Word Ambiguity and Search: Implications for Enterprise Performance Management

Full Paper

David Schuff
Temple University
schuff@temple.edu

Robert D. St. Louis
Arizona State University
stlouis@asu.edu

Karen Corral
Boise State University
karencorral@boisestate.edu

Gregory Schymik
Grand Valley State University
schymikg@gvsu.edu

Abstract

The proliferation of unstructured data is a growing threat to effective enterprise performance management. Enterprise search is a tool to help organizations more effectively manage this document-based information. The success of full-text enterprise search is limited by ambiguity in word meanings, which can result in many documents returned which are not relevant to the searcher. While early work by Zipf provided a first attempt at quantifying the impact of this issue on search, little work has been done to demonstrate the applicability of Zipf's work to contemporary document collections. In this paper we examine whether the frequency-meaning relationship discovered by Zipf holds for contemporary document collections, and whether it consistently holds across different subject domains. We then discuss the implications of our results for the development and use of user-centered KPIs designed to measure the enterprise wide effectiveness of search activities.

Keywords

Document Management, Enterprise Search, KPI, Zipf's Frequency-Meaning Relationship, Semantic Ambiguity

Introduction

The proliferation of unstructured data is a growing threat to effective enterprise performance management. This includes the thousands of digitally stored emails, whitepapers, and technical documentation within an organization's enterprise document repository. Gartner (2013) estimates that 80% of a company's information assets are captured in unstructured content. They also claim that this unstructured content remains "grossly underutilized" and "largely unexplored," resulting in missed opportunities for greater productivity. One cause of this is the difficulty in navigating unstructured data due to the far-reaching and significant impact of ambiguity in language. This paper seeks to better understand how this ambiguity differs across different subject domains and its implications for document search and enterprise performance management.

Effective enterprise search is a key tool for organizations to manage this growing corpus of document-based data. However, only 43% of organizations view the management of unstructured data as a high priority. This is troubling, as the ability to successfully navigate an organization's collection of digital content is critical to employee productivity. McKinsey claims that a typical employee spends 20% of their work week looking for internal information from documents or colleagues; making that information searchable can reduce employee search time by 35%. This lost time could be reallocated to tasks more valuable to the organization.

Even more troubling is the fact that 55% of employees believe that it is hard or very hard to find the information they seek, and 30% are mostly dissatisfied with their organization's search applications. When search fails, the impact to an organization can be significant, as the employee must either make

decisions without complete information or attempt to re-create the knowledge that has been lost. A recent Gartner study estimates the potential impact on enterprise performance, suggesting that a company with at least 1,000 employees can lose up to $2.4 million each year due to ineffective search.

Much of this problem is a result of the inherent difficultly in implementing effective enterprise document search. The source of this difficulty is that, even with recent advances in business analytics, it still is more difficult to clearly define the criteria for search success with unstructured data compared with structured data (Corral et al. 2010). We know if we've retrieved the correct piece of structured data by looking at the columns and rows returned from a deterministic search to verify it matches our information needs. For unstructured data, there is a probabilistic matching algorithm that determines relevance, and true verification can only be done by reviewing the document for relevant, useful information.

A common solution for enterprise document retrieval is full-text keyword search. For these tools, the document's relevance score is based on matches between the keywords used in the search and the content of the document. It is a simple and appealing option because it has low setup cost and can be implemented as a turnkey solution (Schymik et al. 2015). However, full-text search has significant drawbacks (Schymik et al. 2009). The main problem lies in an inherent characteristic of language called semantic indeterminacy. Semantic indeterminacy means that the same word can have different meanings in different contexts. For example, the word "bank" can refer to a financial institution, a collection of related objects, or the side of a river (Corral et al. 2007).

However, the employee conducting the search usually has a single meaning in mind; this means keyword searches will usually return a mixture of relevant and irrelevant documents. The searcher must wade through the results to find the documents that match with the original intent of the search. This problem is especially acute for enterprise document collections, because these tools cannot rely on the referral and personalization mechanisms of web-based search tools to supplement the limited effectiveness of word-level matches.

It is because of this that keyword-based enterprise document search is particularly sensitive to the effects of semantic indeterminacy. Therefore, it is useful to have a more thorough understanding of how the inherent ambiguity in language differs across contexts (i.e., subject domains). Zipf (1945) pioneered work in this area by examining the relationship between a word's frequency and its number of meanings. Understanding that relationship is fundamental to estimating how many search results (relevant and irrelevant) will be returned. It is from this relationship that key performance indicators (KPIs) can be developed that assess the impact of search on organizational performance, since the nature of the result set will necessarily impact the time to find a document, and ultimately, worker productivity.

We build on Zipf's original study by looking at several contemporary document collections to verify that Zipf's relationship still holds, and how robust this relationship is across subject domains. The next section describes Zipf's meaning-frequency relationship of words, and discusses its implications for enterprise search. We then detail the design and results of our experiment verifying whether this relationship exists across multiple corpuses and domains. We conclude with the implications of our findings for enterprise search and enterprise performance management, and outline an agenda for future research.

Word Meanings

As part of his principle of least effort, Zipf relies on his hypothesized "orderly distribution of meanings" (Zipf, 1965). That is, the average number of meanings a word will have in a document collection varies inversely with its rank frequency. To test this hypothesis, Zipf (1945) used E. L. Thorndike's 1932 *A Teacher's Word Book of 20,000 Words*. Thorndike's book was created expressly to provide elementary teachers with guidance of which words should be covered in the curriculum. Thorndike's words[1] came from 41 different sources, including literature for children, the Bible, English classics, elementary-school text books, newspapers, correspondence, and books about "cooking, sewing, farming, [and] the trades." Also, Thorndike used only word lemmas, counting "derivatives under their primary forms" (p. *v*). Zipf used the *Thorndike-Century Dictionary* as the source for the number of meanings for each word. He

[1] The 1932 version was unavailable to us but we had access to the 1927 version of the text. In that Thorndike lists only the first 10,000 most commonly occurring words. His later version was an extension of this work using (what appears from Zipf's comments to be) the same text sources.

counted the number of meanings for a sample of the 20,000 most frequently used words. The average number of meanings per word for every successive set of 1000 words was plotted. When the average number of meanings per word was plotted against its rank on a log/log scale, the results were very close to the expected -0.5 slope (-0.4899 ± 0.0030) once the first 500 words are removed. The first 500 words were removed because they have an insufficient number of meanings. These words are commonly referred to as *stop words* and are usually removed from linguistic analyses.

Zipf (1945)[2] proposes that words with the most definitions are used most frequently due to two opposing forces. The first reason is the "speaker's economy." From the speaker's perspective it is easiest to express his/her meaning using the fewest words possible – at the extreme, using just one word which encompasses all meanings. However, the "auditor's economy" prefers words with fewer definitions to comprehend the speaker's meaning – at the extreme every word has only one meaning. Zipf postulated that these opposing forces explained the inverse relationship between the average number of meanings per word and the rank of the words. Speakers use words with many definitions, but listeners are able to understand only if words are adequately unambiguous.

While using Thorndike's collection was preferable to using the words obtained from some smaller collections which existed at the time, it is an open question if that corpus is still generalizable today. Thorndike's word list was created 84 years ago for use by elementary school teachers from many texts intended for children. It has been shown that specific texts will have some variation in the most frequently occurring words (Manning and Schütze, 1999). Similarly it is possible that the most frequently occurring words can differ based on the nature of the collection of documents. This could have significant implications for search of organizational corpora.

Methodology

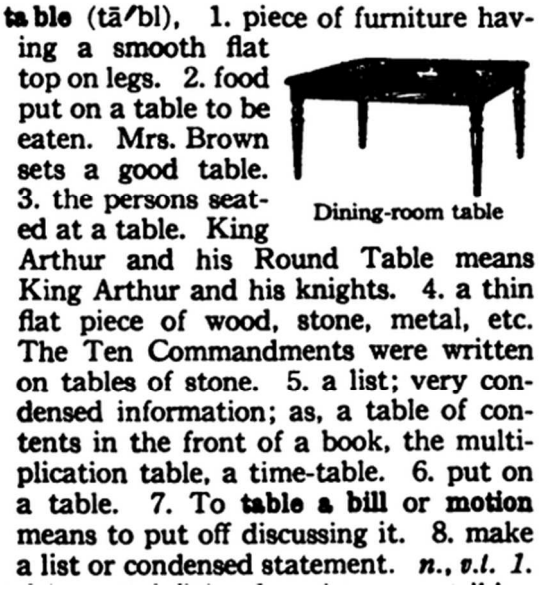
For our study, word frequency data was taken from the Corpus of Contemporary American English (COCA) (available at http://corpus.byu.edu/coca). This corpus is designed for research (Davies, 2010) and consists of data on over 520 million word occurrences from over 190 million texts from a combination of sources. The corpus was established in 1990 with 20 million words and is updated annually with another 20 million word occurrences from a consistent balance of genres (Davies, 2009). The frequency data set includes the rank of the word and its part of speech (noun, verb, adjective, etc.), along with the frequency of occurrence in the total collection and for each genre and subgenre.

COCA is derived from words that occur in spoken conversations, fiction articles and books, popular magazine articles, newspaper articles, and articles in academic journals, as described by Davies (2009). These five source categories were chosen in order to keep consistency with the British National Corpus. The corpus is generated mostly via a set of automated scripts that look for new publications in each of the sources every six months. Words are collected in roughly even amounts across sources – each source in the list is queried for new articles, publications, and texts until the targeted number of words for that source for that time period is collected. The spoken word collection includes transcripts of unscripted conversations from over 150 different radio and television programs such as *Good Morning America* (from ABC), *60 Minutes* (from CBS), and *All Things Considered* (from NPR). The fiction collection includes short stories and plays from literary, children's, and other popular magazines; first chapters of first edition books (starting in 1990); and movie scripts. The popular magazine collection includes articles from nearly 100 different magazines including *Time, Fortune, Cosmopolitan, Christian Century, Sports Illustrated, Men's Health,* and *Good Housekeeping*. The newspaper collection is built from ten US newspapers, taking care to sample across various sections of each paper (local news, sports, opinion, etc.). The list includes: *USA Today,* the *Atlanta Journal Constitution,* the *New York Times,* and the *San Francisco Chronicle.* The academic journal collection sources articles from nearly 100 different peer-reviewed journals. The selection was made so that it covers the entire range of Library of Congress classifications for such journals: education, history, geography/social science, law/political science, humanities, philosophy/religion, science/technology, medicine, and miscellaneous. The distribution of sources serves two purposes. First, it creates a true representation of the language in use at any point in time. Second, it allows year-to-year changes in speech to be accurately monitored.

[2] Zipf explained this idea in more detail in his book, *Human Behavior and The Principle of Least Effort* (1965).

COCA's presents words in lemmatized form, which allowed us to keep our analysis consistent with the work done by Thorndike (1927) and Zipf (1945), since the lemmatized form of a word is what typically appears in a dictionary. The data set lists the lemma, part of speech, and word frequency for each lemma overall and for each genre and subgenre. This means that a given word might have multiple entries in the dataset. For example, the word "her" appears in separate entries as both an adjective and a pronoun, with frequency data associated with each instance of the lemma. The dataset contains stop list words but we removed them prior to completing our analysis[3]. More specific detail on the composition, structure, comparison to other corpuses, and usage of the corpus can be found in Davies (2009).

We used the *Merriam Webster Dictionary and Thesaurus* to determine the maximum number of meanings for each lemma in the COCA. This dictionary is up-to-date and extremely popular – it is ranked first in sales on Amazon. It also provides an API that can be used to automate the process of counting the maximum number of meanings for a word (http://www.dictionaryapi.com). In his 1945 study, Zipf used fourteen different graduate students to count the maximum number of meanings for the 20,000 words that were used in his study. Each student was given a list of words, and asked to look up the maximum number of meanings for each word in the *Thorndike Century Senior Dictionary*. Consider, for example, the word "table". Figure 1 shows the entry in the Thorndike Century Senior Dictionary for the word "table." Zipf's student presumably would have seen that there are eight different meanings for the word, and recorded the number 8 in her/his worksheet. Some data entry errors obviously occurred, but they should have been random, and thus not biased the estimate for the average number of meanings.



ta ble (tā′bl), 1. piece of furniture having a smooth flat top on legs. 2. food put on a table to be eaten. Mrs. Brown sets a good table. 3. the persons seated at a table. King Arthur and his Round Table means King Arthur and his knights. 4. a thin flat piece of wood, stone, metal, etc. The Ten Commandments were written on tables of stone. 5. a list; very condensed information; as, a table of contents in the front of a book, the multiplication table, a time-table. 6. put on a table. 7. To **table a bill** or **motion** means to put off discussing it. 8. make a list or condensed statement. *n., v.t. 1.*

Dining-room table

Figure 1. Entry for the Word "Table" from Thorndike Century Senior Dictionary

If one looks up the meaning of the word "table" in the *Merriam-Webster Dictionary and Thesaurus* (http://www.merriam-webster.com), more information is returned than is in the *Thorndike Century Senior Dictionary*. Part of that information is shown in Figure 2. From this information, it is less clear how many meanings the word "table" has. One could simply look at the numbered entrees and say there are six meanings. However, entry number 3 has three different meanings: meanings a, b, and c. If we count numbered and lettered meanings, then are 11 different meanings. But entry 3b has two meanings. If we count every enumerated meaning, then there are 14 different meanings. Moreover, table can be used as a noun, verb, or adjective.

[3] Stop words were identified from http://www.ranks.nl/stopwords. We used the MySQL Stopwords.

To achieve consistency with respect to counting the maximum number of meanings that a word has, we wrote a Java program to retrieve definitions from the Merriam-Webster API and count the meanings. The program parsed the XML output generated by the API according to these rules:

1. Count only entries for nouns, verbs, and adjectives
2. Count entries as distinct meanings only down to two levels. That is count any entry labeled with a just a number (e.g., 1), or a number and a letter (e.g., 2a), as a unique meaning; but do not count entries labeled with a number and a letter and a number (e.g., 2a(1)) as a unique meaning.
3. Ignore definitions for kids, medical definitions, and British definitions

Entries in the third level of the hierarchy were not counted as unique because they were judged to be too similar to other definitions. Children definitions were not counted because they were judged to be repetitive with the adult definitions. British and Medical definitions were not counted because they were too specialized to represent common usage. These rules allowed us to automate the process of determining the maximum number of meanings for a word. Using these rules, for example, there are 11 distinct meanings for the word "table."

When conducting his study, Zipf took E.L. Thorndike's 1932 *A Teacher's Word Book of 20,000 Words,* ordered the words by their frequency of occurrence, and put them into 20 groups of 1000 words each. He then computed the average number of meanings for the words in each group, and plotted the logarithm of the average number of meanings for the group against the logarithm of the group number, where the group number ran from 1 to 20, with 1 being the group with the most frequently occurring words, and 20 being the group with the least frequently occurring words. This plot resulted in a straight line with a slope of approximately -0.50.

Results

To see whether Zipf's results still hold, and whether they are consistent across different document collections, we examined the fiction, popular magazine, newspaper, and academic word collections in COCA. These four collections represent the entire COCA database except for spoken word transcripts. We did not include those documents since spoken word collections are the least similar to the types of documents found in an enterprise document store. We then eliminated stop words from these collections using the MySQL Stopword List[4]. Within the COCA collection, separate word frequencies are given for the parts of speech such as nouns, verbs, and adjectives. Because we want the total frequency for each word, not the frequency by part of speech, we combined the part-of-speech rows for each word into a single row with a single total frequency. After eliminating the stop words and combining the part of speech rows, we selected the 20,000 most frequently occurring words for each category: fiction, popular magazine, newspaper, and academic. We then ran our Java program on each word in each category to determine the total number of meanings for each word, grouped the words into 1000 word groups, and calculated the average number of meanings for each group. Figure 3 shows the plot of the average number of meanings for each group versus the group number, and Figure 4 shows the plot of the logarithm of the average number of meanings for each word versus the logarithm of the group number. Table 1 shows the regression results for when the dependent variable is the logarithm of the average number of meanings and the independent variable is the logarithm of the group number.

Although the results differ slightly from one collection to the next, it is clear that the relationship between the log of the average number of meanings and the log of the rank is linear for all groups (the R-Squared is 0.99 or higher for each group). Thus the Meaning-Frequency Relationship of Words (sometimes called the orderly distribution of meanings law) that was discovered by Zipf in 1945 still holds, and is robust across a wide variety of document collections.

[4] Available at https://dev.mysql.com/doc/refman/5.5/en/fulltext-stopwords.html

Full Definition of TABLE

1 : TABLET 1a

2 **a** *plural* : BACKGAMMON
b : one of the two leaves of a backgammon board or either half of a leaf

3 **a** : a piece of furniture consisting of a smooth flat slab fixed on legs
b (1) : a supply or source of food (2) : an act or instance of assembling to eat : MEAL <sit down to *table*>
c (1) : a group of people assembled at or as if at a table (2) : a legislative or negotiating session <the bargaining *table*>

4 : STRINGCOURSE

5 **a** : a systematic arrangement of data usually in rows and columns for ready reference
b : a condensed enumeration : LIST <a *table* of contents>

6 : something that resembles a table especially in having a plane surface: as
a : the upper flat surface of a cut precious stone — see BRILLIANT ILLUSTRATION
b (1) : TABLELAND (2) : a horizontal stratum

—on the table

: up for consideration or negotiation <the subject is not *on the table*>

—under the table

1 : into a stupor <can drink you *under the table*>

2 : in a covert manner <took money *under the table*>

See table defined for English-language learners

See table defined for kids

Figure 2. Partial Results for the Word "Table" from Merriam-Webster

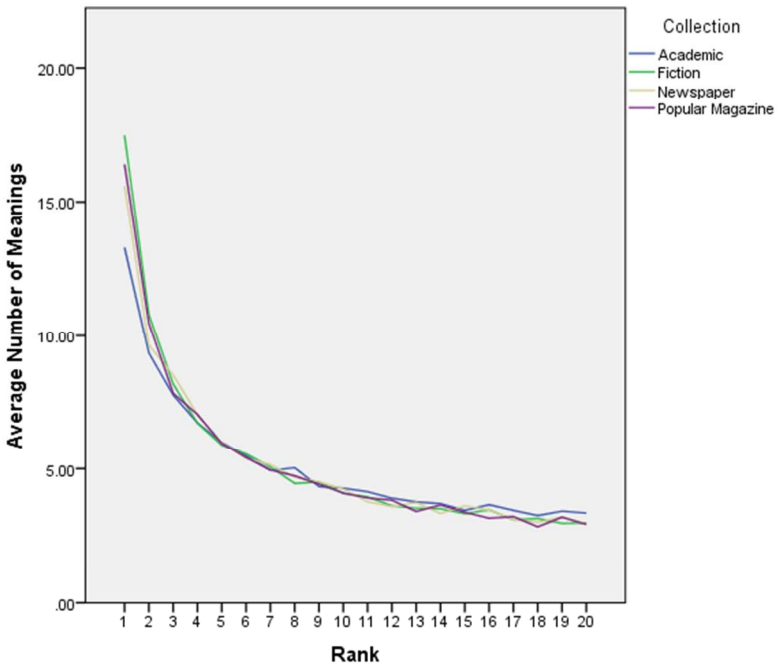


Figure 3. Average Number of Meanings versus Rank

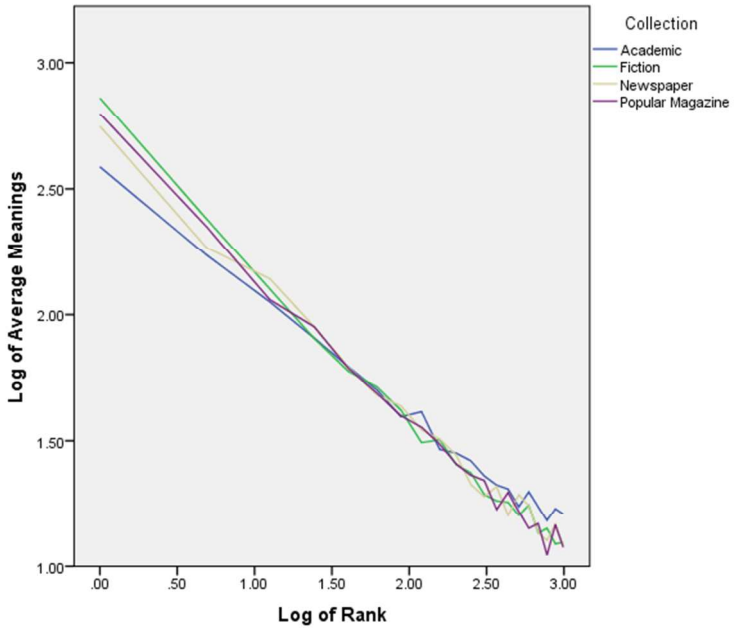


Figure 4. Log of Average Meanings versus Log of Rank

	Academic	Fiction	Newspaper	Popular Magazine
R-Squared	0.992	0.991	0.990	0.992
Constant	2.556	2.763	2.695	2.727
Coefficient	-0.470	-0.578	-0.544	-0.563
P-Value	0.000	0.000	0.000	0.000

Table 1. Regression Results for Log of Average Meanings versus Log of Rank

Conclusions and Future Work

Organizations face a daunting task in managing their unstructured data. Specifically, it will become increasingly challenging for an employee to find the documents they seek through enterprise document search. This can result in lost time and lost productivity, ultimately negatively impacting organizational performance. While KPIs exist for the performance of the database engine behind enterprise search, such as cache utilization and the time required for index "warm up" (Rigsby, 2011), there are few KPIs to evaluate the results of the enterprise search process itself.

The goal of this paper is to better understand the impact of one important limitation of enterprise document search: semantic indeterminacy. This work makes several important contributions to theory and practice. Our findings contribute to linguistic theory by confirming that words follow the orderly distribution of meanings law across four different corpuses – academic articles, newspaper articles, popular magazine articles, and works of fiction. Estimating the number of meanings a word will have in a document collection is important, as it indicates how many irrelevant results may appear in the results of a search. For practice, this can form the basis of several KPIs to enable practitioners to assess search results, such as the percentage of relevant documents returned and the number of documents in the result. It is also strongly related to estimating user-centered KPIs such as the time required to find a target document and the number of queries required to find the target document, both of which have a direct impact on worker efficiency and organizational performance.

Another contribution to linguistic theory is our finding that there is a steep drop-off in the number of meanings – in other words, while a small number of well-known words have many meanings, many less frequently used but more specialized words have few meanings. This may mean there is less impact of multiple word meanings than previously thought. For practice, this key finding has important implications for enterprise search. While semantic indeterminacy has been shown to be a cause of poor search results (Schymik et al., 2015), it may not be the only factor. Instead, it may be that when performing an open-ended search, the searcher may not know how to select the right keywords, opting for terms that ultimately "miss the target." Future research can investigate this issue by examining peoples' keyword choices when performing enterprise search, perhaps through a laboratory experiment.

Future research is also necessary to determine whether our findings hold for larger collections, although the drop-off in meanings occurs early enough to make it likely that our results will be robust. Even more importantly, our results should be replicated based on the actual number of meanings used in the collection instead of the maximum number of meanings in the dictionary. Currently, no one knows how quickly words approach their maximum number of meanings as the size of a document collection increases.

Overall, this study demonstrates strong evidence that providing context is an important component of enterprise search, whether it is to resolve ambiguity in word meanings or provide a guide for searchers leading them to relevant content within the enterprise document store. Enabling employees to spend less time searching for documents and facilitating access to more relevant information will allow organizations to focus on true value-added activities. This will make enterprise document search a tool for achieving higher organizational performance instead of an obstacle standing in the way of productivity.

REFERENCES

Anonymous. 2015. "Big Data and Analytics: The Big Picture," retrieved February 19, 2016 from http://www.idgenterprise.com/resource/marketing-tools/big-data-and-analytics-the-big-picture-infographic/

Chui, M., Manyika, J., Bughin, J., Dobbs, R., Roxburgh, C., Sarrazin, H., Sands, G., and Westergren, M. 2012. "The social economy: Unlocking value and productivity through social technologies," retrieved February 19, 2016 from http://www.mckinsey.com/industries/high-tech/our-insights/the-social-economy

Corral, K., Schuff, D., Schymik, G., and St. Louis, R. 2010. "Strategies for Document Management," *International Journal of Business Intelligence Research* (1:1), pp. 64-83.

Corral, K., Schuff, D., St. Louis, R.D., and Turetken, O. 2008. "A Model for Estimating the Savings from Dimensional Versus Keyword Search," in J. Erickson and K. Siau (Eds.), *Advanced Principles for Improving Database Design, Systems Modeling, and Software Development*, Hershey, PA: IGI Global, pp. 146-157.

Davies, M. 2009 "The 385+ million word *Corpus of Contemporary American English* (1990 – 2008+): Design, architecture, and linguistic insights," *International Journal of Corpus Linguistics* (14:2), pp.150-190.

Davies, M. 2010 "The Corpus of Contemporary American English as the first reliable monitor corpus of English," *Literary and Linguistic Computing* (25:4), pp. 447-464.

Legernes, H. 2013. "Survey results: report on enterprise search adoption survey," retrieved February 19, 2016 from http://conferences.infotoday.com/stats/documents/default.aspx?id=8556&lnk=http%3A%2F%2Fconferences.infotoday.com%2Fdocuments%2F173%2F0945_Legernes(1).pptx

Manning, C. D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press.

Rigsby, J. 2011. "Analytics & Monitoring Key to Effective Enterprise Search," retrieved February 29, 2016 from http://www.cmswire.com/cms/information-management/analytics-monitoring-key-to-effective-enterprise-search-011492.php

Shell, S. 2013. "Findability, not searchability," retrieved February 19, 2016 from http://conferences.infotoday.com/stats/documents/default.aspx?id=8526&lnk=http%3A%2F%2Fconferences.infotoday.com%2Fdocuments%2F173%2FA203_Shell.pdf

Schymik, G., Corral, K., Schuff, D., and St. Louis, R. D. 2015. "The Benefits and Costs of Using Metadata to Improve Enterprise Document Search," *Decision Sciences* (46:6), pp. 1049-1075.

Schymik, G., St. Louis, R., and Corral, K. 2009. "Order of magnitude reductions in the size of enterprise search result sets through the use of subject indexes," in *AMCIS 2009 Proceedings*, paper 195, San Francisco, CA.

Stewart, D. 2013. "Big Content: The Unstructured Side of Big Data," retrieved February 19, 2016 from http://blogs.gartner.com/darin-stewart/2013/05/01/big-content-the-unstructured-side-of-big-data/

Thorndike, E. L. 1927. *A Teacher's Word Book*. 2nd ed. Teachers College, New York: Columbia University.

Zipf, G. K. 1945. "The Meaning-Frequency Relationship of Words," *The Journal of General Psychology*. (33:2), pp. 251-256.

Zipf, G. K. 1965 facsimile of 1949 edition. *An Introduction to Human Ecology*. New York: Hafner Publishing Co.