprising as the cognitive model was constructed from news articles which tend to mention locations often. The other cluster has an interesting label of *cigarette* indicating something to do with a hand. Other words in the cluster are *tobacco*, *smokers*, *smoking*, *Lorillard* (name of a Tobacco company), and other names of people. It appears to have associated the meaning of *palm* with smoking.

For the target word *sentence*, synclustering produced only one candidate WSC for both learning systems. The number of synonyms used for clustering was also

Table 12: The WSC results for using synclusters on the word ***raise*** where the cosine similarity between the cluster centroid and the target word is greater than 0.35.

Grade Level Learning System			News Learning System		
WSC Label	# in Cluster	Cosine to Centroid	WSC Label	# in Cluster	Cosine to Centroid
money	71	0.57	increases	11	0.58
raised	2	0.55	funds	4	0.50
crops	6	0.50	tax	26	0.48
support	6	0.37	interest	4	0.38

expanded to the top 200 with no change in results. Interestingly, the grade level learning system produced a candidate WSC with the descriptor of “spelling”, and the news learning system produced a candidate WSC with the descriptor of “prison”. The centroid cosine similarity to the target word was high, greater than 0.96, for both cases. The grade level learning system learned the definition of *sentence* as a group of words written together expressing a complete thought. The news learning system learned the definition of *sentence* as a punishment assigned to a person in court.

Further experimentation using the target word *raise* yielded some interesting candidate WSCs as shown in Table 12. Although 9 and 27 WSCs were generated respectively for the two learning systems, upon inspection only the ones with a cosine greater than 0.35 were considered as possible valid clusters. In both cases this left four candidate WSCs. The first three senses for *raise* within the grade level learning system all have comparable cosine similarities between the cluster centroids and the target word. This observation indicates different, possibly fine-grain senses, with the same degree of association to the target word *raise*. The grade level learning system captures a sense of “money” for *raise*. This cluster suggests an increase in money, such as taxes, funds, or salary (those three words

were found in the cluster). The other WSCs indicate the sense of raising children, growing crops, and raising support for something.

In the news learning system the candidate WSCs appear to indicate slightly different senses. The first sense “increases” is a meaning of *raise* associated with growth, and the other words in the cluster suggest that as well. The next two WSCs seem refer to raising money. Observing that the corresponding cosines between the centroid for those two WSC and the word *raise* are about the same suggests that they represent the same sense, and upon inspection of the other words in the clusters this appears to be true. The fourth WSC, labelled “interest”, remains somewhat ambiguous in that it could refer to money or attention. The other three words in the cluster do not serve to clarify the sense. Overall, while the senses found for *raise* may not be exhaustive, both learning systems found some distinct senses and exhibited a different learning of meaning between them.

Examining the candidate WSCs induced for the word *line* revealed twelve clusters using the grade level learning system and twenty-four clusters with the news learning system. The news learning system did not find any clusters with a cosine similarity between the centroid and the target word of more than a 0.33. Reviewing the top candidates indicated no clear sense for *line* in the cognitive model constructed from the news input text. However, the grade level learning system induced some strong senses for *line*. Of the twelve candidate WSCs were generated, and five of them had a cosine similarity between their centroid and the target word that exceeded 0.43, as shown in Table 13. A manually produced description of the sense is shown that was derived by considering the other words associated with each cluster.

Table 13: The WSC results using synclusters on the top 200 words for the word ***line*** using the grade level learning system.

WSC Label	# in Cluster	Cosine to Centroid	Manual Description of the Sense
zone	137	0.66	A line marked on a field or court that relates to the rules of a game or sport like a goal line or zone line
assonance	6	0.63	A line of poetry
bait	21	0.53	A line on a fishing rod
horizontal	18	0.49	A mathematical term for a line or lines in particular directions
ahead	7	0.44	A line marking the starting or finishing point in a race

4.4.2 Synclustering Observations

For all eighteen of the target words examined, synclustering produced reasonable results reflecting the word knowledge contained in the LSA semantic spaces that were used in the sense induction process. Not all words resulted in clearly identifiable word senses, but several such as *bank*, *palm*, *sentence*, *raise*, and *line* showed promising results with clearly usable sense identifications for the subsequent task of word sense disambiguation. The empirical evidence suggests that candidate WSCs should have a cosine similarity between the centroid and the target word that exceeds 0.35 to be considered as a sense cluster for the word. Coarse-grained senses appear to be more easily identified, as would be expected, but some fine-grained senses were induced such as with the word *line*.

Testing different threshold parameters and selection quantities for the input synonyms led to the identification of a 0.3 clustering threshold and a selection of 100 synonyms as the best performing configuration. Further research will be needed to refine these input criteria to optimize results. It was apparent in all cases that the two learning systems were not equal in their represented word knowledge.

While they showed some agreement about the senses identified for different words, as was the case with the word *bank*, there were some striking differences such as with the word *sentence* or *line*. Overall the grade level learning system produced more broad-based and consistent results. These observations demonstrate the ability of the WSI process to interrogate the cognitive model being used for sense induction. Synclustering for automatic word sense induction provides a means to induce senses as well as examine and analyze the underlying LSA learning system.

CHAPTER 5
AUTOMATIC DISAMBIGUATION

After senses for a word have been induced, the next step of the LSA-WSD process is to disambiguate target words or determine in which sense a target word is used given a particular context. This sense identification task involves using the induced senses from the WSI task described in Chapter 4 with an automated system to evaluate the sense of the word used in the input context. Given the WSI results discussed in Section 4.3 and Section 4.4, the decision was made to focus on the senses produced with the synclustering approach described in Section 4.4. Using two different methods, the sentence for a given target word was compared to each syncluster to determine which cluster was the closest. This sense was taken to be the identified word sense for the target word in the given context.

To automate this task, a word sense disambiguation (WSD) software was created as part of the overall LSA-WSD system built for this research. The software was written to use either the sentcluster or syncluster sense definitions produced by the WSI software. These, in addition to the input of a target word and a context sentence to be disambiguated, were provided as input to the WSD software, which then determined the sense of the target word as used in the context sentence.

5.1 Methodology

Two basic methods were examined for use in WSD to determine which syncluster identified the correct word sense for the sentence. The hypothesis behind focusing on the syncluster definitions of word senses was that the synonym clusters would serve to break out the various meanings or senses carried by the target word. These synclusters were each described by a centroid vector that maps the average

of all the component vectors in the cluster and by a representative word that corresponds with the closest component vector to the cluster centroid.

The first WSD method tested, the synonym replacement (SR) method, identifies the sense by replacing the target word in the context sentence being evaluated with the synonym identifier for each of the previously induced clusters. The modified context sentence is mapped into the LSA semantic space and then compared to the mapping of the original sentence to measure the amount of change produced by replacing the target word. The expectation is that the best sense synonym for the use of the word would result in a mapping that was most similar to the mapping of the original sentence containing the target word. This similarity is again measured using the cosine similarity between the two mapping vectors.

The algorithm for the SR method is as follows:

1. For the input target word and sentence, keep the original sentence, *sentA*.
2. For each syncluster:
    a. Make a copy of *sentA* into *sentB*.
    b. Replace the target word in *sentB* with the synonym identifier for the syncluster.
    c. Project *sentA* and *sentB* into the LSA semantic space.
    d. Compute the cosine similarity measure between the projections for *sentA* and *sentB*.
    e. Record the cosine similarity.
3. Determine the highest cosine similarity seen and take the corresponding syncluster as the identified word sense.

The second WSD method explored, the context comparison (CC) method, takes the given context sentence with the target word removed and identifies the closest

syncluster centroid. The idea behind this method is that the context words surrounding the target word carry the information suggesting the sense of the target word within that usage. In the LSA semantic space, the mapping of of the sentence without the target word is compared to the mapping of each syncluster's centroid using the cosine similarity measure. The syncluster with the highest cosine is taken to correspond with the sense that best represents the way the target word is being used in the sentence.

The algorithm for the CC method is as follows:

1. Copy the input sentence *sentA* into *sentC*.
2. Remove the target word from *sentC*.
3. Project *sentC* into the LSA semantic space.
4. For each syncluster:
    a. Compute the cosine similarity measure between the projection for *sentC* and the centroid vector for the syncluster.
    b. Record the cosine similarity.
5. Determine the highest cosine similarity seen and take the corresponding syncluster as the identified word sense.

5.2 WSD Experiments

The experiments for the WSD task were based upon the grade level learning system. The target words, and their corresponding word senses determined by synclustering described in Section 4.4.1 were used in these experiments. The target words were *bank*, *palm*, *line*, *raise*, and *sentence*. Because only one sense was induced for the word *sentence*, it was not used in these WSD experiments as there would be no effective selection to examine. Instead, the word *serve* was examined with the candidate WSCs induced using the grade level learning system shown in Table 14.

Table 14: The WSC results using synclusters on the top 100 words for the word ***serve*** using the grade level learning system.

Grade Level Learning System			
WSC Label	# in Cluster	Cosine to Centroid	Manual Description of the Sense
meal	8	0.61	Present food
prepared	3	0.42	Function in a specified way
public	68	0.49	Perform a duty or service
accept	9	0.42	Favorable or adequate, use of satisfying

To evaluate WSD performance, test sentences for each of the five target words were randomly selected from different sources and then hand annotated to identify the correct word sense used in each sentence. Sources for the sentences included the important word set for each target word, the grade level corpus, online dictionaries, and WordNet (Felbaum 2012, WordNet). Sentences were further selected to ensure coverage of all the different senses that had been previously induced by synclustering for each target word. Other sentences containing different senses than those represented by the synclusters for each word, as well as ambiguous sentences, were also included in the test set. This resulted in a test set of twelve or more sentences for each of the target words with a human generated correct word sense identification.

Test sentences for the words *line* and *raise* can be seen Table 16 and Table 15, respectively. The remainder of the test sentences for each target word are shown in the Appendix. Information on the induced word senses for *line* and *raise* can be found in Table 13 and Table 12 in Section 4.4.1.

For each target word, each of the test sentences matching an induced sense was submitted to the WSD software, and a sense was identified for the target word. Both the SR and CC methods were tested, and the results were compared to the

Table 16: Test sentences used in the WSD task for the word ***line*** and their annotated sense determined by a human rater.

Annotated WSC Label	Sentences Using ***line*** in this Sense
zone	Jackie stepped to the line and dropped in both foul shots. Jim plowed forward to stop the quarterback from reaching the goal line.
assonance	The pattern of stressed and unstressed syllables discernible in a line of poetry has been analyzed in order to determine whether the line follows an iambic or a dactylic or an anapestic metrical arrangement. Each stanza has eight lines.
bait	He reeled in the line and bent the pole. He cast out his line.
horizontal	The curved line represents the variation of voltage in the signal. Draw a horizontal line above the vertical line.
ahead	Matthew dashed across the finish line. I crossed the finish line, jogged to a stop, and kneeled on the cinders, breathing deeply.
Different Sense	The workers would build them on a moving assembly line.
Ambiguous Sense	Hold the line a minute, Diane.

Table 15: Test sentences used in the WSD task for the word ***raise*** and their annotated sense determined by a human rater.

Annotated WSC Label	Sentences Using ***raise*** in this Sense
money	With the new job also came a big raise in pay. The federal reserve board is expected to raise interest rates. For the last six years the British parliament had been trying to raise money by taxing the American colonies.
crops	Farmers use water to grow crops and raise animals. Farmers raise hogs and cattle as well as corn and wheat. He raises 2,000 acres of wheat and hay.
raised	It looks like a fine place to raise children. He shouldn't have to raise someone else's kid.
support	The agreement has raised hopes that the war may end soon. His speech was meant to raise the interest of the people in his company.
Different Sense	I would like to help raise the flag each morning.
Ambiguous Sense	The government raised interest.

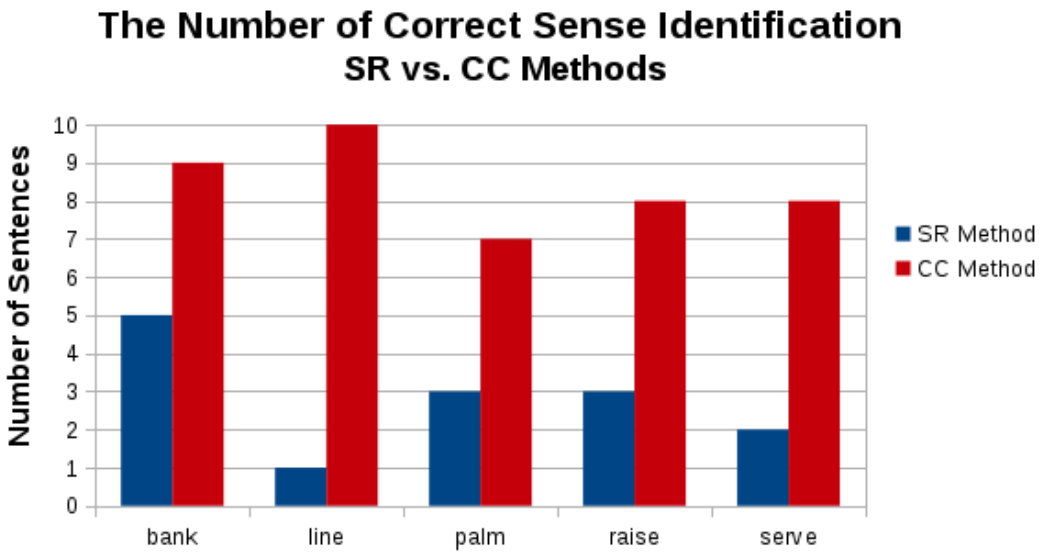

Figure 20: Comparison of the number of correctly identified word senses using each WSD method for the target words used in ten different context sentences

human generated sense identification to determine correctness. The CC method outperformed the SR method for each of the target words examined (see Figure 20).

5.3 Individual Observations

For the target word *bank*, there were two induced senses resulting from synclustering (see Table 9 in Section 4.4.1). Both WSD methods identified the correct sense for all the sentences pertaining to riverbank (identifier "downstream"), but the SR method incorrectly identified the sense for *bank* used in any of the test sentences pertaining to "money", always matching the "downstream" sense in all cases. In contrast, the CC method incorrectly identified the sense for *bank* only once. This was a sentence ("It was a bank robbery in progress.") where the sense of *bank* was "money". The cosine using the CC

method was 0.14 for the “downstream” sense and 0.13 for the “money” sense. Both of these cosines are low, but more importantly, they are about the same, with a difference of only .01. This suggests a low degree of confidence in the WSD result for the target word in the sentence.

In addition to the sentences using the senses induced for the word *bank*, an additional test sentence with a different sense not corresponding to one of the induced senses was processed to determine what the methods would produce. This sentence: “Chuck was listening to the whistles and trills and adjusting dials on a bank of electronic equipment.” makes use of the target word *bank* in the sense of a group or set of similar things. Both methods selected “downstream” as the sense, but the corresponding cosine similarity measurements were 0.78 and -0.02 for the SR and CC methods respectively. With a cosine value near zero (indicating two vectors are unrelated), it can be interpreted that the CC method did not recognize the sense of the word *bank* in the sentence. This would be expected given that operative sense for *bank* in this context had not been learned by the underlying learning system. The same sort of observation was noted for the ambiguous sentences, “That bank’s not safe.” and “Rosa had waved to her at the bank.”. The SR method still had a relatively high cosine values of 0.85 and 0.88 identifying the sense for the word *bank* in those sentences as “downstream”, but the CC method calculated cosine values of 0.05 and 0.04 suggested that no sense could be strongly identified. Interestingly, those cosines produced by the CC method corresponded to the sense of “money” for *bank*.

For the word *line* there were five senses induced by synclustering using the grade level learning system (see Section 4.4.1 and Table 13). Ten test sentences, two for each sense, were processed using the WSD software with both the SR and CC methods. The SR method correctly identified the sense for only one of the test

sentences. Comparatively, the CC method was able to correctly identify all the senses for each of the given the contexts.

Out of all the sentences containing *line*, whether from the ten test sentences that have a matching induced sense or the additional sentences with an unlearned sense or ambiguous usage, the SR method selected “zone” as the sense 85% of the time. The CC method again produced cosine measurements near zero for the test sentences using unlearned or ambiguous senses, suggesting low confidence in the sense identification produced for those contexts. However, given the available learned senses to choose from, with the ambiguous sentence “Hold the line a minute, Diane.”, the CC method identified the sense “bait” (fishing line), which from a human perspective is the most reasonable out of the five induced senses: goal line, fishing line, finish line, poetry line, or mathematical line.

Examining the target word *palm* and its three induced senses, “hand”, “gripping”, and “trees” (see Table 11 in Section 4.4.1), the SR method identified the sense of “gripping”, for the word *palm* in all cases, of which only 30% were correctly identified. The CC method identified 70% of the senses correctly, missing the correct identification for the sense of *palm* in the following three sentences:

1. Yellow sap oozed onto my palm.
2. She stroked my golden curls with a hand so large it seemed to palm my whole head.
3. I suspected that he had palmed a playing card.

The first sentence uses the word *palm* in the “hand” sense, but the CC method identified the sense as “trees” with a cosine value of 0.23 between the syncluster centroid representing the “hand” sense and the original context sentence with the target word removed. The word “sap” in the sentence suggests the “tree” sense, but from a human standpoint it is obvious that sentence uses *palm* in the context

of a hand. The other two sentences (2 and 3) use the “gripping” sense, which is the holding, pick up, or stroking with a hand, for the word *palm*. For these the CC method identified the sense for *palm* in sentence 2 as the “hand” sense with a cosine similarity value of 0.51 and a cosine value of 0.23 for the sense of “gripping”. These measurements are not close, so the identification of the incorrect “hand” sense is clear. For sentence 3 the CC method produced cosine similarities near zero for all senses (0.03, -0.02, and -0.04) indicating that no sense could be discerned with a high degree of confidence. For sentences with unlearned or ambiguous usage of the word *palm*, the results were similar to the those observed for *bank* and *line*. The cosine similarity values were near zero and in each case indicating the sense was not identifiable with any degree of confidence.

For the word *raise* (see Section 4.4.1, Table 12), the SR method achieved an accuracy of 30% on the test sentences and again predominantly identified the same sense, only indicating another sense in two cases. The CC method identified the correct senses for 80% of the test sentences, incorrectly identifying the sense in two cases where the word *raise* was used in the “money” sense. The two sentences where the sense was incorrectly identified were the following:

1. With the new job also came a big raise in pay.
2. The federal reserve board is expected to raise interest rates.

Both sentences were identified as using *raise* in the “support” sense, which is the sense related to “support of or interest in something”. Intuitively it would seem that the CC method should identify these sentences as using the word *raise* in the “money” sense, but further investigation into the cluster reveals that there are no words suggesting “increase” in the “money” sense cluster. The word “increase” did not appear in the top 100 synonyms for *raise*, but out of those top 100 synonyms, the word closest in meaning to the word “increase” is the word

"improve" ranked 92nd. The word "improve" is included in the "support" syncluster, and that is the sense that was identified for both sentences.

The WSD results for the target word *serve* (see the senses described in Table 14) are similar to those for the other words examined. The SR method identified the same sense, in this case the "accept" sense, for the word *serve* in all but two test sentences, producing a correct identification only 20% of the time. The CC method identified the correct sense in 80% of the test cases. Of the two cases where the sense identification was incorrect, one was the sentence "The woman will serve on a jury for a murder trial." In this case the correct sense was the "public" sense but the sense was incorrectly identified as the "accept" sense, with the "public sense" being ranked a very close second. The other sentence, "Two additional spheres attached to the bottom of the station would serve as observation points for studying the undersea environment.", again was incorrectly identified with the "accept" sense instead of the correct "prepared" sense. The separation between these two senses is a fine-grained distinction, and the error could be attributed to the human rater. For sentences with unlearned or ambiguous usage for the word *serve*, the CC method again indicated that the sense of the target word was not identifiable as evidenced by the near zero cosine values.

5.4 WSD Conclusions

Comparing the performance of the SR and CC methods, it was apparent in all the cases tested that the CC method produced much more accurate sense identification results. For each target word, the SR method predominantly associated a single sense with that word in most of the tested contexts (80-100%). Interestingly, the dominant sense selected was not always the one with the largest syncluster or was it necessarily the syncluster with the highest cosine between its centroid and the target word, such as in the case for *raise* and *palm*. The original

hypothesis for the SR method was that replacing the target word with representative synonyms from each syncluster would identify the proper sense syncluster because the correct sense synonym would perturb the mapping of the context sentence to the least degree. The results of these experiments however, indicate that there may exist one or more senses that are carried by the target word within its vector mapping more than all of the others. It can be concluded that the SR method is failing to separate the senses of the target word adequately for a correct identification to be made. Further investigation of this condition is left as a subject for future research.

In contrast, the CC method did well in correctly identifying the sense of the target word in a given context. Additionally, the cosine similarity measure produced by the CC method appears to provide information about the confidence of the sense identification, and even an indication of when a sense cannot be clearly determined. More experimentation with this method is needed to further develop these promising results. The confidence in the sense identification is a measure of the learning embodied in the LSA semantic space being used for the WSD task. Observations of the performance in WSD task further refine the characterizations of the learning system that were developed in the WSI task.

CHAPTER 6
CONCLUSIONS AND RECOMMENDATIONS

Although work in the field of WSD in natural language processing began in the late 1940s, truly unsupervised, fully automated, and language agnostic technology is still lacking. Much advancement has been made, but a recent (2015) survey on WSD indicates that expensive supervised WSD approaches are still required to attain near human levels of ability in disambiguating (Pal and Saha 2015). Less expensive unsupervised methods are still lacking in ability. This research was intended to develop a new WSD method using LSA that was unsupervised and fully automated with sufficient accuracy levels for disambiguation of targeted words, and the results using LSA-WSD process for WSI and WSD demonstrate success in accomplishing that.

There exists no one annotated source (dictionaries, WordNet, etc.) that has an exhaustive list of possible word senses for a word, nor do the existing sources agree on the possible senses for the same word. For example, WordNet has thirty noun meanings for the word *line* and six verb meanings. Merriam-Webster dictionary has fifty noun meanings and ten verb meanings for the word *line*, while dictionary.com has fifty-four noun senses and nine verb senses and includes a sense for a “drug” *line* that does not appear in the other sources. This demonstrates that no one source, even ones produced by expert lexicographers, provides a consistent, comprehensive definition for a word. This is especially true as language changes over time. With this in mind, it must be admitted that the expectation of complete and comprehensive sense discovery cannot be achieved with unsupervised, automated WSI systems either. However, these WSI systems have certain advantages. They can usually be adapted to different languages and easily change when language changes. Also, WSI systems give insight into the

AI upon which the system is built. This research using the LSA-WSD system, that incorporates an LSA-based learning system, discovers word senses utilizing solely the input text which is representative of a specific body of knowledge. It can be used with any language and is easily updatable with new linguistic examples. Furthermore, the results of the LSA-WSD system, both for sense discovery and sense identification, indicate how well the LSA-based learning system has learned a word's senses. This can be leveraged for many different applications and uses.

6.1 Conclusions

The research presented in this dissertation for measuring word importance in a sentence, word sense induction, and word sense disambiguation produced notable results that show promise and will guide future research in using the LSA-WSD system for word sense disambiguation as well as analyzing LSA-based learning systems.

In the sentence word importance research presented in Chapter 3, it was demonstrated that the importance of a word to a sentence can be determined by computing its cosine impact value (CIV). It was shown that those words having a CIV of less than 0.9 were primary contributors to the meaning of the sentence being examined. Additionally, most words typically had a small impact on the meaning of a sentence, but the majority of sentences in the corpora examined did contain at least one important word. Overall, it was observed that all the words used in a sentence contribute something to the meaning of the individual sentence. For longer sentences individual word importance was diminished. There were very few unique words that were deemed important for any sentence within a document corpus. Corpus size and content did not have an observable effect on the word importance. The experiments and analyses done for word importance were used to provide insight to which sentences to use for the sentclustering approach

discussed in Chapter 4, but the subject of word importance by itself is interesting and warrants further investigation.

To address the task of word sense induction described in Chapter 4, sentences where the target word was important to the sentence's meaning were selected as the best candidates for use in the sentclustering approach to produce word senses for the individual target words. Sentclustering was able to discover multiple senses for a target word, but routinely produced multi-sense clusters that it was unable to separate. Sentclustering was abandoned as a viable means for accomplishing the WSI task.

In contrast, the synclustering approach described in Section 4.4 was shown to produce reasonable results for sense discovery. This approach leveraged the word knowledge contained in the LSA semantic space to induce word senses for a given learning system. With synclustering, word senses for any word in the learning system can be discovered without the need to select specific sentences for examination. The results from the synclustering research suggested that candidate WSCs should have a cosine similarity between the centroid and the target word that exceeds 0.35 to be considered as a sense cluster for the word. Coarse-grained senses were more easily identified, but a few fine-grained senses were discovered as well. Multiple base corpora were tested in this research, and it was observed that the more general grade level learning system produced more broad-based and consistent results for WSI. The synclustering approach for automatic WSI provided a means to induce word senses as well as examine and analyze the knowledge contained in the underlying LSA-based learning system.

It is worth noting that the LSA-WSD software can be used as an unsupervised system, or as a semi-supervised system for WSI as it allows the user to judge the WSC derived using synclustering in the system. The user has the ability to refine

the candidate WSCs by choosing a specific cosine cluster inclusion threshold as well as the number of top terms to use for clustering to capture the best senses for the target word derived from the learning system. The user can also indicate through input parameters which candidate WSCs to keep as induced senses.

For the word sense disambiguation task of sense identification, the subject of Chapter 5, two methods, synonym replacement (SR) and context comparison (CC), were proposed and considered within the LSA-WSD system using senses induced with the synclustering approach. The CC method was observed to produce more accurate sense identification results. The SR method tended to associate a single word sense for a target word with the majority of the tested context sentences. The SR method failed to separate the senses of the target word adequately for a correct identification to be made. In comparison, the CC method performed well, identifying the correct sense for a given target word within the context sentences 84% of the time. Additionally, the cosine similarity measure produced by the CC method provided information that suggested a degree of confidence for the sense identification, and even an indication of when a sense could not be clearly determined.

6.2 Future Research

Word sense disambiguation is a large and complicated undertaking. This research has shown promising results for unsupervised sense discovery and sense identification. Further research will help to refine the LSA-WSD process for the tasks of WSI and WSD. Additional work may also leverage this research for probing of the cognitive model used with the LSA-WSD system and help to further characterize the knowledge base that it represents.

The investigation of word importance within in any particular sentence is the first research of its kind using the LSA-based learning system. While the sentclustering

approach did not prove to be a useful method for WSI, word importance had some interesting findings that raise additional questions for future research. The same notion of word importance could be applied to sentence importance within a document, or other sub-part related to a larger collective expression of text. There are also possible applications of word importance in educational settings such as evaluating vocabulary acquisition or assisting in composition.

Although using sentclusters to produce word senses did not result in well defined WSCs, further refinement may be possible that would yield improved results. Human annotated sentences, or some other selection criteria for sentences, might provide better candidates for input to a sentclustering process for use in WSI because the LSA-based learning system might be able to better derive word senses from that input. The main problem encountered with the sentclusters was the production of multi-sense clusters. Secondary processing of these clusters might be a viable means for further refining these clusters into distinct senses.

Within the LSA-WSD system, the synclustering approach produced promising results for inducing word senses. More research should be done both on inducing senses for other words as well as using more synonyms and different thresholds for the already studied target words to refine the initial results presented in this dissertation. Further research will help to define the input criteria, number of synonyms to use and the cluster inclusion threshold, to further optimize results.

For the WSD task, only two possible methods were explored (SR and CC). Other methods for disambiguating the sense of a word in given context using the word senses already induced are possible and left as a subject for future research. Additionally, more sentences need to be tested to further validate and generalize the promising results of the CC method for distinguishing senses.

Finally, the development of the LSA-WSD system has successfully produced a viable unsupervised system for automating both the sense discovery and sense identification tasks of WSD. The flexibility and adaptability of the system allows for the LSA-WSD system to be used for different applications and purposes. There are still many languages in which WSD has not been attempted, and this system can be used to explore WSD in those venues. The system can also help to define the body knowledge and use of language captured in the underlying learning system and to guide the creation of these systems for general application.

LIST OF REFERENCES

Aggarwal C. C., & Reddy C. K. (Eds.). (2014). *Data Clustering Algorithms and Applications*. Boca Raton, FL: Taylor & Francis Group.

Agirre, E., & Edmonds, P. (2007a). Introduction. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and Applications* (pp. 1–22). Springer Science & Business Media.

Agirre, E., & Edmonds, P. (Eds.). (2007b). *Word Sense Disambiguation: Algorithms and Applications*. Springer Science & Business Media.

Bhala, R. V., & Abirami, S. (2014). Trends in word sense disambiguation. *Artificial Intelligence Review*, *42*(2), 159–171. http://doi.org/10.1007/s10462-012-9331-5

Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T. K., & Foltz, P. W. (2014). Models of Vocabulary Acquisition: Direct Tests and Text-Derived Simulations of Vocabulary Growth. *Scientific Studies of Reading*, *18*(2), 130–154. http://doi.org/10.1080/10888438.2013.821992

Boyd-Graber, J., Blei, D., & Zhu, X. (2007). A topic model for word sense disambiguation. In *EMNLP-CoNLL* (pp. 1024–1033).

Deerwester, S., Dumais, S. T., Furnas, G., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of The American Society for Information Science*, *41*(6), 391–407.

Di Marco, A., & Navigli, R. (2012). Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. *Computational Linguistics*, *39*(3), 709–754. http://doi.org/10.1162/COLI_a_00148

D'mello, S., & Graesser, A. (2012). AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS) - Special Issue on Highlights of the Decade in Interactive Intelligent Systems*, *2*(4). http://doi.org/10.1145/2395123.2395128

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2), 229-236.

Edmonds, P., & Cotton, S. (2001). SENSEVAL-2: Overview. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems* (pp. 1–5). Stroudsburg, PA, USA: Association for Computational Linguistics.

English Oxford Living Dictionaries. (n. d.). Retrieved from https://en.oxforddictionaries.com/

Edmonds, P., & Kilgarriff, A. (2002). Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering*, *8*(4), 279–291.

Everitt, B. S., Landau S., Leese M., Stahl D. (2011). *Cluster Analysis*, *5th Edition*. Chichester, West Sussex, United Kingdom: John Wiley & Sons, Ltd.

Fellbaum, C. (2012). WordNet. In *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Inc.

Foltz, P. W. (1996). Latent Semantic Analysis for Text-based Research. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 197–202.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *28*(2-3), 285–307. http://doi.org/10.1080/01638539809545029

Foltz, P. W., Streeter, L. A., Lochbaum K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation* (pp. 68-88). Lawrence Erlbaum Associates.

Gaur, V., & Jain, S. (2013). Word Sense Induction and Disambiguation Using Principal Component Analysis and Latent Semantic Indexing. *International Journal of Scientific and Research Publications*, *3*(11).

Gliozzo, A., Magnini, B., & Strapparava, C. (2004). Unsupervised Domain Relevance Estimation for Word Sense Disambiguation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 380–387).

Graesser, A., Penumatsa, P., Ventura, M., Cai, Z., & Hu, X. (2007). Using LSA in AutoTutor: Learning Through Mixed-Initiative Dialogue in Natural Language. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. (pp. 243-262). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and Their Applications*, *15*(5), 22–37.

Hirst, G. (2007). Foreword. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and Applications* (pp. xvii–xix). Springer Science & Business Media.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.). Prentice Hall.

Katz, P., & Goldsmith-Pinkham, P. (2006). *Word Sense Disambiguation using Latent Semantic Analysis*. Retrieved from /www.sccs.swarthmore.edu/users/07/pkatz1/cs65f06-final.pdf

Kilgarriff, A. (1998). SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. In *In LREC* (pp. 581–588).

Kilgarriff, A., & Palmer, M. (Eds.). (2000). *Special double issue on the Senseval Word Sense Disambiguation Exercise* (Vol. 34:1–2). Computer and the Humanities.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street: Computer-Guided Summary Writing. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. (pp. 263-278). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kintsch, W. (2007). Meaning in Context. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 89–105). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kireyev, K., & Landauer, T. K. (2011). Word Maturity: Computational modeling of word knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 299–308). Porland, Oregon.

Kulkarni, A., & Pedersen, T. (2006). How Many Different “John Smiths”, and Who Are They? In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2* (pp. 1885–1886). Boston, Massachusetts: AAAI Press.

Landauer, T. K. (1998). Learning and Representing Verbal Meaning: The Latent Semantic Analysis Theory. *Current Directions in Psychological Science*, *7*(5), 161–164.

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In *Psychology of Learning and Motivation* (Vol. 41, pp. 43–84). Elsevier Inc.

Landauer, T. K. (2007). LSA as a Theory of Meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 3–34). Mahwah, New Jersey: Lawrence Erlbaum Associates.

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.

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

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word Maturity: A New Metric for Word Knowledge. *Scientific Studies of Reading*, *15*(1), 92–108. http://doi.org/10.1080/10888438.2011.536130

Landauer, T. K., Laham, D., & Foltz, P. W. (1998). Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report. In M.

I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems* (pp. 45–51). MIT Press.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003a). Automatic Essay Assessment. Assessment in Education: Principles, Policy & Practice, 10(3), 295-308. http://doi.org/10.1080/0969594032000148154

Landauer, T. K., Laham, D., & Foltz, P. W. (2003b). Automated Scoring and Annotation of Essays with the Intelligent Essay Accessor. In M. D. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-disciplinary Perspective* (pp. 87–112). Lawrence Erlbaum Associates.

Landauer, T. K., Lochbaum, K. E., & Dooley, S. (2009). A New Formative Assessment Technology for Reading and Writing. *Theory Into Practice*, *48*(1), 44–52. http://doi.org/10.1080/00405840802577593

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Landauer, T. K., & Way, D. (2012). Improving text complexity measurement through the Reading Maturity Metric. Presented at the National Council on Measurement in Education, Vancouver, Canada.

Levin, E., Sharifi, M., & Ball, J. (2006). Evaluation of Utility of LSA for Word Sense Discrimination. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers* (pp. 77–80). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lewis, D., Yang, Y., Rose, T., & Li, F. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, *5*, 361–397.

Martin, D. I., & Berry, M. W. (2007). Mathematical Foundations Behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. (pp. 35-56). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Martin, D. I., & Berry, M. W. (2010). Latent Semantic Indexing. In M. J. Bates & M. N. Maack (Eds.), *Encyclopedia of Library and Information Sciences (ELIS)* (Third, pp. 2195–3204). Oxford: Taylor & Francis.

Martin, D. I., & Berry, M. W. (2015). Text Mining. In N. Higham (Ed.), *The Princeton Companion to Applied Mathematics (*pp. 887-891). Princeton, New Jersey: Princeton University Press.

Martin, D. I., Martin, J. C., & Berry, M. W. (2016). The Application of LSA to the Evaluation of Questionnaire Responses. In M. E. Celebi & K. Aydin (Eds.), *Unsupervised Learning Algorithms*. (pp. 449-484). Switzerland: Springer International Publishing.

Martin, J. C. (2016). (Doctoral dissertation). Quantitative Metrics for Comparison of Hyper-dimensional LSA Spaces for Semantic Differences. Retrieved from http://trace.tennessee.edu/utk_graddiss/3942.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys (CSUR)*, *41*(2), 10. http://doi.org/10.1145/1459352.1459355

Navigli, R., & Lapata, M. (2010). An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, *32*(4), 678–692. http://doi.org/10.1109/TPAMI.2009.36

Pal, A. R., & Saha, D. (2015). Word sense disambiguation: a survey. *arXiv Preprint arXiv:1508.01346*. http://doi.org/10.5121/ijctcm.2015.5301

Pedersen, T. (2007). Unsupervised Corpus-Based Methods for WSD. In E. Agirre & P. Edmonds (Eds.), *Word Sense Disambiguation: Algorithms and Applications* (pp. 133–162). Springer Science & Business Media.

Pedersen, T., & Kulkarni, A. (2005). Identifying similar words and contexts in natural language with senseclusters. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 20 No. 4, p. 1694). Menlo Park, CA; Cambridge, MA; London: AAAI Press; MIT Press.

Pedersen, T., & Kulkarni, A. (2007). Unsupervised Discrimination of Person Names in Web Contexts. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 299–310). Berlin, Heidelberg: Springer-Verlag. http://doi.org/10.1007/978-3-540-70939-8_27

Pedersen, T., Kulkarni, A., Angheluta, R., Kozareva, Z., & Solorio, T. (2006). An Unsupervised Language Independent Method of Name Discrimination Using Second Order Co-occurrence Features. In *Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 208–222). Berlin, Heidelberg: Springer-Verlag. http://doi.org/10.1007/11671299_23

Pedersen, T., Purandare, A., & Kulkarni, A. (2005). Name Discrimination by Clustering Similar Contexts. In *Proceedings of the Sixth International*

Conference on Intelligent Text Processing and Computational Linguistics (pp. 220–231).

Pilehvar, M. T., & Navigli, R. (2014). A Large-scale Pseudoword-based Evaluation Framework for State-of-the-art Word Sense Disambiguation. *Compututional Linguistics*, *40*(4), 837–881. http://doi.org/10.1162/COLI_a_00202

Pino, J., & Eskenazi, M. (2009). An Application of Latent Semantic Analysis to Word Sense Discrimination for Words with Related and Unrelated Meanings. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 43–46). Stroudsburg, PA, USA: Association for Computational Linguistics.

Purandare, A., & Pedersen, T. (2004a). Discriminating among word meanings by identifying similar contexts. In *AAAI* (pp. 964–965).

Purandare, A., & Pedersen, T. (2004b). SenseClusters: Finding Clusters that Represent Word Senses. In *Demonstration Papers at HLT-NAACL 2004* (pp. 26–29).

Purandare, A., & Pedersen, T. (2004c). Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces (pp. 41–48). Presented at the Proceedings of CoNLL-2004, Boston, Massachusetts.

Resnik, P., & Lin, J. (2010). Evaluation of NLP Systems. In A. C. member Lecturer, C. F. Reader, & S. Lappinessor (Eds.), *The Handbook of Computational Linguistics and Natural Language Processing* (pp. 271–295). Wiley-Blackwell.

Resnik, P., & Yarowsky, D. (1999). Distinguishing systems and distinguishing senses: new evaluation methods for Word Sense Disambiguation. *Natural Language Engineering*, *5*(02), 113–133. http://doi.org/null

Schumacher, K. (2007). Four Methods for Supervised Word Sense Disambiguation. In Z. Kedad, N. Lammari, E. Métais, F. Meziane, & Y. Rezgui (Eds.), *Natural Language Processing and Information Systems* (pp. 317–328). Springer Berlin Heidelberg.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Comput. Linguist.*, *24*(1), 97–123.

SemEval-2018: International Workshop on Semantic Evaluation 2018. (2018). Retrieved from http://alt.qcri.org/semeval2018/

Stevenson, M., & Wilks, Y. (2005). Word-Sense Disambiguation. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* (pp. 249–265). OUP Oxford.

Tomar, G. S., Singh, M., Rai, S., Kumar, A., Sanyal, R., & Sanyal, S. (2013). Probabilistic LSA for Unsupervised Word Sense Disambiguation. *International Journal of Computer Science Issues*, *10*(5 No. 2), 127–133.

Turing, Alan. (1950). Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460.

Van de Cruys, T., & Apidianaki, M. (2011). Latent Semantic Word Sense Induction and Disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language*

Technologies - Volume 1 (pp. 1476–1485). Stroudsburg, PA, USA: Association for Computational Linguistics.

WordNet - About WordNet. (n.d.). Retrieved April 13, 2016, from https://wordnet.princeton.edu/

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics* (pp. 189–196). Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/981658.981684

Yarowsky, D. (2000). Word-Sense Disambiguation. In R. Dale, H. Somers, & H. Moisl (Eds.), *Handbook of Natural Language Processing* (pp. 629–654). CRC Press.

Zhou, X., & Han, H. (2005). Survey of Word Sense Disambiguation Approaches. In *FLAIRS Conference* (pp. 307–313).

APPENDIX

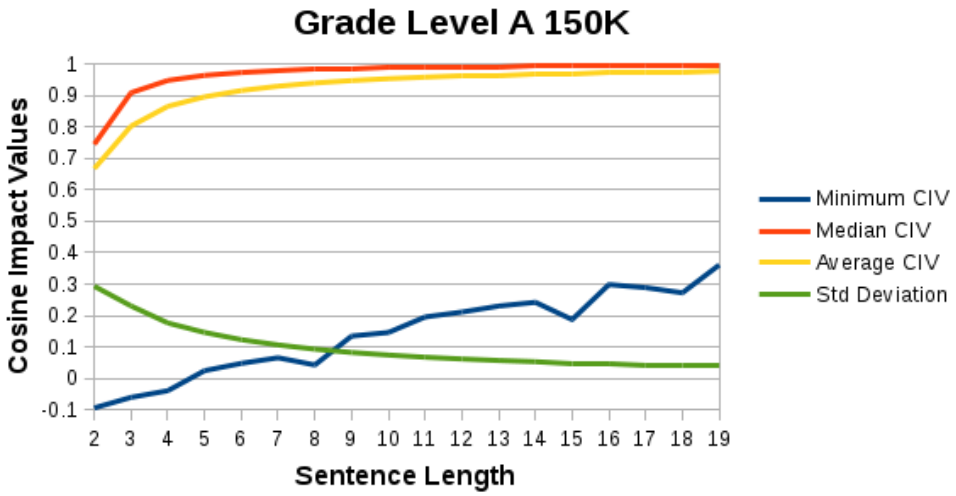

Figure 21: The minimum, median, and average CIVs for all words in sentences of lengths 2-19 for document set grade level A containing ~150,000 documents.

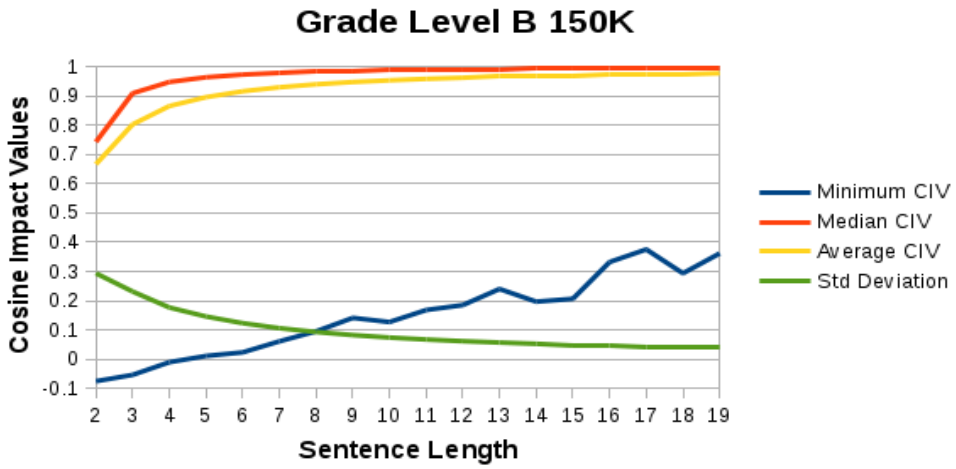

Figure 22: The minimum, median, and average CIVs for all words in sentences of lengths 2-19 for document set grade level B containing ~150,000 documents.

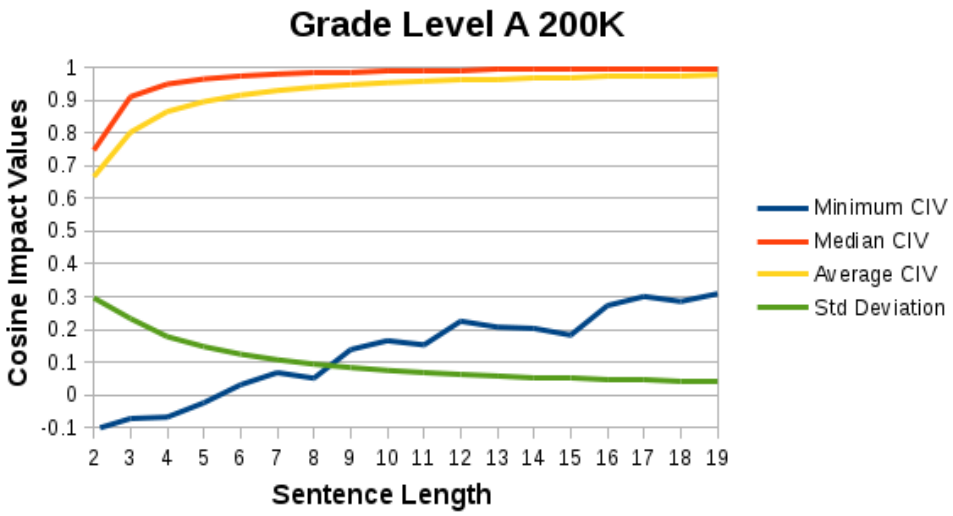

Figure 23: The minimum, median, and average CIVs for all words in sentences of lengths 2-19 for document set grade level A containing ~200,000 documents.

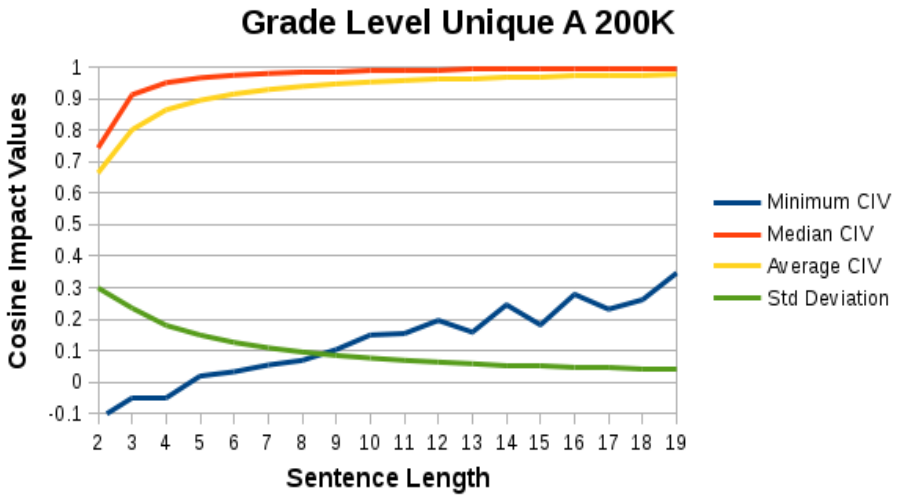

Figure 24: The minimum, median, and average CIVs for all words in sentences of lengths 2-19 for document set grade level unique A containing ~200,000 documents.

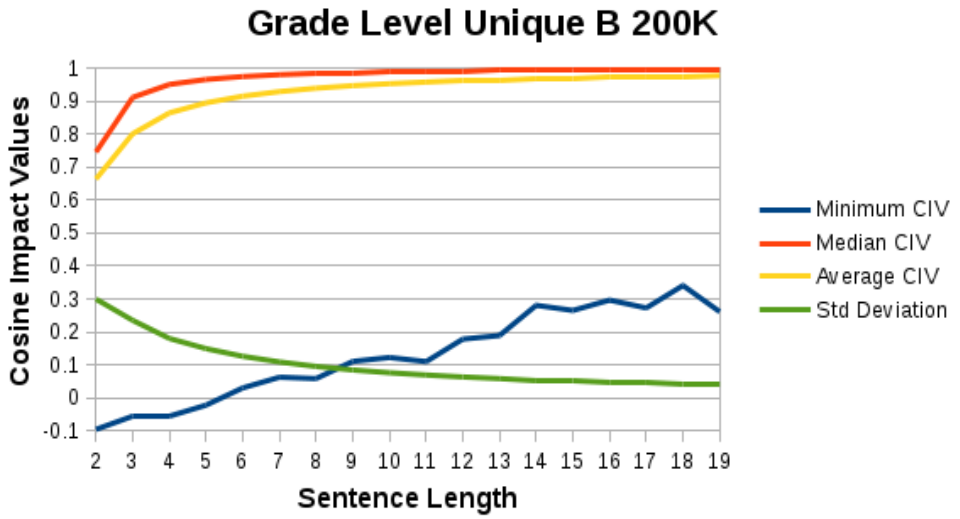

Figure 25: The minimum, median, and average CIVs for all words in sentences of lengths 2-19 for document set grade level unique B containing ~200,000 documents.

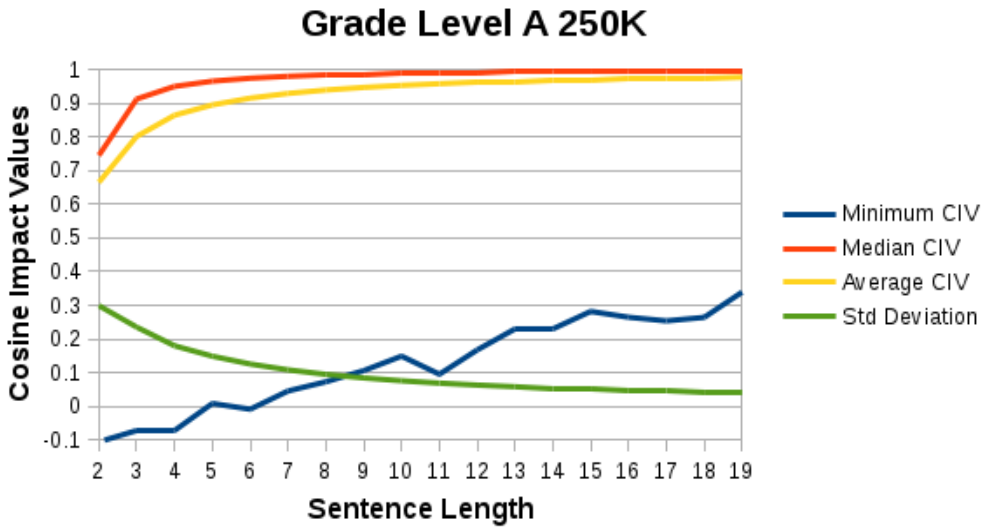

Figure 26: The minimum, median, and average CIVs for all words in sentences of lengths 2-19 for document set grade level A containing ~250,000 documents.

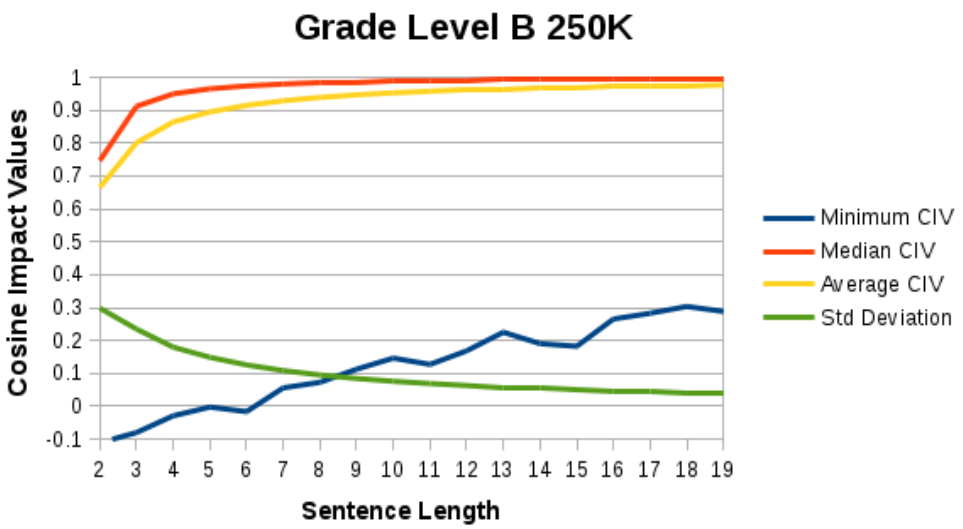

Figure 27: The minimum, median, and average CIVs for all words in sentences of lengths 2-19 for document set grade level B containing ~250,000 documents.

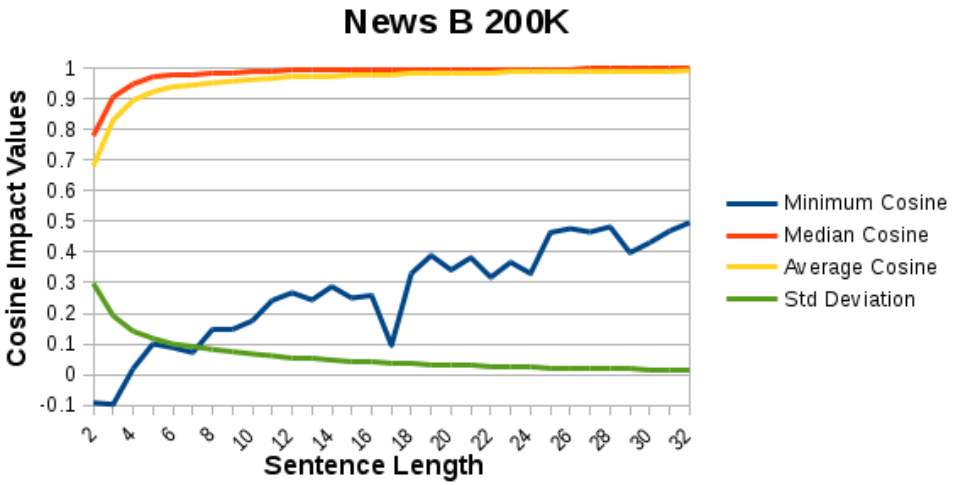

Figure 28: The minimum, median, and average CIVs for all words in sentences of lengths 2-32 for document set news articles B containing 200,000 documents.

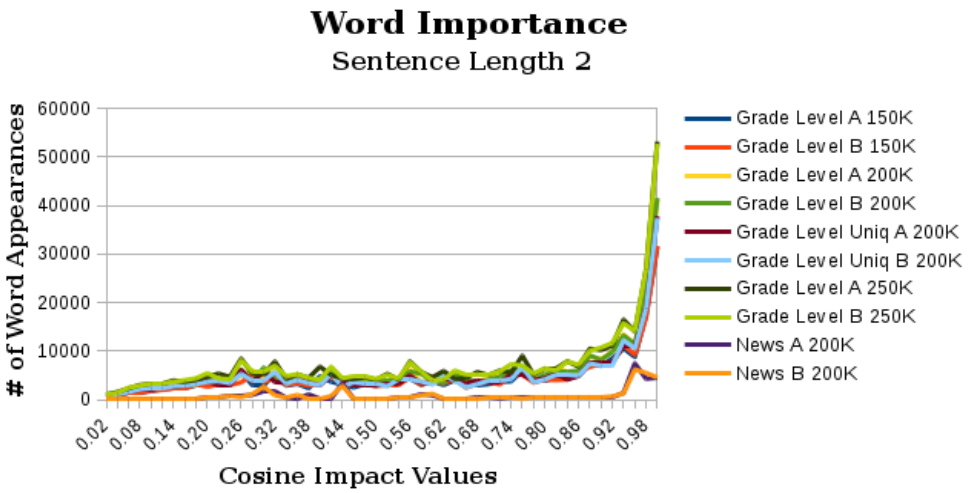

Figure 29: For each document set, the CIV for all the words in sentences of length two in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

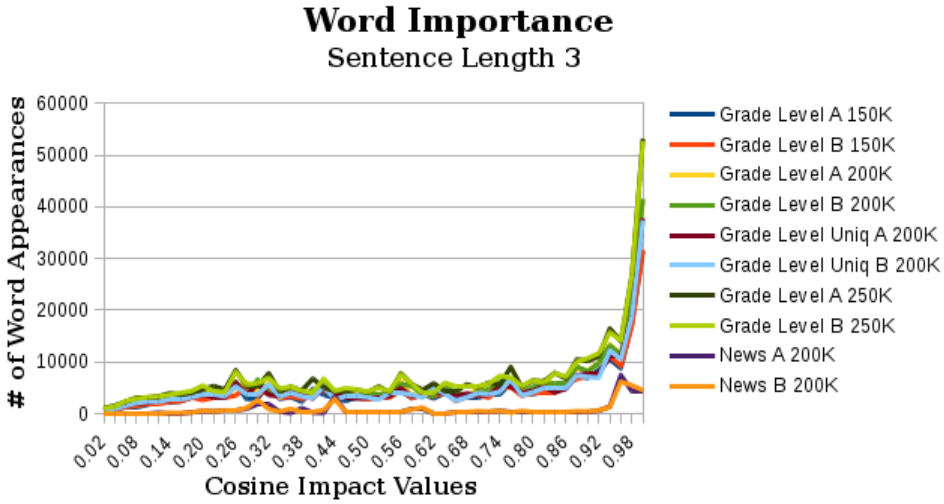

Figure 30: For each document set, the CIV for all the words in sentences of length three in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

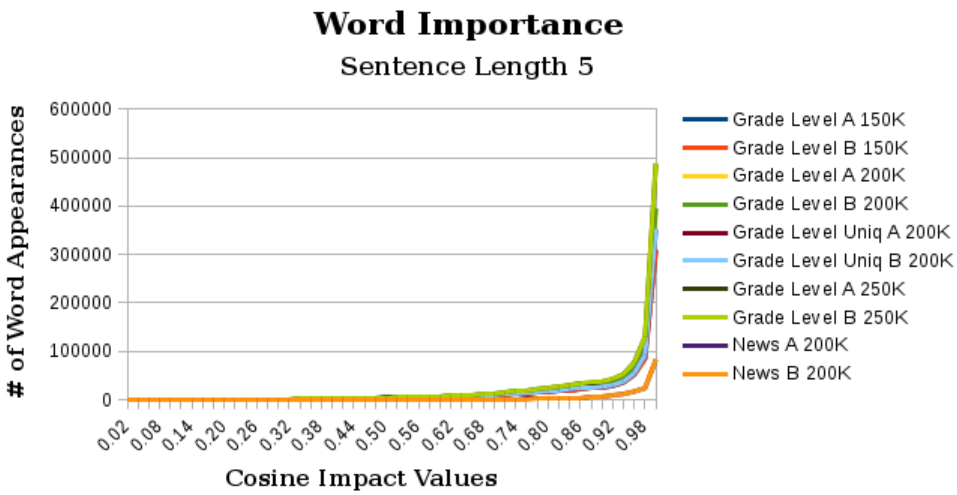

Figure 31: For each document set, the CIV for all the words in sentences of length five in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

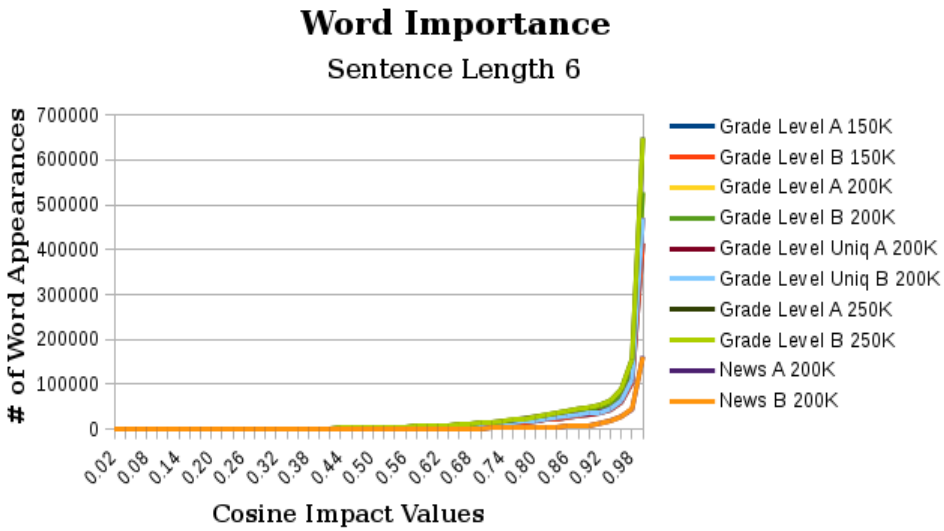

Figure 32: For each document set, the CIV for all the words in sentences of length six in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

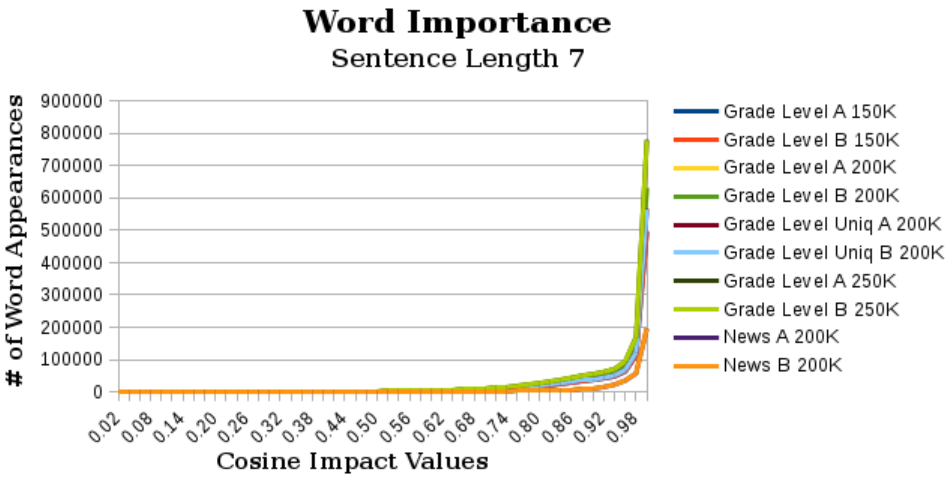

Figure 33: For each document set, the CIV for all the words in sentences of length seven in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

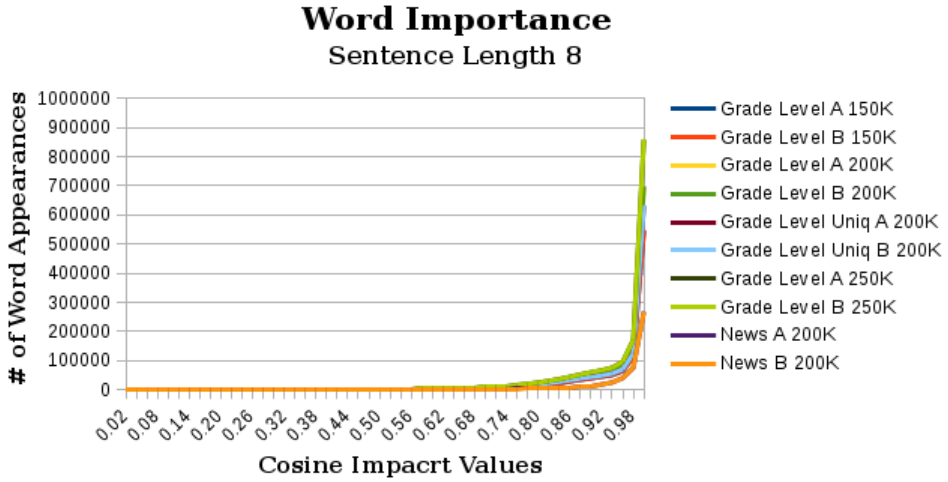

Figure 34: For each document set, the CIV for all the words in sentences of length eight in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

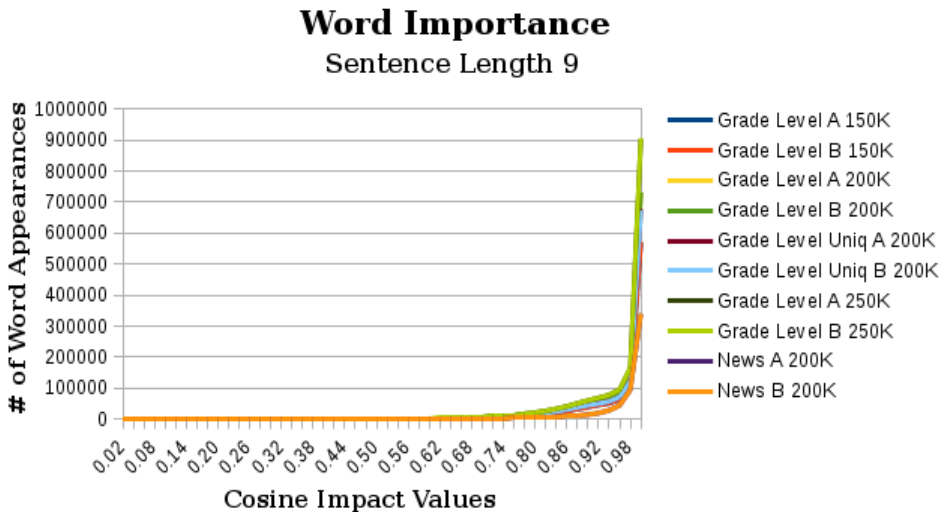

Figure 35: For each document set, the CIV for all the words in sentences of length nine in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

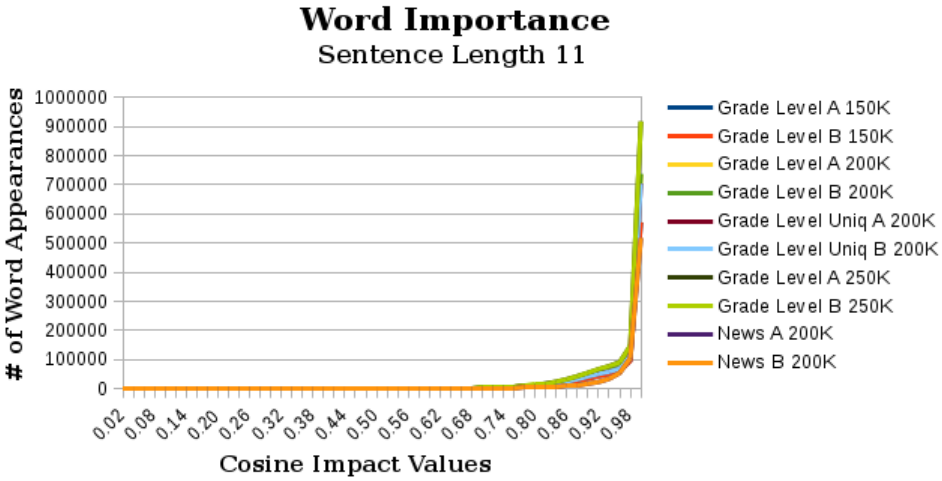

Figure 36: For each document set, the CIV for all the words in sentences of length eleven in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

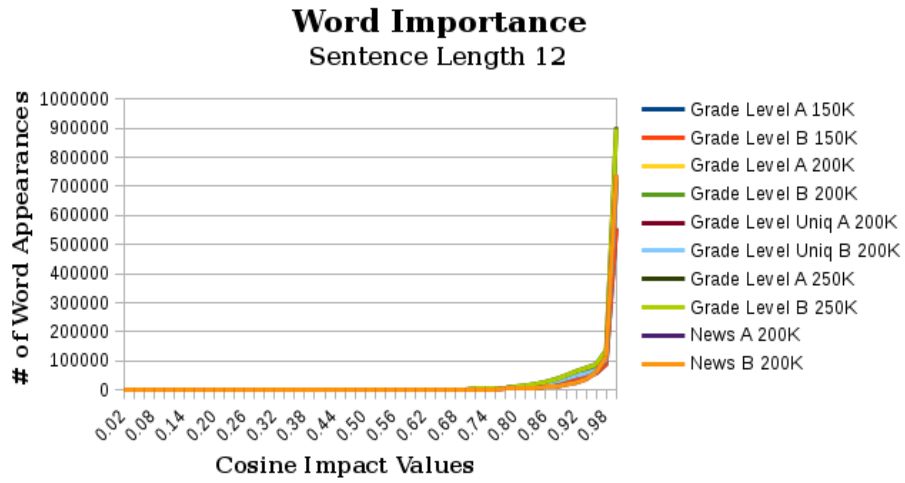

Figure 37: For each document set, the CIV for all the words in sentences of length twelve in that document set was calculated. The number of words at each CIV interval is shown in this line graph

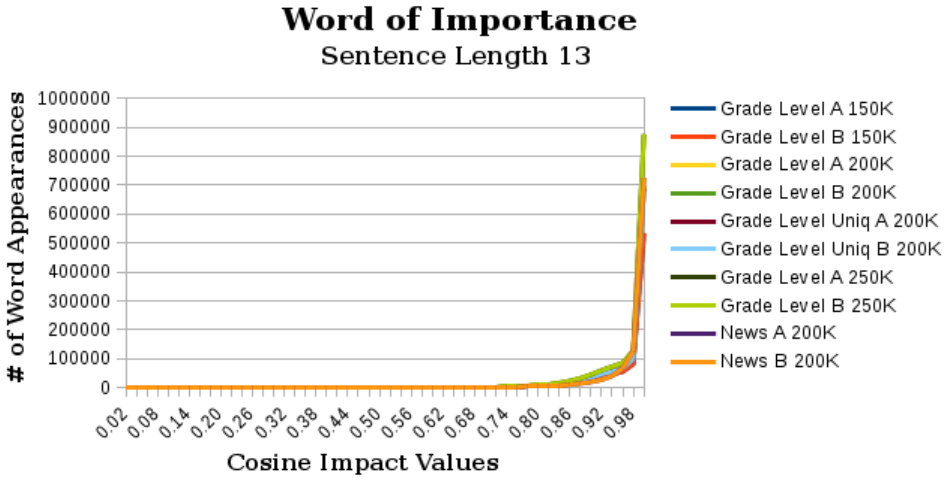

Figure 38: For each document set, the CIV for all the words in sentences of length thirteen in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

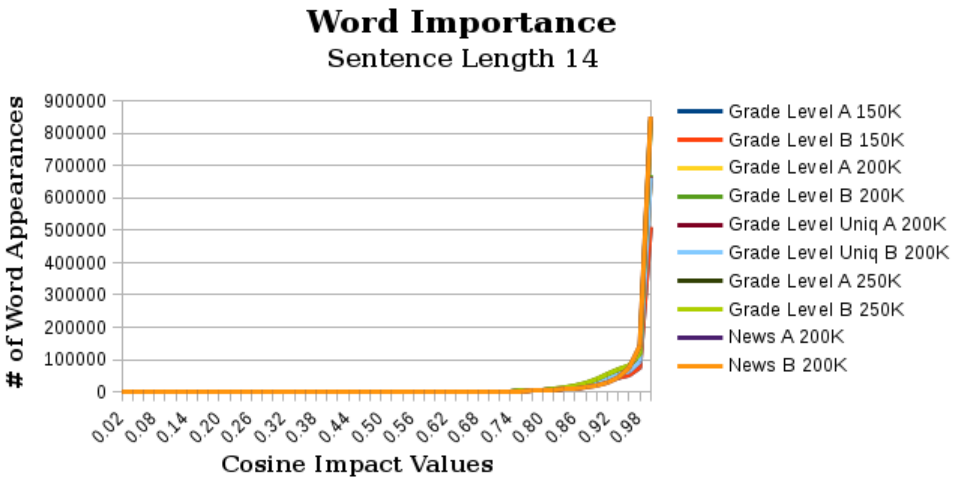

Figure 39: For each document set, the CIV for all the words in sentences of length fourteen in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

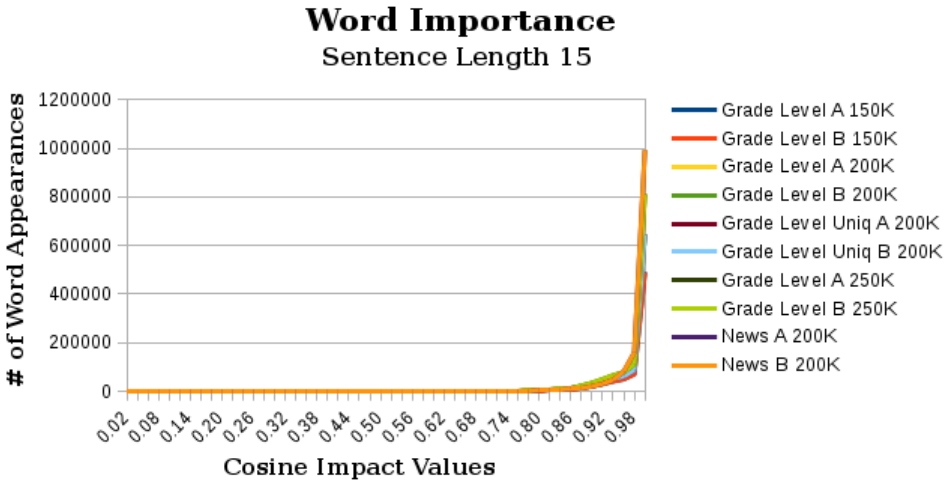

Figure 40: For each document set, the CIV for all the words in sentences of length fifteen in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

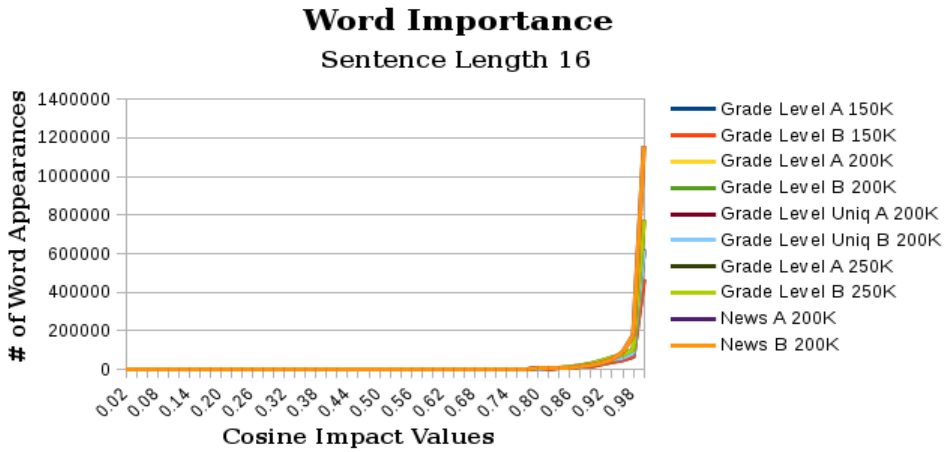

Figure 41: For each document set, the CIV for all the words in sentences of length sixteen in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

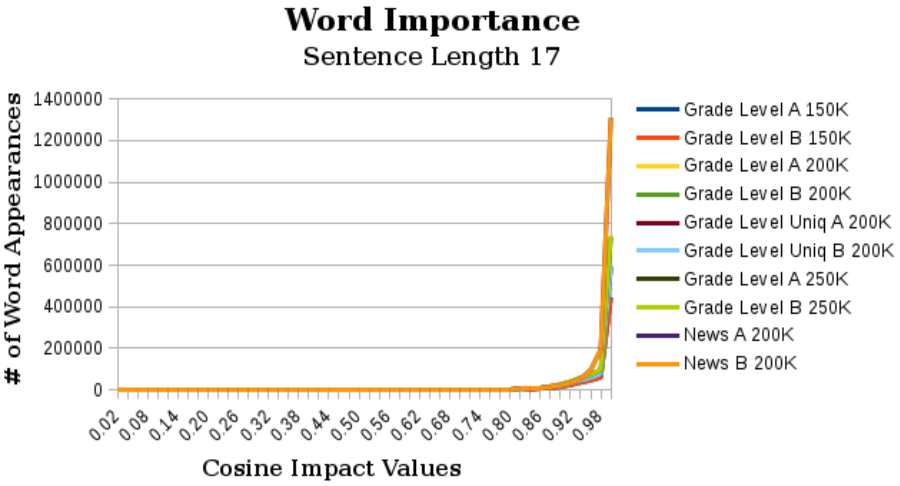

Figure 42: For each document set, the CIV for all the words in sentences of length seventeen in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

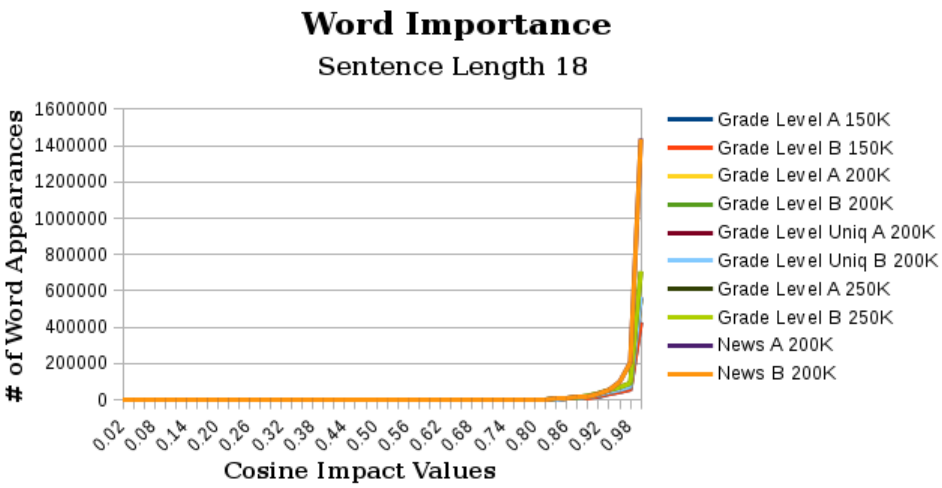

Figure 43: For each document set, the CIV for all the words in sentences of length eighteen in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

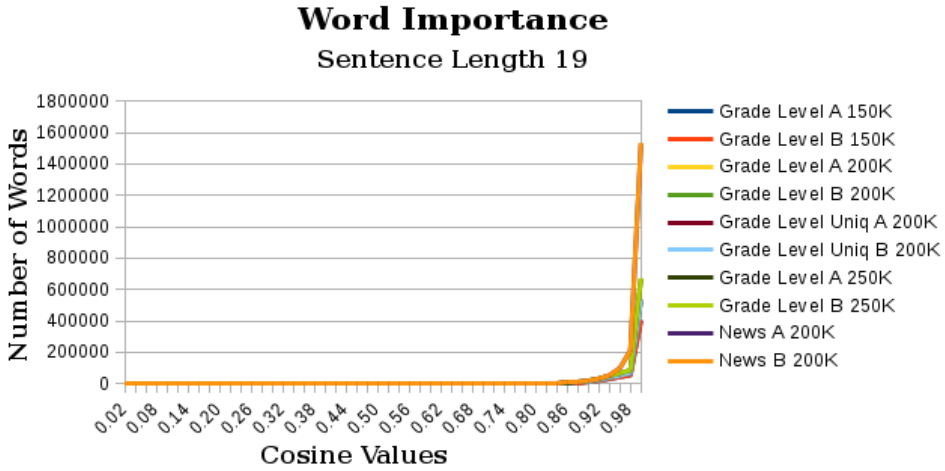

Figure 44: For each document set, the CIV for all the words in sentences of length nineteen in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

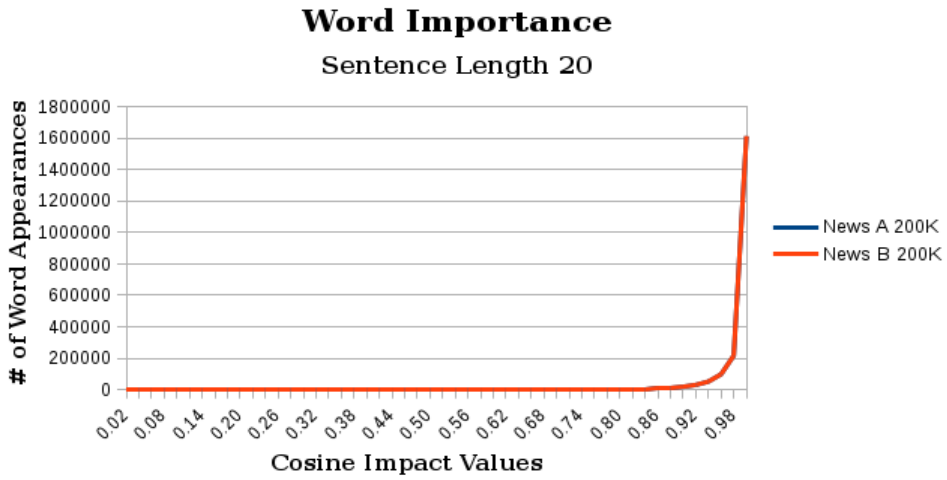

Figure 45: For each news articles document set, the CIV for all the words in sentences of length twenty in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

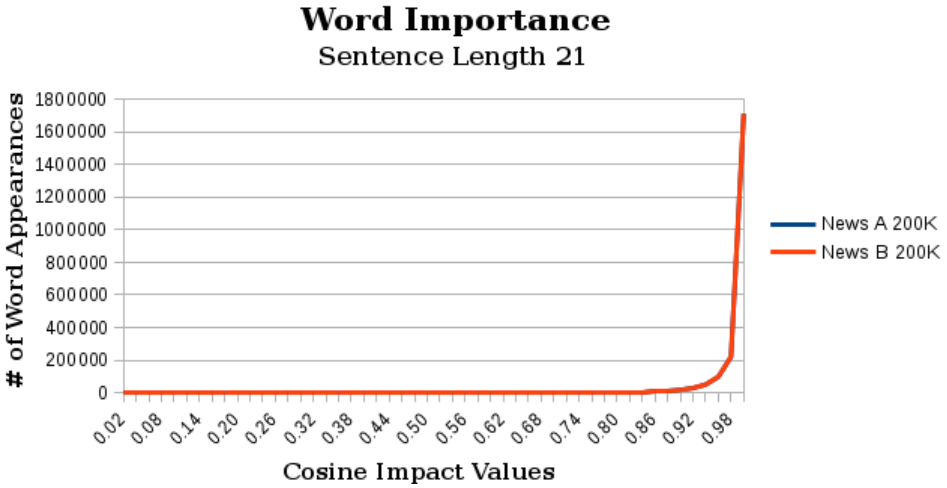

Figure 46: For each news articles document set, the CIV for all the words in sentences of length twenty-one in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

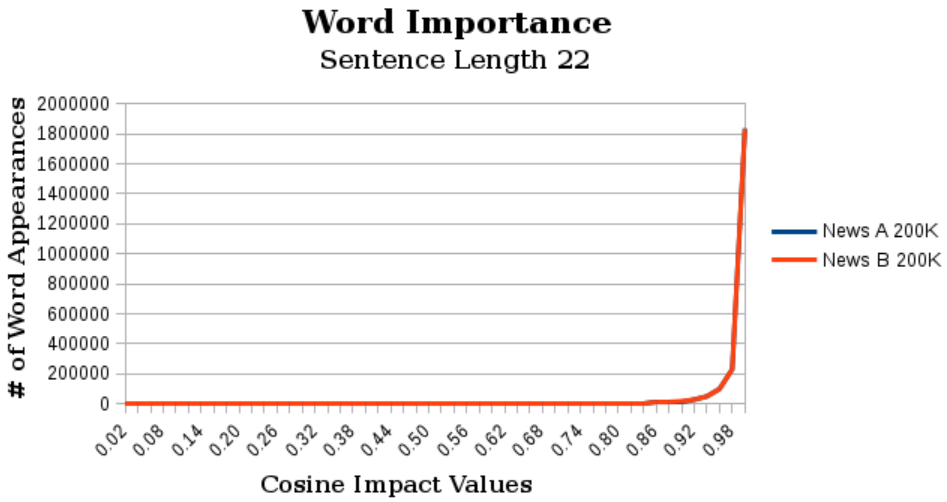

Figure 47: For each news articles document set, the CIV for all the words in sentences of length twenty-two in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

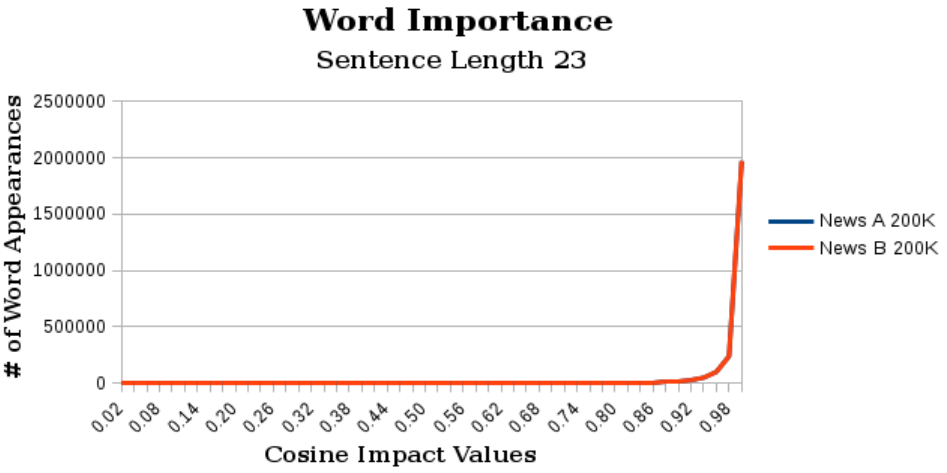

Figure 48: For each news articles document set, the CIV for all the words in sentences of length twenty-three in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

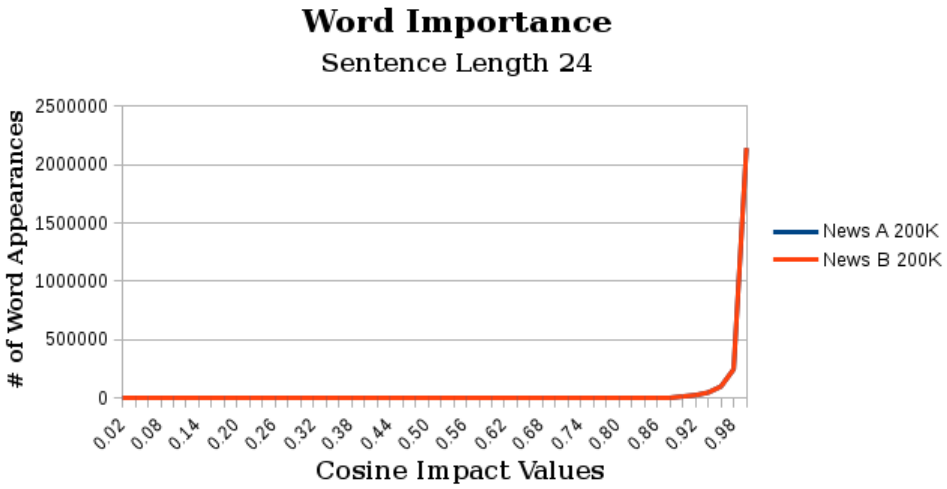

Figure 49: For each news articles document set, the CIV for all the words in sentences of length twenty-four in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

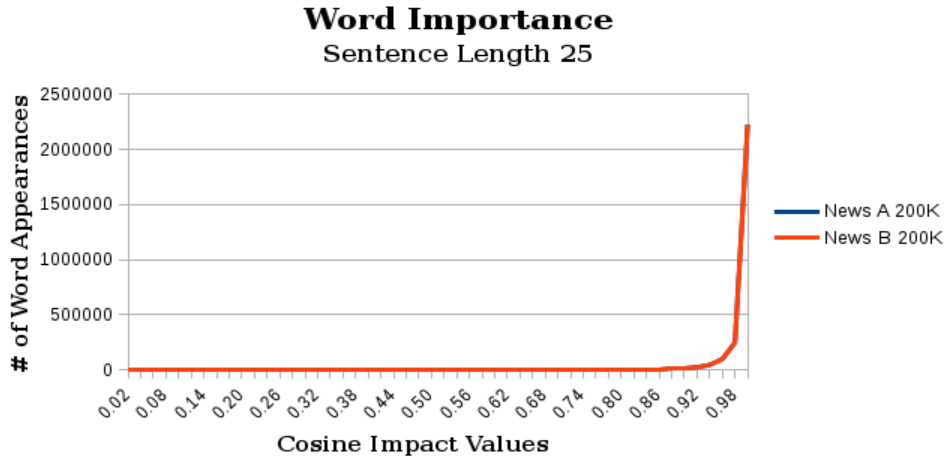

Figure 50: For each news articles document set, the CIV for all the words in sentences of length twenty-five in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

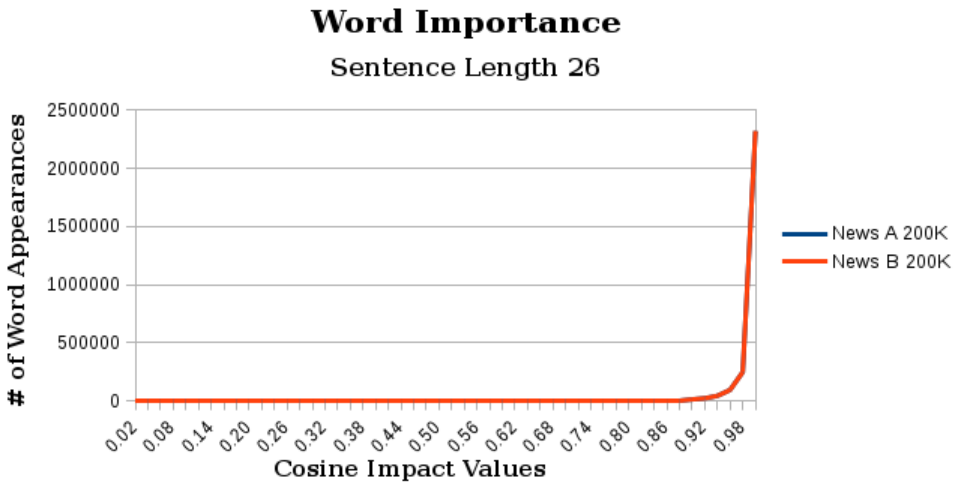

Figure 51: For each news articles document set, the CIV for all the words in sentences of length twenty-six in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

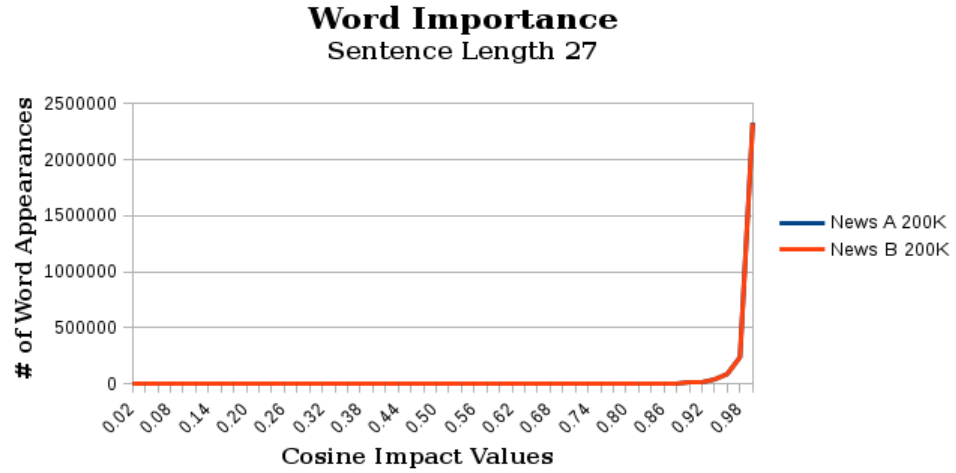

Figure 52: For each news articles document set, the CIV for all the words in sentences of length twenty-seven in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

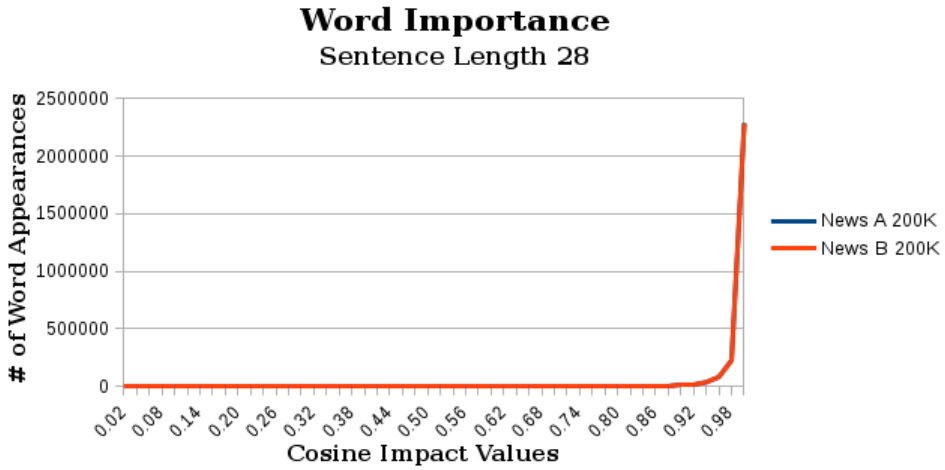

Figure 53: For each news articles document set, the CIV for all the words in sentences of length twenty-eight in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

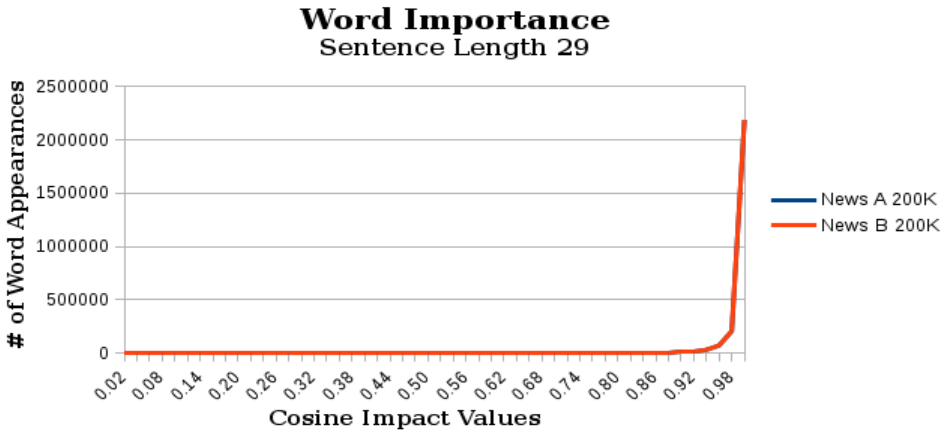

Figure 54: For each news articles document set, the CIV for all the words in sentences of length twenty-nine in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

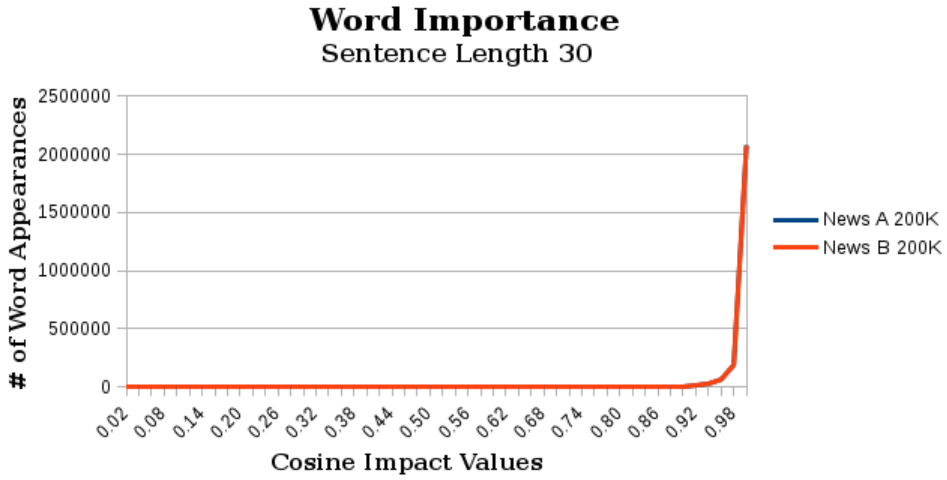

Figure 55: For each news articles document set, the CIV for all the words in sentences of length thirty in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

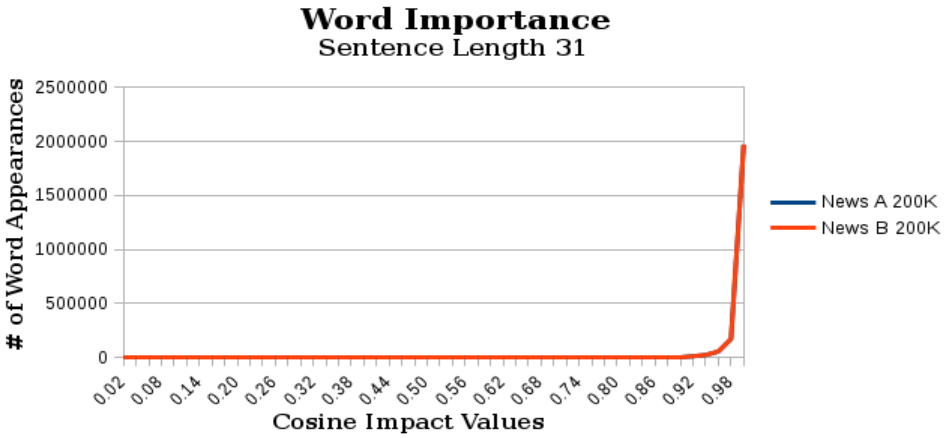

Figure 56: For each news articles document set, the CIV for all the words in sentences of length thirty-one in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

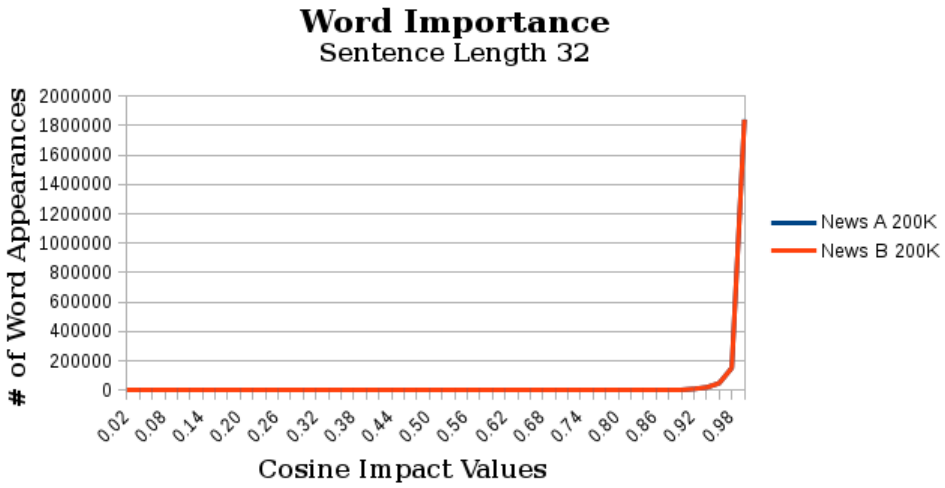

Figure 57: For each news articles document set, the CIV for all the words in sentences of length thirty-two in that document set was calculated. The number of words at each CIV interval is shown in this line graph.

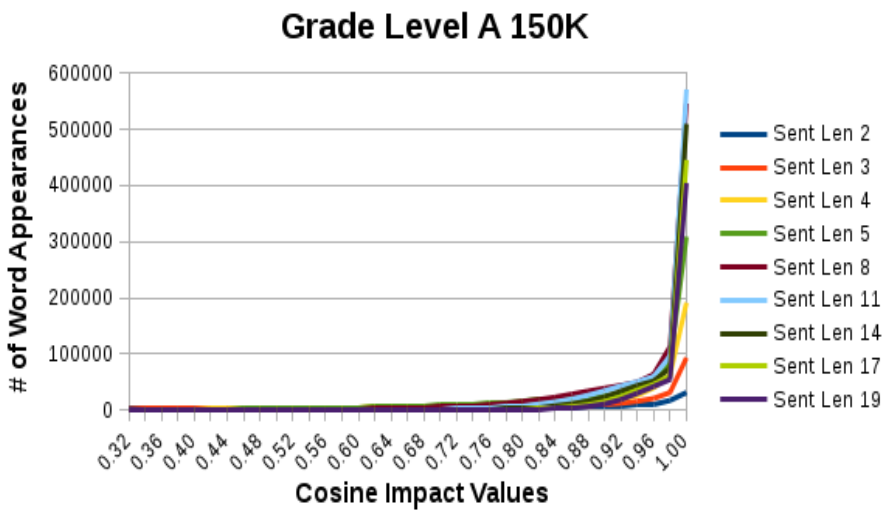

Figure 58: Distribution of CIVs for words at different sentence lengths for the grade level A document set containing ~150,000 documents.

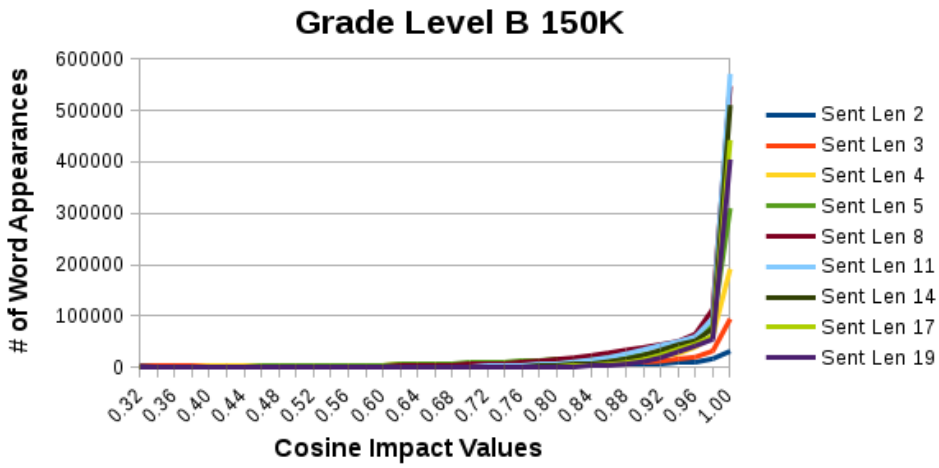

Figure 59: Distribution of CIVs for words at different sentence lengths for the grade level B document set containing ~150,000 documents.

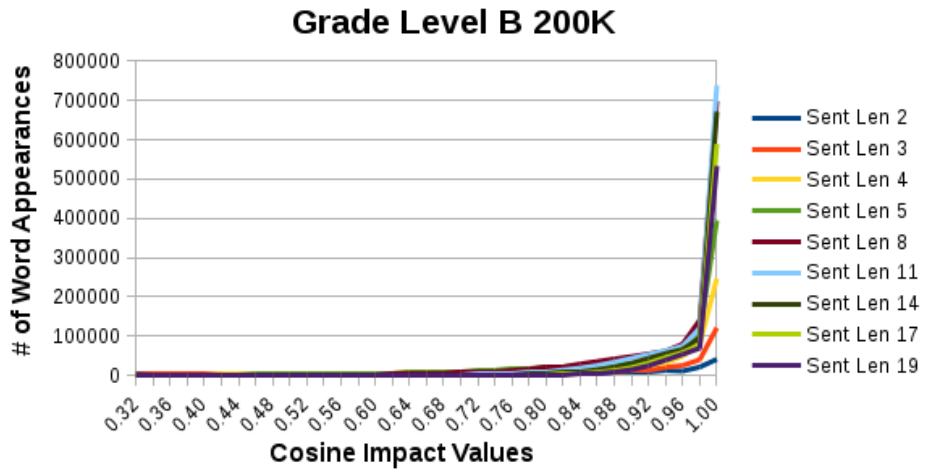

Figure 60: Distribution of CIVs for words at different sentence lengths for the grade level B document set containing ~200,000 documents.

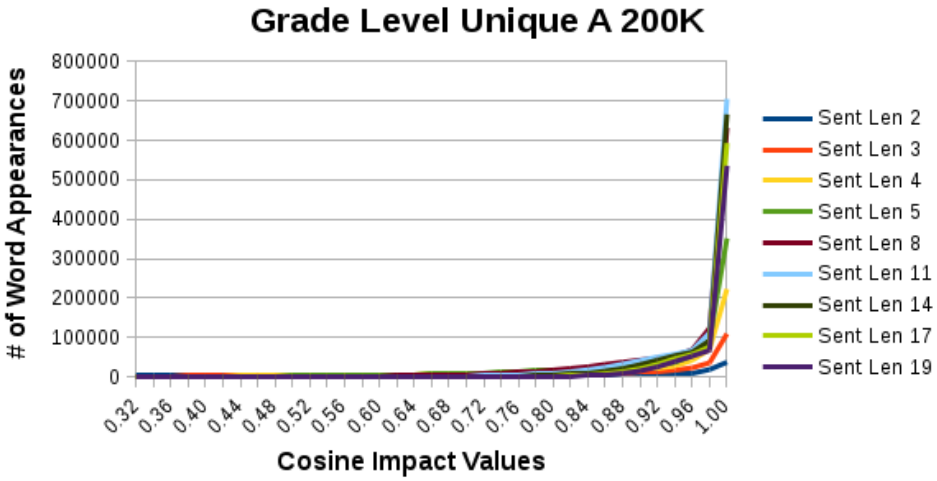

Figure 61: Distribution of CIVs for words at different sentence lengths for the grade level unique A document set containing ~200,000 documents.

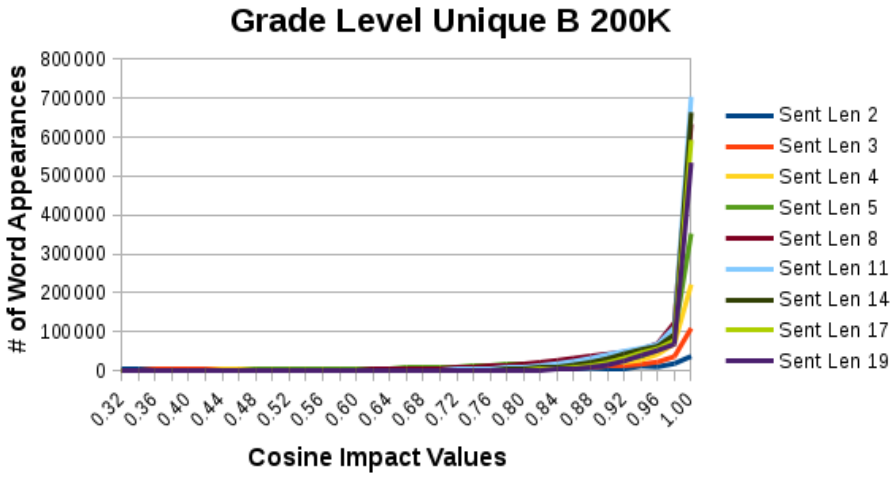

Figure 62: Distribution of CIVs for words at different sentence lengths for the grade level unique B document set containing ~200,000 documents.

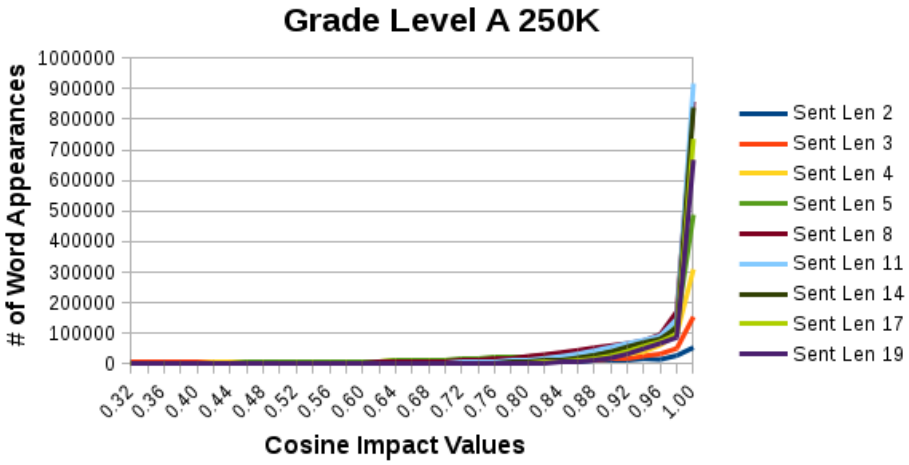

Figure 63: Distribution of CIVs for words at different sentence lengths for the grade level A document set containing ~250,000 documents.

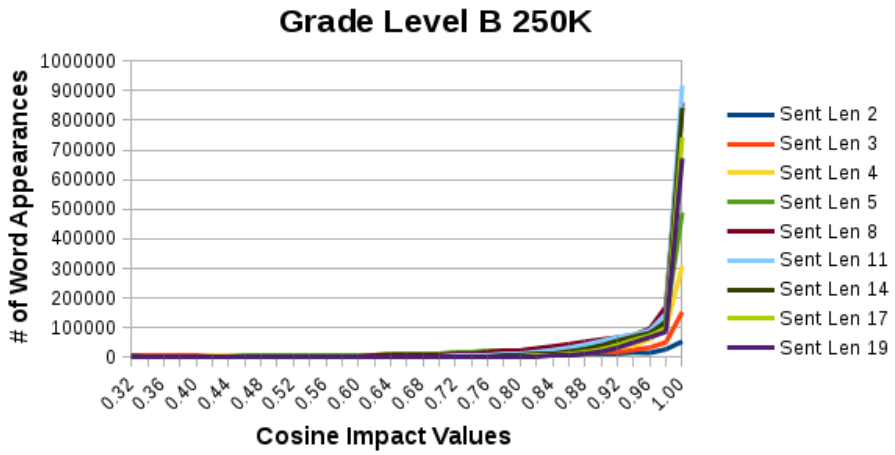

Figure 64: Distribution of CIVs for words at different sentence lengths for the grade level B document set containing ~250,000 documents.

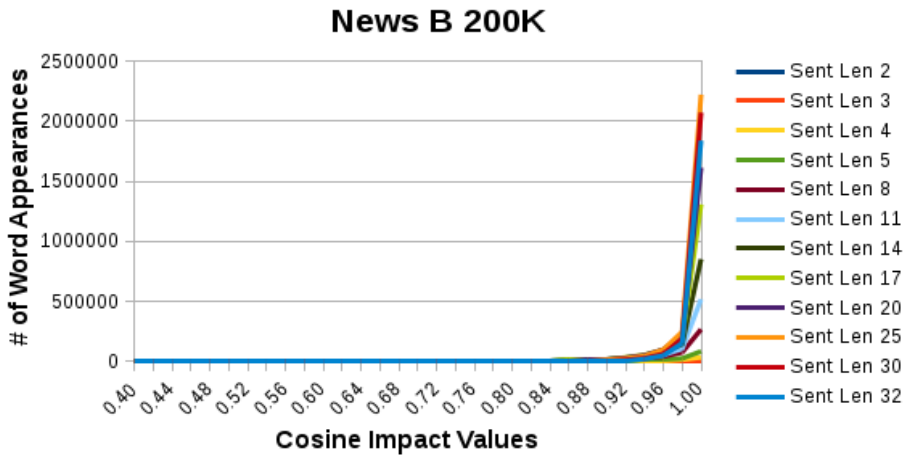

Figure 65: Distribution of CIVs for words at different sentence lengths for the news articles B document set containing 200,000 documents.

Table 17: The WSC results for using synclusters on the word ***pretty*** where the cosine similarity between the cluster centroid and the target word is greater than 0.35.

Grade Level Learning System			News Learning System		
WSC Label	# in Cluster	Cosine to Centroid	WSC Label	# in Cluster	Cosine to Centroid
guy	22	0.50	really	87	0.73
nice	11	0.40	comfortable	2	0.39
weird	23	0.37	locker	1	0.37

Table 18: Test sentences used in the WSD task for the word ***bank*** and their annotated sense determined by a human rater.

Annotated WSC Label	Sentences Using ***bank*** in this Sense
downstream	He was searching the bank for the stakes he drove in the soft mud to hold the trotlines. He scooped the plover up and waded back to the bank. Then he spotted the Indian's dugout, concealed among the thick reeds lining the bank. Zoe wandered farther down the creek bank. Nikki crawled to shore and fell on the bank.
money	This customer uses a bank machine. It was a bank robbery in progress. My piggy bank was starving. The bank, however, refuses to acknowledge the transaction. He was the bank clerk.
Different Sense	Chuck was listening to the whistles and trills and adjusting dials on a bank of electronic equipment.
Ambiguous Sense	That bank's not safe. Rosa had waved to her at the bank.

Table 19: Test sentences used in the WSD task for the word ***palm*** and their annotated sense determined by a human rater.

Annotated WSC Label	Sentences Using ***palm*** in this Sense
hand	A pile of glittering dust appeared in his palm. Yellow sap oozed onto my palm. He held them in the palm of his hand. She patted it with her hand.
gripping	She stroke my golden curls with a hand so large it seemed to palm my whole head. I suspected that he had palmed a playing card. Palming a basketball requires strong fingers and a lot of grip strength.
trees	The tree house was perched on the top of a palm tree. Later that day, just before lunch, tortoise wrapped himself in palm leaves. Palm fronds flared from the handlebars.
Different Sense	The palm of victory from the fierce and brutal ape.
Ambiguous Sense	He would look at his palm.

Table 20: Test sentences used in the WSD task for the word ***serve*** and their annotated sense determined by a human rater.

Annotated WSC Label	Sentences Using ***serve*** in this Sense
meal	Filipinos serve noche buena buffet style. The recipe serves four people. Isabel offered to serve her dinner in bed.
prepared	Somebody put deck chairs against the rail to serve as steps. Two additional spheres attached to the bottom of the station would serve as observation points for studying the undersea environment.
public	In later years, he was chosen by other presidents to serve the government, too. I feel very proud to have been elected to serve the citizens of this country. The woman will serve on a jury for a murder trial.
accept	This book will serve a useful purpose. They do not represent a complete set, but they should serve to give the reader a good idea as to the nature of this system.
Different Sense	When it was wanda's turn to serve, she couldn't get the ball over the net.
Ambiguous Sense	I am ready to serve.

VITA

Vita goes here, written in third person.